Dunstkreis: Detecting Radical Right Populist Speech Contagion in German State Legislatures with a Data-Driven AI/ML Pipeline

Aaron Rudkin

August 24, 2026

Abstract

Scholars and commentators alike have expressed concern about the risks associated with the election of radical right populist parties to legislatures. I study the effect of populist entry on legislator speech by building a data-driven, low-assumption *populism detector* pipeline, using document embeddings to encode the features of legislative speech, large language models to annotate training data, and model selection to estimate the relationship between the features and labels. Drawing on the exogenous entry timing of the *Alternative für Deutschland* into the sixteen German state legislatures (*Landtage*) and a novel corpus of all speeches given in these legislatures from 2008 to 2024, I find that non-AfD parties adopt more populist legislative speech in response to AfD entry. The effect is positive, moderately large ($\approx$ 3 p.p., 61% relative to baseline), and robust to inferential threats, including spillover. These findings substantiate concerns about downstream effects of populist entry and highlight speech contagion as a mechanism sustaining institutional equilibrium. In addition, this article offers a workflow and guidance for integrating artificial intelligence tools into computational text analyses.

Keywords: radical right populism, text-as-data, ideological measurement, legislative speech, legislatures, institutions, large language models

1 Introduction

Researchers and activists have long worried about the growing popularity of radical right populist (RRP) parties in Europe and the broader West: if RRP parties win elections and form governments, they may preside over an erosion of democratic or pluralist values and undermine democratic institutions. But the graver threat to representation is that they may achieve these goals even *without* winning, through their ability to force parties to accommodate their demands via electoral or legislative pressure (A. Downs 1957; Sartori 1976). In decades past, minimum electoral thresholds were observed to deter extremist entry (Blais 1991; Jackman and Volpert 1996). RRP parties now routinely exceed these electoral thresholds and have entered national legislatures in over 90% of OECD countries. Even as mainstream parties impose *cordons sanitaires* with respect to allowing these parties to join governments (Art 2007; Van Spanje and Van Der Brug 2007), RRP parties nonetheless gain the ability to vote, propose amendments, question members of government, and participate in plenary debates, making it difficult to fully shut them out (Mares 2026). I study the downstream effect of RRP party entry and presence on the parliamentary speech of incumbent parties.

Past research has been mixed: incumbent parties only sometimes accommodate RRP parties or their ideas, and those that do are typically understood to be engaged in a spatial ideological adaptation. While RRP parties appear to have fair success in forcing their ideas to be considered in campaigns and elections, their record in legislatures is apparently more modest. However, the record so far suffers from two key limitations: first, most research on RRP entry or presence relies on simple before-after comparisons, which lack causal identification and conflate the effects of "demand-side" pressures with the "supply-side" effects of interest. If mainstream parties adopt RRP positions in response to voter demands, this may threaten liberal values but is not inherently an affront to democratic representation. If, on the other hand, the *entry* (or *presence*) of RRP parties inside the legislature drives additional adaptation by incumbent parties, this militates towards a need to understand exactly how institutional designs contribute to such contagion. Second, existing research on RRP parties has not focused its attention on how best to measure RRP speech, and recent technological advances can overcome long-standing technical hurdles.

I leverage plausibly exogenous variation in German sub-national election timing and results to study the entry of the *Alternative für Deutschland* (AfD) into German *Landtage* (state legislatures) and its subsequent effect on the legislative speech of mainstream (non-RRP) parties. I make three contributions to the literature. First, I assemble a comprehensive, cleaned, and annotated corpus of all speech in all Landtage from 2008 onward. Second, I provide a workflow to detect radical right populist speech using document embeddings, model selection, and LLM-based speech annotation, as well as adversarial, LLM-as-a-Judge based coder reconciliation. Finally, I speak to two substantive debates in political science: broadly, to speech and idea contagion as an important mechanism underpinning institutional equilibrium formation (Schmidt 2008) in the wake of any novel political entrant, and in a more targeted way, on the direct effect of populism entry and presence. I find that the AfD's entry produces a moderately large increase ($\approx 60\%$) in radical right populist speech by mainstream German parties.

The remainder of the paper proceeds as follows: I begin in **Section 2** by reviewing the broader literature on populism, RRP parties, and legislative speech. In **Section 3**, I review the case selection. Next, I describe the construction of the Landtag corpus and the populism detector (**Section 4**). I discuss identification and estimation strategies in **Section 5**, the results in **Section 6**, and I assess the external validity of the method and the findings, as well as other limitations and future directions for research in **Section 7**.

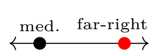
2 Past Work on Radical Right Populism and Its Effects

Populism and the Far Right: There is little consensus about what to call the "far right" or how much the label matters. Terms such as "extreme right", "populist", "radical right", "nationalist", "nativist", "xenophobic", and a vertiginous array of other labels are said to have salient conceptual differences but are nevertheless used interchangeably to refer to a discrete and generally well-understood (if porous) set of political actors. I first offer brief definitions of "populist" and then "radical right", drawn from recent literature, before proceeding to a summary of empirical research on RRP parties.

I adopt the view[1], crystallized in Mudde (2004), that populism is:

> ... an ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups, 'the pure people' versus 'the corrupt elite', and which argues that politics should be an expression of the *volonté générale* (general will) of the people. Populism, so defined, has two opposites: elitism and pluralism. Elitism is populism's mirror-image: it shares its Manichean worldview, but wants politics to be an expression of the views of the moral elite, instead of the amoral people. Pluralism, on the other hand, rejects the homogeneity of both populism and elitism, seeing society as a heterogeneous collection of groups and individuals with often fundamentally different views and wishes... As a thin-centered ideology, [populism] can easily be combined with very different [...] ideologies, including communism, nationalism, or socialism.

This frame, which views populism as discursive in nature (Hawkins 2009), has the advantage of divorcing populism from any left- or right-wing ideology. A populist can define "the people" in an inclusionary or exclusionary way, with or without reference to class or nationality (March 2012; Mudde and Kaltwasser 2013). Buttressing this definition, past studies of public opinion have observed a separation between support for populism and host ideological features (Akkerman, Mudde, and Zaslove 2014).

The "far right" portion of the label similarly requires refinement. Encoded in the term itself is a notion of spatial ideology, namely 

, that there exists an ideological left-right dimension (whose content might be redistributive preference or any other definition),

1. Other scholarship focuses on the centrality of charismatic, personalist leadership and low institutionalization to populism (Barr 2009; Knight 1998; Weyland 2001); some view populism as a neutral or even positive "corrective" to "elite consensus" (Canovan 1999; Mair 2002; Mény and Surel 2002). The former approach is not suitable for the measurement of legislative speech and the latter differs from my chosen definition only in terms of valence, not measurement characteristics.

that further there exists some "center" or median voter on that dimension, and that the far-right are those political actors who are far from the median on the right side. Such a definition is problematic in the comparative context: whose median and when? What happens when the median drifts over time, as it is surely wont to do (Fiorina and Abrams 2008; Stimson 2018)?

Instead, a label used to refer to an ideology should have a fixed meaning. Ignazi (1992), in an early and influential work, characterizes the far-right as being united by a belief in a "right-wing antisystem" – "a belief system that does not share the values of the political order in which it operates" (quoting Sartori 1976). Mudde (2007) synthesizes this and other conceptions of the far-right into a four-part "ladder of abstraction". At the base is nationalism: an ideology that assumes or aims to achieve "the congruence of the cultural and the political unit". Adding xenophobia to nationalism creates nativism, a view that "states should be inhabited [*ed. almost?*] exclusively by members of a[n ethnic] native group... and that nonnative elements (persons and ideas) are fundamentally threatening" (Mudde 2007). Adding a belief in political authoritarianism (in the sociological—psychological sense, as in Altemeyer 1981 or Oesterreich 2005) to nativism results in radical right parties[2], with radical here meaning "opposition to some key features of liberal democracy". Finally, adding a wholesale rejection of democracy produces extreme right parties.[3] In the rest of the paper, I use the "RRP" to refer to the constituent ideologies, "RRP parties" to the parties that espouse such ideologies, and "mainstream" or "non-RRP parties" to refer to other parties.

Responses and Effects: Research proposes three basic approaches that mainstream parties can take in response to the presence of a novel or fringe entrant (Atzpodien 2022; W. Downs 2001; Meguid 2008). First, they can *accommodate.* Accommodation involves political parties moving

2. Mudde prefers "populist radical right" which he distinguishes from "radical right populism" only linguistically, by averring that "populism" modifies "radical right" and thus should come in that order. I follow Mudde's definition but use the phrasing more commonly found in the literature, RRP.

3. Note that this definition studiously avoids a requirement that RRP parties have a particular view on the state's role in the economy, in large part because classically fascist parties and many otherwise well-fitting parties today endorse redistributive or interventionist economic policy (Minkenberg 2000).

to the right and adopting RRP positions or framing (which the literature calls "contagion"). This is typically argued to be a defensive strategy to ward off electoral threats, though nothing about contagion requires adaptation to be strategic or even intentional. Immigration is sometimes claimed to be a uniquely spatially cross-cutting issue, and for that reason, it is said to be able to induce accommodation *across* the political spectrum (Abou-Chadi and Krause 2020; Alonso and Fonseca 2012). Second, parties can adopt an *adversarial posture*, accepting the RRP issue as salient while maintaining or enhancing anti-RRP posturing on those issues (Matsunaga and Winzen 2025; Meguid 2005; Merrill and Grofman 2019). The expectation is that an adversarial posture drives electoral polarization around RRP issues, betting on the notion that RRP positions are, at best, expressed by a passionate minority of voters. Finally, as has historically been the case, parties can *ignore* RRP parties. This can vary from formal *cordons sanitaires* (refusing to pass legislation with RRP support or admit RRP parties to government) to the use of procedural tools to impede RRP participation in the legislature (Patton 2020)[4]. The prevailing view is that parties choose to ignore challenger parties when the challenger is comparatively weak, and choose accommodation or opposition depending on whether models of spatial threat predict more votes lost to the challenger party or to incumbent parties on the party's other flank. But this typology of responses speaks more clearly to party behavior in a campaign context than in a legislative one; even a party that tries its best to ignore an RRP entrant must still eventually choose to speak on issues of RRP interest, and theory does not provide a principled basis for predicting that an incumbent party chooses accommodation or opposition beyond spatial voting considerations.

I now turn my focus to the direct literature on effects of RRP entry on legislative speech. Once in a legislature, the RRP party gains access to *ex officio* capabilities (including funding for policy researchers and lawyers, allocated debate time, access to oral and written question

4. In Germany, the *cordon* holds, though a hung parliament in Thuringia led to a partial rupture when AfD votes were required to elect a Premier (*Ministerpräsident*); the resulting premiership lasted three days and led to abrupt shifts in vote intention toward parties of the Left. In a few dozen cases across Europe, formal *cordons* have collapsed, and parties have permitted RRPs to enter government (De Lange 2012), though they are rarely able to deliver extreme policy goods (Akkerman 2012).

mechanisms [*Anfragen* in German legislatures], the ability to demand reports or hearings, and nuisance tactics such as points of order [*Geschäftsordnungsantrag*]) that frequently leave incumbents on their back feet and unable to disrupt their influence (Heinze 2022). Absent intervention, RRP rabble-rousing may undermine the legitimacy or effectiveness of parliaments, and the only defense may involve the use of legal–institutional levers (Mares 2026). Internal tension about whether to respond to RRPs (and how) sometimes paralyzes mainstream parties (W. Downs 2001).

Despite this, empirical studies tend to find mixed results. In some instances, RRPs succeed in hijacking debates on immigration and related issues (Schwalbach 2023), though Meijers and Van der Veer (2019) find that parties do not increase issue focus on RRP pet issues. Valentim and Widmann (2023) find that RRPs in parliaments use strongly negative emotional language, but mainstream parties counter with adopting more positive emotional language, while Calvo, Bäck, and Carroll (2024) find that mainstream parties speak negatively about RRP colleagues upon first entry. After RRP/Euroskeptic entry to the European Parliament, pro-EU politicians speak more skeptically about EU issues but continue to vote pro-EU as ever, suggesting that speech and roll call voting can become unglued in some circumstances (Bélanger and Wunsch 2022; Wunsch and Bélanger 2024). In the works closest to my own, Atzpodien (2022) finds that over time, German parties speak more like the AfD in migration debates, though this analysis cannot account for whether the adaptation is demand- or supply-driven. Wagner, Schafer, and Yavuz (2025) find a durable rightward pull on immigration. Lewandowsky et al. (2021), by contrast, find that mainstream parties rarely use populist rhetoric in legislatures. I improve on past efforts with better identification, a large and long-running speech corpus, and direct measurement of RRP speech. My findings disagree with the consensus: RRP entry in German state legislatures produces a moderately large and sustained shift towards (radical right) populism in legislative speech across the ideological spectrum and across topics unrelated to immigration and integration. Before development my measurement strategy, I turn briefly to the matter of case selection.

3 The German Landtag Case and the AfD

I study the case of the entry of the AfD into the German state legislatures. Here, I describe the AfD, the role of Landtage in Germany, and how the paper's identification strategy exploits Landtag treatment timing. The AfD is preeminent among the RRP parties in Europe. Originally established in 2013 as a Euroskeptic, anti-eurozone, anti-bailout[5] party, the party rapidly moved to the (further) right (Pytlas and Biehler 2024) under the leadership of Frauke Petry and in response to pressures from high-demander groups within the party's broader institutional and social coalitions, including the *Flügel* party faction and the civil society anti-Islam group PEGIDA.

The AfD can be fairly characterized as "radical right populist" both deductively and inductively. Deductively, I begin with Mudde's typology: the AfD is radical right because it combines a nationalist and nativist view of the German *volk* with authoritarian and anti-pluralist policies, but it does not position itself in overt opposition to electoral democracy (in contrast to previous extreme right movements, including the National Democratic Party [*Nationaldemokratische Partei Deutschlands*, *NPD*] and arguably, *Die Republikaner*, though this may be a strategic position rather than a sincere one). Qualitative researchers broadly classify the AfD as "right-wing populist" (Arzheimer 2015; Berbuir, Lewandowsky, and Siri 2015) and "radical right" (Arzheimer and Berning 2019) on similar grounds. The German anti-extremism intelligence agency, *Bundesamt für Verfassungsschutz*, has designated the AfD and its youth wing as *gesichert rechtsextremistisch* (confirmed right-wing extremist) organizations, stipulating that "[t]he ethnicity- and ancestry-based understanding of the people prevailing within the party is incompatible with the free democratic order" (quoted in Kirby 2025), though German courts have intervened to prevent further action against the party. The party once fired its chief spokesman, a self-identified fascist, when he gave an interview in which he opined, "[t]he worse it goes for Germany, the better for the AfD...

5. In the wake of the 2008 financial crisis, a series of Eurozone countries (typified by Greece, but also including Ireland, Spain, and others) suffered banking and sovereign debt bond collapses that threatened the health of the Euro. The resulting bailouts totalled over 500 billion Euro.

[it's good when migrants arrive,] because the AfD does better. We could later shoot them all. That's not at all the issue. Or, gassing, or whatever you want. It's the same to me." (Translated in Deutsche Welle 2020)

Inductively, it is possible to situate the AfD in comparative perspective. It is a close ally of the far-right Freedom Party of Austria (*FPÖ*) (Der Standard 2016). Within the European Parliament, it co-founded the Europe of Sovereign Nations Group, the furthest-right political group, after it was ignominiously expelled from the previous furthest-right political group (Hix, Whitaker, and Zapryanova 2024; McElroy and Benoit 2007) over concerns it was too extreme, placing it to the right of France's National Rally, Italy's League (*Lega*), the Danish People's Party, and the Dutch Party for Freedom; and in the company of the Dutch Forum for Democracy and the French Reconquest party. Political Science datasets code each of the aforementioned parties as "populist" (Rooduijn et al. 2024; Zaslove, Huber, and Meijers 2024), "right-wing" (Döring and Manow 2012), "radical right" (Rovny et al. 2025), and "nationalist" (Merz, Regel, and Lewandowski 2016) where present.

The AfD is therefore an archetypal case to study. It has seen consistent and growing electoral success. Since it began winning seats in 2013, it has maintained a nearly unbroken record of crossing the minimum vote threshold to be seated in both the federal *Bundestag* and state legislatures. Today, it is the second-largest party in the *Bundestag*, elected in 14 of the 16 states, and the largest party in one.

Why study *Landtage* rather than the *Bundestag*? Studying state legislatures allows me to take advantage of exogenous variation in treatment timing. The *Bundestag* is a single legislature that was treated the moment the AfD entered. Comparisons before and after the AfD's entry risk confounding a temporal trend in populism (the demand effects I discuss above) with the actual effect of interest: the AfD's entry. Even designs that rely on narrow temporal discontinuities still face this bundled treatment problem. The state legislatures, on the other hand, are treated sequentially over a five-year period, allowing for overlapping periods during which some states are treated and others are untreated (and where treated states were exposed to varying "doses" of treatment). All Landtage use proportional representation in awarding seats and a minimum electoral threshold of 5%. Because the AfD's emergence and later spike in popularity occurred in response to national and international events, and because German elections take place on fixed schedules, I treat election timing as conditionally exogenous. In **Section 5** below, I further discuss treatment timing and identification considerations.

The Landtage are also a substantively fruitful set of legislatures. Germany is a highly decentralized country (by some estimates, the most decentralized major economy; see Marks, Hooghe, and Schakel 2008). *Länder* (the German sub-national unit, states) enjoy wide policy autonomy over populism-exposed policies, including education, housing, and public administration; and concurrent jurisdiction in healthcare and environmental regulation. While immigration is the purview of the *Bundestag*, states weigh in on associated issues through their autonomy over policing, criminal justice, refugee facilities, courts, and social services. This means that the substantive content of populist speech in Landtage should capture the full breadth of RRP priorities. I now describe the Landtage data collection and measurement strategy in **Section 4**.

4 Data and Measurement

Figure 1 below shows a high-level overview of the data pipeline; readers interested in skipping to the empirical strategy can find it below in **Section 5**. The central engineering challenge of this article is twofold: first, I need to gather a corpus of all speeches from each of the 16 German Landtage. Because these are not available for bulk-download, I create data scrapers for each of the 16 states. The resulting speeches require significant processing, depicted in a flowchat in **Appendix A.1.1**. In **Appendix A.1.2**, I describe one particular technical challenge of general interest: identifying speech boundaries and extracting text from multi-columnar PDF documents, which I solve by developing a novel probabilistic algorithm. The resulting corpus, after dropping speeches under 35 words in length (primarily introductions, rejoinders, oath-swearings, or other procedural text), as well as speeches by the active presiding officer, is a corpus of 583,621 speeches. In **Appendix A.1.3**, I provide descriptive statistics of the corpus that include average speech length, speech count per state, state party dominance, and date ranges covered.

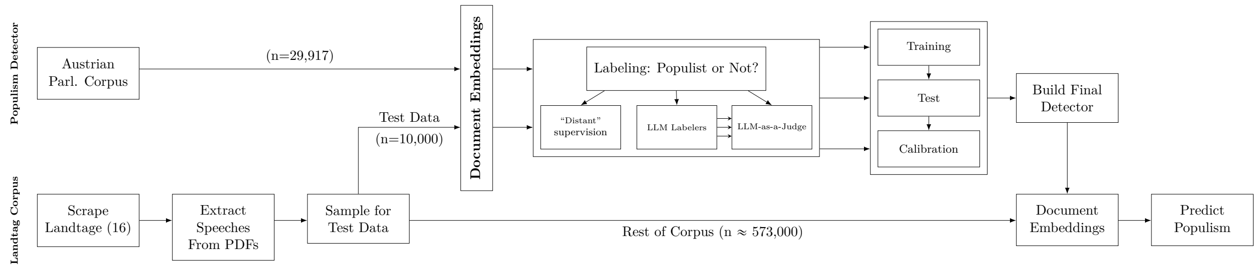


Figure 1: Populism Detector Workflow: In order to estimate the effect of AfD's entry into Landtage on levels of far-right populist speech, I need a measure of far-right populist speech. To build a populism detector, I use a corpus of Austrian parliamentary speech, along with test data taken from the German Landtag corpus. I use document embedding models to embed the speeches and label them using several labeling techniques. I fit many classes of model to the data and calibrate them. The best-performing embedder–labeler–model–calibration detector is kept and yields a measure of populism for the entire corpus.

Second, I need to measure the extent of radical right populism in any given speech, validate the measure, and apply the measure to my Landtage corpus. I describe a general-purpose methodology for this sort of text classification now.

4.1 Populism Detector

Historically, text classification meant using sentiment analysis, a method in which the researcher defines *ex ante* the exact words that constitute the dimension of interest and combines them across the body of the text (Pang, Lee, and Vaithyanathan 2002; Young and Soroka 2012). Although it has been used in the context of measuring populism (Jagers and Walgrave 2007), such an approach is suboptimal: it is sensitive to the researcher's choice of keywords, and there is no obvious "stopping rule" for declaring a list exhaustive or complete (Grimmer and Stewart 2013; Van Atteveldt, Van der Velden, and Boukes 2021);

it is monolingual and requires that the dictionary compiler has mastery of idiom in every linguistic context (Araujo et al. 2016), though this can sometimes be mitigated by machine translation (Proksch et al. 2019); it tends to conflate a speech's *sentiment* with its *stance* (e.g., capturing speech about immigration as populist irrespective of viewpoint; see Bestvater and Monroe 2023); and it is unlikely to capture features of speech (e.g., sarcasm) that modify a word's meaning using context (Joshi, Bhattacharyya, and Carman 2017). In the alternative, researchers can use scaling methods, which classify speech along constructed spatial dimensions according to either anchor texts (as in the popular Wordscores method, see e.g., Klemmensen, Hobolt, and Hansen 2007; Laver, Benoit, and Garry 2003; Lowe 2008), or dynamically (as in Wordfish, Slapin and Proksch 2008). These methods also have limitations with respect to encoding concepts, often relying on users to manually subset texts to a single substantive dimension or risk scaling features unrelated to the concept of interest.

Alternatively, researchers can use human coders to label documents directly, typically according to a rubric (e.g., Laver and Budge 1992; Rooduijn, De Lange, and Van Der Brug 2014). Coding social scientific concepts is fraught, and the best way to mitigate human errors is to adopt a multi-coder approach and measure inter-coder reliability (Scott 1955). Where coders diverge, a reconciliation process is used to produce a single set of labels (Campbell et al. 2013; DeBell 2013). All of this is extremely costly, even given the existence of lower-cost offshoring or crowdsourcing options, and most corpora are simply too large to code by hand. This leads to the approach typical of more recent social science text-as-data work: treating the task as a supervised learning problem, labeling a small portion of the documents, and then using a statistical learner to map words to labels for the purpose of predicting unlabeled data. The model may be as simple as looking for "wedge words" that differentiate between pre-labeled classes (Mosteller and Wallace 1963), or it might involve using a more flexible learner.

Sentiment, scaling, and human-coding approaches all introduce what critics of social science call "researcher degrees of freedom" or "forking paths" (Gelman and Loken 2013; Simmons, Nelson, and Simonsohn 2011): channels through which researchers can (un/)intentionally, through a series of individually defensible choices, obtain desired and significant results that are false or fragile. Practical concerns supplement epistemological ones. Each of these approaches is unilingual. Each approach relies on the so-called "bag of words" assumption about text (Qader, Ameen, and Ahmed 2019), wherein a computer disregards all information about word order or context and simply counts the occurrences or frequency of each word (or n-gram). Though such an assumption is "clearly false" (McCallum and Nigam 1998), it has been the bedrock of text-as-data work for decades.

Recent advances in computation and artificial intelligence mean that researchers can move beyond these limitations. We no longer need to treat individual words, divorced of context, as the features we examine. Enter word embeddings, a method that involves ingesting a large corpus of text, pre-training a model that observes how often words occur near one another, and then assigning each word a position in a high-dimensional space based on what it has learned. "King" and "Queen" have a relationship because they tend to appear near each other and *also* near the same other words (Church 2017). If words can be embedded, then so can sentences (Arora, Liang, and Ma 2017; Reimers and Gurevych 2019). If sentences can be embedded, then so can documents (Dai, Olah, and Le 2015; Lau and Baldwin 2016; Wu et al. 2018). Document embedding models assign each document a position in a high-dimensional space (in the selected model, 768 dimensions; in other models considered, as many as 3,096) according to the meaning it carries. The intuition is that documents that are semantically similar are closer together in space, while those that are semantically distant are spatially distant. In **Figure 2**, I demonstrate this intuition by reducing the 768-dimensional corpus to 3 dimensions using UMAP dimensionality reduction (McInnes, Healy, and Melville 2018).[6] The resulting features encode the input concept more reliably and render it easier for statistical models to draw boundaries between classes.

6. This is a toy example; I selected the documents to make the stylized point. There is no general sense in which this particular *3-dimensional representation* of this data encodes the same relationships as the full embedding space.

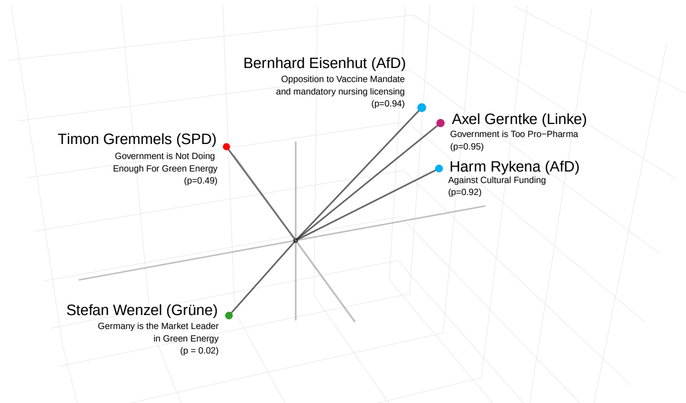


Figure 2: Visual Representation of Document Embeddings: This figure depicts a 768-dimensional document embedding of Landtag speeches, projected down into 3-dimensional space, for ease of visualization. Each speech is assigned a point in the space according to its semantic content. Here, three "high populism" speeches load heavily high on the vertical axis and close together. A medium-populism speech by an *SPD* speaker loads orthogonally to the populist cluster, and a low-populism speech by a *Grüne* (Green Party) speaker almost diametrically opposed.

The solution to the cost of human coders at scale remained elusive until the development of large language models (LLMs), which are multi-billion-parameter models trained on large corpora for the purpose of generative text production. Where do LLMs fit into a modern text-as-data workflow? "ChatGPT, explain how large language models can reliably replace human coders for qualitative text judgment tasks. Please specifically cite papers from *Political Analysis* that speak to the efficacy of this approach."

> The core argument for using large language models (LLMs) as replacements for human coders in qualitative text analysis is not that they are "human-like" in a general sense. It is that, for many annotation tasks in political science, they can now reproduce the outputs of trained human coders with reliability comparable to — and sometimes exceeding — inter-coder agreement among humans themselves... The most directly relevant article in *Political Analysis* is "Positioning Political Texts with Large Language Models by Asking and Averaging" by Gaël Le Mens and Aina Gallego (Le Mens and Gallego 2025)
>
> They show several things that matter for replacing human coders in practice:
>
> - LLMs work on short texts and long documents.
> - They generalize across languages.

- They can outperform conventional supervised text-as-data pipelines when the task depends on semantic interpretation rather than keyword frequency.

Though the prose is trite and voiceless, and the citation is slightly inapposite to the claims being made, the broader point is understood. The bulleted list highlights the properties of LLMs that make them most suitable for the task: they are inherently multilingual, they can label texts as well or better than human coders, and they work cheaply at any scale. A variety of somewhat more impartial sources have demonstrated the efficacy of LLMs in a social science research context (Benoit et al. 2025; Bol and Bono 2025; Di Leo et al. 2025; Gilardi, Alizadeh, and Kubli 2023; Heseltine and Hohenberg 2024; Mellon et al. 2024; Ornstein, Blasingame, and Truscott 2025; Wang 2023). Below, I propose "LLM-as-a-Judge", a technique commonly used in frontier AI research to mitigate concerns with LLM-based labeling using, as the label suggests, more LLMs.

The major drawback of LLMs in the present context is that, while they perform well in classification tests, they are notoriously innumerate. There exists no obvious mechanism to extract probabilistic uncertainty from an LLM (Shorinwa et al. 2025); token generation probabilities are no substitute (Toney and Wails 2025); and LLM qualitative self-assessments of certainty are poor (He et al. 2026; Xiong et al. 2024; Ye et al. 2024). Therefore, while LLMs are good at classifying documents, they are unable to produce continuous (or even ordinal) predicted probabilities for binary classification. So, they can be, and here are, used to label training documents, but not to estimate how populist any given text is — or even how likely it is to be strongly populist. Awareness of this limitation separates my approach from other attempts to use LLMs for this purpose (see, e.g., (Jung et al. 2025)).

Thus, a modern text classification methodology consists of learning a general mapping (a predictive model) from text (represented as document embeddings) to concepts (labeled by LLMs). I turn now to the process of actually selecting each of these components.

Training Corpus: To train the model, I need a training corpus. To avoid overfitting, machine learning practitioners typically separate the data used to train (and evaluate) models from the data that the models will eventually predict, and stress the need for ample training data. The best way to do this without consuming the original sample is to find a separate, non-Landtag-based corpus to train on. But this generates a countervailing risk of transfer loss (Weiss, Khoshgoftaar, and Wang 2016), where training data differs systematically from prediction data. Both overfitting and transfer loss are failures of the model to generalize, but they occur for different reasons. In overfitting, the data-generating process for the training and eventual prediction data is the same, but the model is sufficiently flexible relative to the size of the input data that it fits chance patterns in the training data that are absent from the test data. Transfer loss occurs when the feature mappings learned from the training data differ in the prediction data because the two datasets come from different data-generating processes.

I balance these concerns by building a training corpus out of two sources. First, I use a small portion of the Landtag corpus (10,000 speeches), and second, a separate corpus of speeches from the Austrian national parliament.[7] The Austrian corpus represents $29,917$ speeches from 2010–2016.

Compared with training on non-German-language data, training on the Austrian corpus minimizes transfer loss from translation or language differences (though there are some dialect differences between Austrian German and standard German). Austria has a similar political context to Germany. It has active RRP parties (including the FPÖ). Austria is viewed as having a weaker *cordon sanitaire* than Germany, both culturally and politically (see, e.g., Art 2007), and its primary conservative party, the Austrian People's Party (*ÖVP*), once formed a coalition with the *FPÖ*. Remaining transfer loss is mitigated by the inclusion of the diverted Landtag data. No data from the training corpus is used in the eventual empirical analysis.

Data Pipeline: Formally, the problem that I set out to solve is as follows:

7. The Austrian database is an off-the-shelf product, AUSTROPARL (Leonhardt and Blätte, AustroParl Corpus of Parliamentary Debates), and is processed through the polmineR (Blätte 2020) package in R; I perform only trivial cleanup on these speeches.

$$
\begin{aligned}
\text{Given:} \quad & X \in \mathcal{X} && \text{input texts,} \\
& \mathcal{L} && \text{loss function,} \\
& f \in \mathcal{F} && \text{document embeddings,} \\
& g \in \mathcal{G} && \text{estimation models,} \\
& h \in \mathcal{H} && \text{calibration models,} \\
& j \in \mathcal{J} && \text{labelers}
\end{aligned} \tag{1}
$$

$$
(f^*, g^*, h^*, j^*) = \arg\min_{f,g,h,j} \mathcal{L}\big(j(X),\, h(g(f(X)))\big) \tag{2}
$$

$$
\widehat{\Pr}(Y_i = 1) = h^*\Big(g^*\big(f^*(x_i)\big)\Big) \tag{3}
$$

In plain terms, I select a large pool of possible embedding, estimation, calibration, and labeler options. I loop over each combination to determine which minimizes error in a data-driven way. The resulting pipeline is used to generate predicted probabilities for the full corpus, which serve as the dependent variable for the empirical analysis. I proceed by introducing each element in turn.

Document Embeddings ($\mathcal{F}$): I begin by transforming the text of these parliamentary speeches into high-dimensional vector representations using document embedding models. I select six document embedding models using a process described in **Appendix A.2.1**. The embedding model ultimately chosen for the analysis, GEMINI-EMBEDDING-001, uses 768-dimensional embeddings, so the resulting training data for a given model is a matrix of size $(29,917 + 10,000) \times 768$. In **Appendix A.2.2**, I report estimates using the 2^{nd}–5^{th} place models and find the results hold. Each dimension of the embedding is entered as a predictor on the right-hand side of the populism detector model equation.

Training Labels ($\mathcal{J}$): Next, I need to assign labels to each document in the training corpus, which will serve as the dependent variable for the populism detector. I prompt LLMs to classify a text as (radical right) populist or not. LLMs have good internal knowledge of the concepts, but the prompt includes supplementary examples of categories of nativist, anti-elitist, or anti-pluralist speech that are most directly characteristic of RRP, in a manner similar to a rubric for human coders. Given the bipartite definition of radical right populism as having both the "thin" populist ideology and "thick" nativist ideology, the LLM aims to capture *both* dimensions holistically. I use two commercial LLM providers and one local LLM for this task. I describe this stage in **Appendix A.3.2**, including the models selected and the prompt provided to the LLMs. In addition to non-RRP and RRP speech, I ask the LLMs to identify and exclude purely boilerplate procedural speeches (such as a presiding officer recognizing a speaker or adjourning debate) from training.

The LLMs agree 80% of the time, which is comparable to the performance of human coders on similar tasks. Nevertheless, because they sometimes disagree, I also use an "LLM-as-a-Judge" system to perform inter-coder reconciliation: here, a more powerful LLM is given the speech, the rubric, and the decisions made by each of the LLMs, and asked to decide which position is correct, thus producing an authoritative set of labels.[8] LLM-as-a-Judge is now common in frontier AI research (Zheng et al. 2023), for example, modern LLMs *internally* consist of multiple agent models whose outputs are synchronized and adjudicated by a Judge model (Wang et al. 2025), and "debate" between "adversarial" LLMs has been shown to improve reasoning quality (Du et al. 2024; McAleese et al. 2024).[9]

Each of the labelers ($\mathcal{J}$) produces the $(29,917 + 10,000 \times 1)$ binary response vector of the populism detector model equation. Ultimately, the process selects the "LLM-as-a-Judge" model as the optimal labeler, though the paper's main results hold irrespective of this choice (**Appendix A.3.4**).

8. Many researchers are understandably concerned with LLM hallucinations (Huang et al. 2025) as a particularly catastrophic type of error. These are more common when the task is open-ended (e.g., the production of a paragraph of text) than in cases like this article, where the goal is to simply produce a binary class label. Regardless, multi-coder approaches, and thus approaches that synthesize the outputs of multiple coders, mitigate the remaining risk under the assumption that hallucinations, as opposed to conceptual errors in the rubric or close cases, will occur at different times and in different ways across multiple models. Under independence (or orthogonality), the risk of *joint* failure plummets; and even under relatively high levels of dependence (as might be expected from difficult edge cases), it is bounded to be strictly lower than the failure rate of any one model.

9. I add a final non-LLM-based approach as a benchmark for baseline performance in the absence of detailed knowledge of the text: I produce a set of labels via a weak form of distant supervision (Mintz et al. 2009) — using what I already know and see to label documents — wherein I simply classify all speeches by RRP parties as radical right populist speech (and conversely, classify all remaining speeches as non-populist).

Model Selection ($\mathcal{G}$): Next, I train a predictive model to build a mapping from the document embeddings to the dependent variable. The semantic content of the document embedding space is unknown *a priori*, and there is little precedent for which class of models will perform best at the task. Document embeddings were developed to encode the semantic content of *all human text ever created*, so it is true that many (if not most) of the features will be irrelevant to the much narrower task of classifying one dimension of political speech. Given my agnosticism about the functional form or features of interest, I use the data itself to decide. I run a "horse race" in which 25 model classes in $\mathcal{G}$ (described in **Appendix A.4.1**) are fitted against the training data. The models include conventional statistical models (linear probability models, k-nearest neighbors, logistic regression); machine learning models (support vector machines, Elastic Net, XGBOOST, random forests); and neural networks.[10]

Where applicable, I use cross-validation to tune model parameters. The models are trained without reference to the speaker's party or other characteristics. I report details about model performance in **Appendix A.4.2**. The selected optimal model is Neural Network 3, although the results are not sensitive to the choice of model (see **Appendix A.4.3**).

Calibration ($\mathcal{H}$): The models output a predicted probability $\widehat{Pr}(Y_i = 1)$ for each document in the corpus. Models can be highly effective at discriminating between classes (suppose, for instance, a model that assigns every populist speech a higher probability of being populist than every non-populist speech) while also producing relatively poorly fitting probabilities, especially when the underlying data is highly class-imbalanced (Guo et al. 2017; Silva Filho et al. 2023); calibration adjusts output probabilities to correct for this. In the final model, I select the Platt calibrator, though the results do not depend on this choice.[11]

10. Although LLMs do not produce output probabilities, some transformer models ("AI") are designed to simultaneously produce contextual embeddings for input data and learn the mapping of features to outputs (Huang et al. 2020). Those models can theoretically serve as embedder and estimator at the same time. My approach instead partitions the problem into the "representation learning" problem (learning how words relate to one another via the document embeddings) and the "boundary-learning" problem (learning what values of those features translate into the classification label). More sophisticated estimation models do not perform better at the latter problem when the former is already solved (Bengio, Courville, and Vincent 2013; Gorishniy et al. 2021).

11. I omit much discussion of calibration. I test isotonic regression and Platt (Platt 2000) calibrators against

Loss Function ($\mathcal{L}$): Because RRP speech is rare in the corpus, the loss function must take into account the class imbalance. I average three performance metrics: PR-AUC (a metric that rewards the model for precisely estimating *RRP* speech); class-balanced accuracy (a metric that rewards the model for estimating both categories); and Brier score (a metric that captures how well-calibrated the output probabilities are). The choice of f, g, h, and j is only weakly sensitive to the weights placed on each of the loss functions, see **Appendix A.5**.

4.2 Detector Results and Outputs

To summarize the entire pipeline briefly: I prepare the Austrian and Landtag document corpora, use document embedding models to transform the speeches into high-dimensional vectors, and use LLM labelers to label the training data. I test a set of 6 embedding models, 25 estimation models, and 2 calibrators (or none at all) to determine which "black box" is best, according to a blended loss function. The pipeline is low-assumption, in the sense that it does not commit to a document feature set, exact labeler, model class, or calibration approach *ex ante*, and data-driven in the sense that the choices it makes stem from performance on the data rather than researcher discretion. The selected models are Gemini embeddings, LLM-as-a-Judge labels, Neural Network 3, and a Platt calibrator. This combination produces a PR-AUC of 0.816 (on a scale from 0.5–1.0), a class-balanced accuracy of 0.880, and a Brier score of $1 - 0.046$.

Since "ground-truth" here is provided by the LLM labels, these are loss functions that capture the relative compatibility of the embedding and model with learning the labels, not "correctness" (but human-coded documents also suffer from this weakness). Thus, the next step requires that I validate the detector.

Validation: My aim is to ensure that the model can detect RRP speech at a sufficient level of accuracy to answer the substantive question of interest. I briefly consider the question of

the uncalibrated probabilities; in the final model, Platt calibration is selected. In the selected model, the average populist speech is assigned a probability of 0.608 of being populist, and the average non-populist speech is assigned a probability of 0.114 of being populist, indicating relatively high separation.

measure validity. No formal methodology exists to "prove" the content or construct validity of a measure (see psychometric research on this point, including Messick 1987, 1995), and both document embeddings and neural networks are opaque to interpret. Nevertheless, I persist. In the main text here, I present three simple validation checks:

First, do the parties coded as having the most populist speech demonstrate criterion and construct validity by matching external assessments of RRP parties? In **Table 1**, I show that this is the case. The RRP parties (so-defined by ParlGov, PopuList, the Chapel Hill Expert Survey, and the Manifesto Project) appear at the top, and the orders continue in the expected manner. Minor parties omitted from the table also load according to expectations.

Austria					Germany				
Party	n	Mean	$\texttt{Pop}_{75\%ile}$	$\texttt{Pop}_{20\%}$	Party	n	Mean	$\texttt{Pop}_{75\%ile}$	$\texttt{Pop}_{20\%}$
Stronach	1,558	0.442	59.0%	61.4%	NPD	4,576	0.549	81.0%	70.5%
BZÖ	2,041	0.418	54.0%	56.4%	AfD	48,492	0.335	63.2%	47.8%
FPÖ	6,062	0.333	45.1%	47.1%	Linke	60,348	0.057	22.8%	9.1%
NEOS	1,103	0.198	30.6%	33.1%	FDP	50,776	0.050	21.4%	7.9%
GRÜNE	3,522	0.154	22.5%	24.8%	CDU/CSU	96,646	0.050	20.5%	7.9%
ÖVP	7,354	0.075	10.2%	12.5%	Grune	83,253	0.039	16.5%	5.9%
SPÖ	8,102	0.071	9.9%	11.8%	SPD	88,233	0.038	16.1%	5.7%

Table 1: Mean Predicted Populism by Party, Face Validity Check: This table depicts the major parties in the Austrian and German datasets sorted in descending order of predicted populism. Parties in red, identified *ex ante* as being RRP parties, appear at the top. $\texttt{Pop}_{75\%ile}$ is the proportion of speeches by the party that have populism predictions beyond the 75th percentile of the country-wide distribution; $\texttt{Pop}_{20\%}$ is the proportion of speeches whose predicted probability of populism is at least 20%.

Despite low mean populism and few non-RRP speakers whose mean populism is above any given threshold, 67.9% of speakers give at least one medium-populism ($p \geqslant 0.2$) speech, and 27.2% give at least one high-populism speech; within-speaker heterogeneity seems plausible and consistent with expectations. *Die Linke*, a left-populist party, appears below the RRP parties, but the SPD, a social democratic party, appears lowest of all, indicating that the detector detects both *radical right* and *populist* speech and does not inherently confused speech of the political left with populism.

Second, are the topics discussed in high-populism speeches consistent with expectations about the substantive content of RRP? If the populism detector is doing its job, then the text it identifies as being high-populism should relate to the subjects and express the themes expected from the concept being measured (i.e., demonstrate content validity). I fit a structural topic model (Roberts et al. 2013), constraining the number of topics to $k = 10$ to facilitate the presentation of this diagnostic. In **Figure 3**, I show that this is also the case: both the topics discussed and the words that characterize the topics are consistent with expectations.

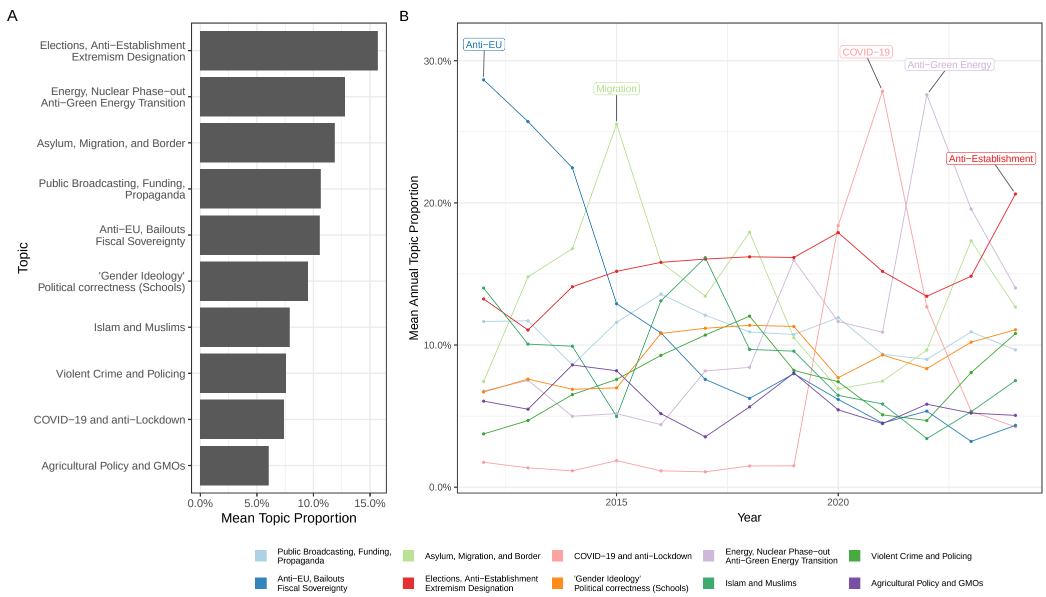


Figure 3: High-Populism Topics: This figure is a face validity check of the coding of high-populism speeches by the populism detector, presenting the results of fitting a structural topic model with $k = 10$ topics to a sample of 8,000 high-populism speeches drawn from the Austrian and German corpora. **Panel A**: Indicates the mean cross-sectional topic prevalence across the sample. **Panel B**: Presents the variation in mean topic prevalence over time.

Finally, does the detector measure *all* of what we would expect radical right populism to encode. Because the individual LLMs and the LLM-as-a-Judge provide brief reasoning for coding any speech a “1” (populist), I can investigate how those reasons concord with the definitions of “populism” and “radical right” adopted in Section 2. In brief, I write rules to map the free-text decisions made by the LLMs to elements of the definitions. Every training observation coded as 1 maps to at least one category. The categories are rough, non-exclusive, and I place no particular stock in the rank order; the goal of this validation exercise is to establish that each of the primary definitional components of "populism" and "radical right" is featured. The results are displayed in **Table 2**.

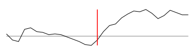
Reason	N	%
Anti-Elite	4,774	77.8%
Manichean (Good vs. Evil)	3,183	51.9%
Nativism/Xenophobia	2,754	44.9%
Nationalism	2,010	32.8%
Anti-Pluralism	360	5.9%
Rejection of (Electoral) Democracy	316	5.1%
Political Authoritarianism	163	2.7%

Table 2: Populism Coding Reasons in Training Data: This table summarizes why speeches in the training corpus were coded as populist ($N = 6,137$). I extract and map the LLM coding reasons to one or more definitional elements of "populism" or "radical right". Categories are non-exclusive and sum to more than 100%.

In the appendices, I continue the validation exercise by presenting clipped examples of high-populism speeches from the AfD and other parties (**Appendix A.6.1**) and the top keywords associated with high-populism speeches (**Appendix A.6.2**). All results acquit the detector well. Having developed the measure and predicted the extent of populist speech in the corpus of interest, I can finally turn to answering the paper's motivating question: what is the effect of the entry of the AfD into German state legislatures on mainstream political parties?

5 Identification and Estimation

A politician might decide to be more populist in her rhetoric for any number of demand-side reasons: she might undergo a personal evolution on a policy matter; she might respond to changes in the salience of various issues; she might respond directly to feedback from her electorate or selectorate; she might fear losing re-election. In this article, I aim to identify supply-side effects, which emerge only from the *entry* and *presence* of RRP parties in legislatures.

When the AfD formed, it polled at or near the electoral threshold nationally and in many states. But its election to each of the legislatures became inevitable when real-world events, including the apex of the Syrian refugee crisis in Germany, intervened, abruptly driving AfD vote intention up.  Ever since, the AfD has remained well above the electoral threshold. Landtag elections are held on a fixed and staggered schedule (and no states held snap or early elections during the AfD's rise[12]). Thus, the entry date of the AfD on a state-by-state basis was often substantially delayed, with wide variation by state, in a way that is orthogonal to local temporal variations in demand. The first treatment of states varied from October 2014 (Thuringia) — before the uptick — to January 2019 (Hesse). My identification strategy here is based on staggered adoption of treatment and conditionally exogenous treatment timing.

As noted above, I operationalize treatment as a continuous dose in my primary specifications. The intensity of treatment varies by state: in states such as Schleswig-Holstein and Bremen, the AfD barely crossed the electoral threshold (5.5% and 5.9% vote shares, respectively, in their first election). In others, the AfD received much larger vote shares (e.g., 20.8% in Mecklenburg-Vorpommern). Vote shares correlate nearly 1:1 with seat shares because all states use proportional representation systems where seats are distributed proportionally after the elimination of minor parties that do not clear the electoral threshold. All models in the paper use seat shares as the treatment variable because that is how institutional resources are allocated. I show treatment entry timelines and seat shares in **Figure 4**.[13]

12. There is equally no reason to believe that state governing coalitions conspired to defer any would-be elections to avoid an incoming "wave" — of all the governing coalitions after the AfD's emergence and before their entry, the most likely candidate would be Schleswig-Holstein's Albig cabinet (2012–2017), which governed with a single seat majority. But by all accounts, this was a harmonious government (Die Welt 2017) with minor-party support from the six Pirate Party legislators, so it would be difficult to construct a story where unlikely bedfellows stuck together to defer AfD entry.

13. The AfD did not run in Bremen in 2023 because an internal administrative dispute led the party to submit two competing candidate slates, resulting in a state court disqualifying both slates.

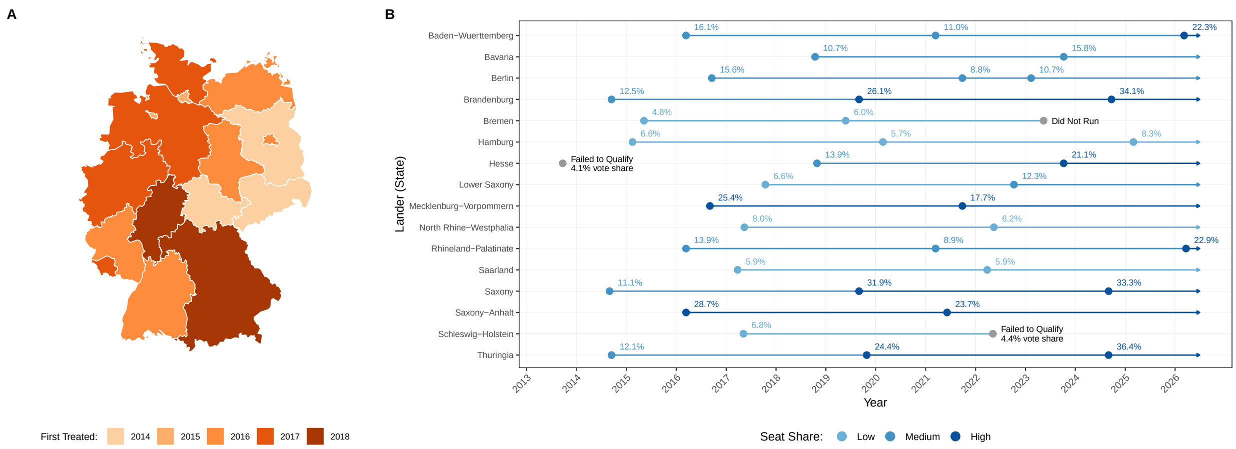


Figure 4: Treatment Timing and Dose: This figure shows the treatment timing and dose for each Länder (State). **Panel A:** A choropleth map of Germany shows variation in the first treatment period. A chance geographic pattern in election timing (i.e., states in the northeast holding elections in 2014) presents an identification concern. **Panel B:** A plot of the treatment and dose history of each state. Once the AfD begins being elected in 2014, it enters each Landtag at their next election. Some states have consistently low (high) treatment, while others vary materially over time.

Consider a simple model of the world where there are two imaginary states that border each other, have similar economies and resident demographics, and share a local media market. Each state experiences national and international political events simultaneously. Even regional events are experienced by both states simultaneously. By construction, demand-side factors for populist speech should move politicians in both states in the same direction. Now, suppose that an election occurs in one state but is years away in the other. The direct treatment, the AfD's actual entry into the legislature, applies only to the treated state. Thus, the change in trajectory of the treated state compared to the non-treated state is identified.

In the real world, although nature was kind enough to exogenously allocate election timing, the *observed* allocation was cruel. **Figure 4** shows a clear geographical pattern in electoral timing cohorts: several of the eastern states go first. Chiefly, the pattern emerged because the former *DDR* (East Germany, GDR in English) states began holding elections at the same time, shortly after German reunification, and in the absence of any early elections, their election timing has remained nearly synchronous since. Though the cause of the timing order effect is unrelated to my question of interest, this chance correlation is a problem if these states share characteristics that also predict the potential outcomes of the states — and they may well, since rank-order election timing negatively correlates with AfD vote share in the 2013 *Bundestag* election and *NPD* vote share in 2005. My design corrects for this by identifying those economic, cultural, and geographic characteristics that potentially predict election timing and radical right populism in speech. I discuss the literature on demand for RRP Parties in **Appendix A.7.1**. Inclusion or exclusion of these variables do not drive results.

A greater concern is spillover. States are treated by the entry of the AfD, but the same election that causes AfD entry is itself a treatment signal: demand for RRP policies that was previously only latent has been made manifest by the election. Since the untreated state(s) can surely observe election results, this constitutes a potential SUTVA violation due to treatment spillover. I do not pretend to know whether the indirect and direct effects would point in the same direction, so instead I model the most likely pathways for spillover, which follow measurable geographic or cultural channels.

Finally, *when* does the effect of AfD entry become realized after treatment? In a vacuum, the time cadence of any treatment is unknown. It may be the case that parties would adapt to the new normal immediately. However, it may also be the case that the AfD takes time to learn and use parliamentary processes, and checking for immediate results would yield none. Previous research suggests that legislators took time to understand how the AfD would participate in debates and how to respond to them (Heinze 2022). There may also be policy cycle considerations (Franzese Jr 2002) (e.g., the types of issues debated and the manner of speaking about them at the beginning of a parliamentary session may differ systematically from those debated later in terms of their tendency to yield RRP speech). This cautions against regression discontinuity designs, which rely on comparing speeches immediately before and after the election to identify an immediate treatment effect.

Driven by these concerns, my estimation approach uses the generalized form of the difference-in-differences technique via the canonical two-way fixed effects (TWFE) estimator:

$$\widehat{Y}_{i,t} = \tau\, D_{j[i],t} + \beta\, \mathbf{X}_{it} + \gamma_j + \delta_t + \epsilon_{it} \tag{4}$$

where $\widehat{Y}$ is the dependent variable (the probability that the speech is populist, according to the populism detector), D is the treatment status, $\mathbf{X}$ are confounders, γ is a unit fixed effect, δ is a time fixed effect, and ϵ is an error term; and where i indexes a speech, j a state, $j[i]$ the state in which the speech takes place, and t time. D_{it} is a continuous treatment equal to the proportion of seats earned by the AfD in the controlling election in the state. The confounder term, $\mathbf{X}_{it}$ includes a variety of state-year confounders (population, state-level GDP, state-level unemployment, sectoral employment, education rates, local foreign-born population[14], local crime, and local poverty rate). These variables are available from Eurostat and German government sources for the relevant time period (see **Appendix A.7.2**).[15] τ is, thus, the coefficient of interest, representing the effect of treatment on the likelihood that a speech is populist. The fixed effect structure excludes the variation across states and the variation across specific periods in time, allowing me to capture the variation attributable to treatment.[16]

In addition to the primary TWFE model, I further present results from a stacked ("event study") treatment model. This model includes lead- and lag-treatment terms, which allow me to rule out anticipatory effects *and* examine the time course of the ultimate treatment effects. Formally:

14. Sub-national migration data is unreliable during much of the analysis time period, as is data on the non-EU foreign-born population, so the foreign-born population proportion is the most reasonable proxy for exposure to the Syrian refugee crisis.

15. It is arithmetically non-obvious how speech *length* translates to embedding geometry or the underlying LLM labeling, so out of an abundance of caution, I also control for logged speech length to ensure that any relationship between verbosity and classification is not a confounder.

16. Results are invariant to the choice of time fixed effects (year, quarter, week); to the inclusion or exclusion of unit or time fixed effects; to the inclusion of party or individual fixed effects; and to the inclusion of state-level or region-level time trends; see **Appendix A.9.4**.

$$\widehat{Y}_{i,t} = \sum_{k=-w}^{w} \tau_k \, D_{j[i],t+k} + \beta \, \mathbf{X}_{it} + \gamma_j + \delta_t + \epsilon_{it} \tag{5}$$

where k denotes the number of lag (lead) periods since (until) treatment in the reference observation, w is the size of the window, and τ_k is now a vector of coefficients.

A recent literature suggests that TWFE estimators can produce biased estimates under treatment effect heterogeneity by cohort or unit due to the well-known implicit "negative weight" problem (De Chaisemartin and d'Haultfoeuille 2023; Goodman-Bacon 2021; Imai and Kim 2021). Proposed replacement estimators solve these known problems but typically rely on more onerous identification assumptions; in addition, most require binary treatments. For a recent overview of the tradeoffs associated with the use of more modern HTE estimators; see Chiu et al. (2026). In **Appendix A.8**, I repeat the core analysis and event study using the modern Counterfactual estimator (Liu, Wang, and Xu 2024) (with a binary treatment) and find the results are substantively unchanged; however, owing to reduced power, treatments at $t+3,4$ lose significance.

6 Results

Primary Results: I begin by running the primary TWFE analysis, first across all non-RRP parties, next subsetting to just parties that are part of a state's government (or opposition), and finally subsetting to individual parties. I report the results of these models in **Table 3**. The results are positive, significant, and moderately large. The primary all-party effect coefficient is 0.097 (9.7 percentage points); both government and opposition parties are positive and significant (indicating that the effect is not driven solely by changes in the discursive mode opposition parties use to criticize government); and four of the five major parties show positive and significant effects. To the extent prior theory predicts adaptation solely by parties of the center-right and right, my results refute this: mainstream left parties show higher RRP in response to RRP entry, and the largest effect is among the already-populist left-wing party *Die Linke*.

Dependent Variable:	Predicted Populism							
	All Parties	All Gov.	All Opp.	CDU/CSU	SPD	Linke	Grüne	FDP
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
AfD Intensity	0.097***	0.077***	0.146***	0.068***	0.085***	0.127***	0.076***	0.007
	(0.006)	(0.007)	(0.009)	(0.011)	(0.010)	(0.014)	(0.011)	(0.025)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects								
State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	393,951	178,464	214,008	96,646	88,233	60,348	83,253	50,776

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table 3: Effect of AfD Entry on Parliamentary Speech, by Party: This table depicts the paper's primary workhorse model: a 2WFE estimator which calculates the effect of AfD entry into German Landtage on the propensity for populist speech in non-AfD parties. Treatments are continuous, with dose equal to the proportion of votes/seats earned by the AfD. Column (1) reports the effect across all political parties; columns (2)–(3) subset to speakers whose parties are in governing coalitions and those whose parties are not; column (4) restricts to the conservative (but non-populist) CDU/CSU; and columns (5)–(8) report party-specific subsamples for the SPD, Die Linke, Bündnis 90/Die Grünen (Grüne), and FDP. Other than the FDP, a right-liberal party, the estimates are positive, moderately large, and significant across groups.

The direct effect proposed above can be driven either by *replacement* (where low-RRP incumbents retire or are defeated by higher-RRP replacements) or *adaptation* (where the same people change their speech patterns), or both. The presence of replacement would make (some of) the effect a mechanistic translation of election results, weakening my claims. The proof of the pudding is in the eating. I add legislator fixed effects to capture within-legislator variation in treatment. The results attenuate slightly but remain substantively unchanged, also in **Appendix A.9.4**.[17]

17. I omit much further analysis about how elections might endogenize incumbent characteristics, but every state uses a variant of mixed-member proportional representation, such that parties have significant ability to rescue high-value incumbents by adjusting their (mostly) closed party list rank.

The analysis continues by interpreting the magnitude and reach of this result, estimating a dynamic (time-varying) treatment effect, exploring the risk of spillover, implementing placebo tests, and finally discussing the overall robustness of the results.

Interpreting Result Magnitude: The model's coefficient requires two important pieces of contextual information to interpret: the range of treatment doses in the data and the base rate of populist speech in the data. Mathematically, I report the feasible change in the rate of populism as $(\bar{y} + \tau \times s_{90})/\bar{y}$ where $\bar{y}$ is the base rate, τ the observed treatment effect, and s_{90} the 90^{th} percentile observed treatment for scaling purposes. I chose the 90th percentile as a large but feasible treatment effect. Scaling is important because the regression coefficient itself characterizes an impossible quantity (the effect of moving from no AfD presence to all elected legislators being members of the AfD). This is impossible not only in practical terms (the highest individual state performance of any party in German history is the Bavarian CSU in 2003 at 69%), but also numerically (since I estimate on the subsample of speeches *excluding* AfD speeches, the sample would be empty if the AfD did sweep all the seats).

Instead, the AfD — when it clears the minimum electoral threshold — receives a seat share varying from 4.8% to 36.4% in the sample. The 90^{th} percentile treatment value is a 29.7% seat share, so the coefficient size is reduced by not quite a factor of four. The shrunken coefficient size is counterbalanced by the fact that there is a low overall rate of populist speech in the base case. In the all-parties subsample, the average predicted probability of a pre-treatment speech is 4.8%, hence, an adjusted coefficient size of 2.8 percentage points increases the propensity for populist speech by 61% relative to the baseline, which I interpret as moderately large — not nearly enough to make the CDU after treatment sound like the AfD, but large enough to make them sound more populist than the left-populist *Die Linke* did at baseline; or large enough to make Brandenburg (middle of the pack) sound more populist than Berlin (3rd most-populist) did at baseline.[18]

18. Although treatment is heterogeneous by party and state, it is inelastic enough that rank-order populism levels are generally quite stable in pre-/post-comparisons, more a “tug-of-war” than a “leapfrog”.

Speech Affected: Speech could become more populist because the salience of topics changes in favor of topics that generate more populist speech (agenda-setting on the part of the RRP party), or it could become more populist because speech within a given topic becomes more populist (accommodation). I test the mechanism by unpacking how the treatment effect varies across speeches by topic. To do so, I fit a structural topic model with $k = 25$ topics to the entire Landtag corpus, excluding RRP parties (n.b. this is a separate topic model from the high-populism-only topic model presented above to validate the detection of populism in speech).[19] Because the model is fitted excluding AfD speeches, the topics themselves are not responsive to RRP framing except insofar as other parties are responsive to their framing.

I then partition the corpus by *modal* topic and produce 25 topic-level treatment effects. This has the attractive, though conservative, quality of disentangling topic-mix effects from within-topic effects, since each model only compares speeches that are alike in topic. 17 of the 25 topics capture substantive policy, while seven capture largely procedural dimensions of speech (one is a so-called junk topic; see AlSumait et al. 2009). I plot the effect size coefficients for the substantive topics in **Figure 5**.[20] To further characterize how populism is picked up in these topics, I sample low-, medium-, and high-populism speeches from two of the topics ("Economy, Taxes, Banks" and "Energy, Climate") and reproduce them in **Appendix A.6.4**. Although I would caution against over-interpretation of any one of these effects, given the substantially reduced model power and the vicissitudes of topic labeling in general, there is a positive effect across a wide variety of topics. The topics that might be expected to drive the results (including crime and migration) are actually modest in size and in some cases non-significant.

Two disclaimers are in order. First, the test here is deliberately calibrated to avoid

19. The goal of this exercise is not to properly label the exact contours of Landtag speech; it is to get a sense of whether the effect is being uniquely driven by narrow types of speeches. I omit the intensive descriptive diagnostics typical of topic analyses for this reason, but the result (a large majority of topics having positive and significant coefficients) holds for $k = 10$ and $k = 50$ topics as well.

20. This plot suppresses a topic centered around Low German, regional minorities, and language politics because it has a high coefficient and an extremely wide confidence interval crossing zero due to very few speeches, as it comprised $< 0.3\%$ of the corpus.

finding the known effect that RRP entry could drive the *salience* (relative topic mix) of RRP pet issues; the question is whether RRP presence makes individual speeches more populist. Second, a null or weak result on crime and migration does not conclude that these topics are not constitutive of RRP ideology or not discovered by the model; rather, the impact of RRP presence is not as strongly felt in speeches on those topics by non-RRP parties.

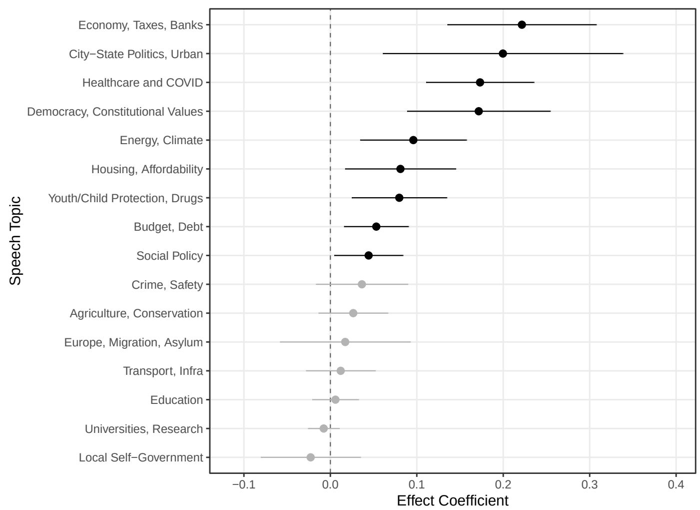


Figure 5: Effect Heterogeneity by Speech Topic Subsample: This effect plot shows the result of running the primary model on subsamples for each speech modal topic. Although these models are somewhat underpowered by nature, the results show both that positive effects are found across a wide variety of topics, and topics which might reasonably be expected to be driving the effects in total actually have modest and in some cases non-significant coefficients.

Event Study and Treatment Cadence: Next, I estimate the time cadence of the treatment effect by including lag- and lead-treatment terms in the regression. I take the immediate pre-treatment period as a reference and look at how the propensity for populist speech varies in the periods leading up to the reference period and in the treated periods following the reference period. Anticipatory effects are weakly negative or null, while the treatment effect is positive throughout and grows in the mid–late parliamentary term. I present the results graphically in **Figure 6**. These results are consistent with either a policy cycle (in terms of either rhetoric or the substantive content of debated bills) or with there being an on-boarding period during which the AfD is still learning to use parliamentary institutions to extract accommodation responses.

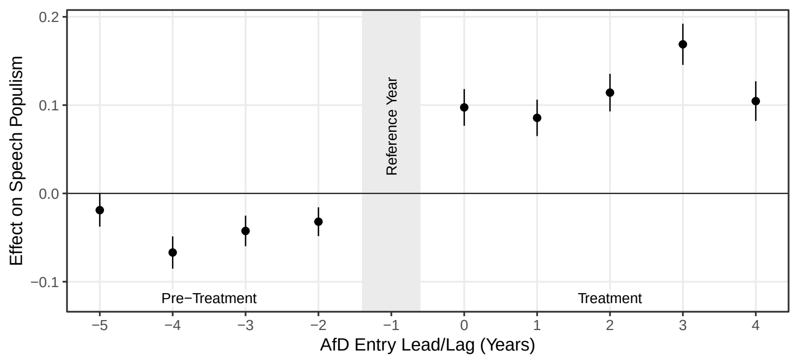


Figure 6: Event Study: This effect plot shows the result of running the primary model with stacked lead/lag treatment terms to discover the trajectory of the treatment. In both panels, pre-period years show a slightly negative and inconsistently significant effect compared with the reference year ($t = -1$ relative to treatment). The treatment effect is large and positive across the treatment period.

Spillover: As noted above, a primary concern is that the effects might be biased by the presence of spillover from the indirect portion of the treatment. To the extent that these spillovers are experienced in a nationally uniform way (i.e., the first election of the AfD treats the entire country), the time fixed effect design absorbs this. But sub-national spillover remains a threat. I consider spillover by almost every taxonomy of geographic region (east–west; *Mainlinie* [north–south]; by nearest major metropolitan area; by cardinal region; and by broadcast media market[21]). In each case, I consider a state spillover-treated when any state in its spillover group is treated. I also model spillover by whether the state is majority Catholic (four states are) as a crude surrogate for cultural connections not captured by geography. Finally, I model border-proximity-based spillover in two different ways: one where a state is spillover-treated if any bordering state is treated, and one where a state is spillover-treated by an average of the treatment status of other states, weighted by the pairwise inverse distance

21. Germany has nine regional public broadcasters that operate independently as the *ARD* consortium. These predate the later formation of a national broadcaster, *ZDF*. Regional broadcasters share some content but also produce their own extra-regional and national content. This estimate is intended to proxy for the states most likely to share a broader media environment with the reference state.

between the states. In **Table 4**, I present the magnitude of the direct effect and spillover effects for each channel and find that the results are not sensitive whatsoever to these spillover violations; a few spillover channels are statistically significant, but none are substantively significant.

Dependent Variable:	Predicted Populism							
Spillover	Catholic	West/GDR	Fed. Region	*Mainlinie*/South	Media Market	Metro Region	Contiguous?	Inv. Dist
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
AfD Intensity	0.097***	0.097***	0.097***	0.096***	0.094***	0.099***	0.093***	0.094***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Spillover	0.000	-0.003	0.000	0.002	0.002*	-0.004**	-0.007***	-0.014**
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects								
State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	393,951	393,951	393,951	393,951	393,951	393,951	393,951	393,951

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05, .: 0.1*

Table 4: Results with Spillover Modelled: This table repeats the primary all-party analysis while allowing for explicitly modeled spillover between states. The first six columns model spillover model states as members of cultural or geographic groupings; states receive spillover treatment when any state in their grouping is treated. The remaining two columns model spillover spatially: in model (7), states are treated if any bordering (contiguous) state is treated. In the final column, states are given a weighted spillover based on distance to all treated states. In each case, the primary effect survives the spillover treatment unaffected. Spillover effects are consistently small, often non-significant, and sometimes negative, lending support to the claim that spillover is not a threat to inferential validity.

Placebo Testing: I use permutation inference to examine how the observed results stack up against various proposed “placebo” treatments. In each analysis, I keep the actual distribution of populist speech, control variables, and fixed effects but draw a new treatment vector to mechanically break the link between treatment and outcome. I fit the model across many draws and assess the distribution of τ. If the observed coefficient is in the extreme right tail of the distribution, the findings are bolstered because they are unlikely to have emerged through chance correlations of the treatment distribution and outcomes. I test four different placebos: (1) adding or subtracting a random time increment to each state’s first date of treatment; (2) randomly permuting which doses are assigned to which state-elections; (3) randomly permuting which treatment dose and timing are assigned to which state; and (4) adding (or subtracting) a random time increment to each state's time course and a random dose scaling factor to each state's dosage. The four placebos combined manipulate all of the information contained in the actual treatment assignment process.

In **Figure 7**, I show the sampling distribution of these tests. All four reject the null hypothesis, indicating that the observed effect is unlikely to have emerged from a chance alignment between treatment timing or dosage and outcome trends.[22]

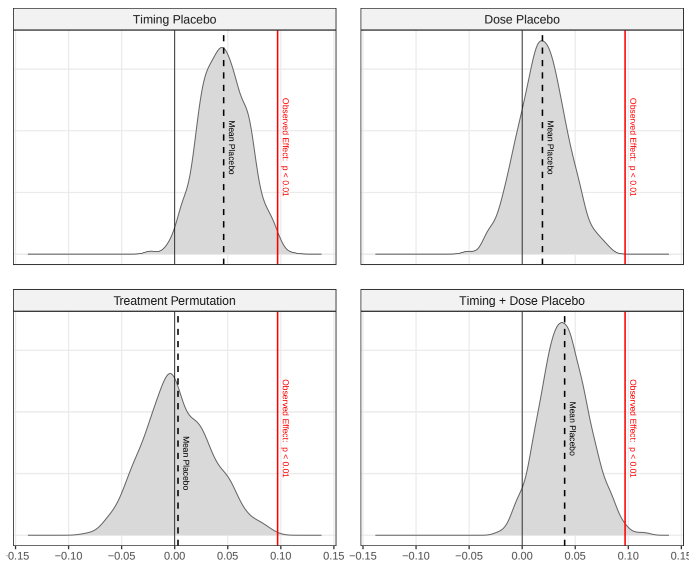


Figure 7: Placebo Test: This figure presents the results of four permutation-based placebo tests, all of which show that the observed effect is an outlier in size. In order, the tests: (1) randomly shift the treatment timing of each state; (2) randomly permute the doses assigned to each state-election; (3) randomly reassign a state's entire treatment and dose trajectory to another state; (4) randomly adjust each state's treatment timing and randomly scale the state's dose. Each test's sampling distribution consists of 1,000 permutations.

Robustness: The results described above do not depend on the choices made in constructing the populism detector. Alternative embeddings (**Appendix A.2.2**), labelers (**Appendix A.3.4**),

22. Three of the four placebo distributions have non-zero centers. This is not inherently a problem (the observed result still lies in the tail of the distribution). In a panel setting, timing placebos are sometimes non-zero centered because the placebo-treated units can absorb some of the pre-treatment trend, particularly under unit heterogeneity. Likewise, the dose placebo disrupts the dose-response curve, but not the distinction between untreated and treated units.

estimation models (**Appendix A.4.3**), and loss functions (**Appendix A.5**) all yield comparable results.

Nor are they driven by the construction of the treatment or outcome variables. In **Appendix A.9.1**, I test binary (effectively, AfD entry) and binned (low, medium, high presence) treatment variables, and in **Appendix A.9.2**, I binarize the dependent variable at various thresholds. In both situations, the outcomes are similar. They are further invariant to dropping the state-year confounders or multiply-imputing missing values instead of interpolating them (**Appendix A.7.3**). The results obtain across states. In **Appendix A.9.3**, I show that dropping any one state, or historical West or East Germany, has no effect on the all-party outcomes.

In **Appendix A.9.4**, I examine alternative fixed effect structures (state + quarter, state + week, state + year + party, state-year trend, state + year + region, and state + year + region-year trend) and find that the results are not sensitive to these either.

In summary, I find that the AfD's entry into and presence in German state legislatures has made the remaining parties moderately more populist in response, and this is true across a wide variety of measurement and estimation approaches.

7 Discussion

This article presents three contributions to political science. First, it develops a methodology for the construction of general-purpose text classifiers using document embeddings, LLM-based labeling (including inter-coder reconciliation using LLM-as-a-Judge), and automated model selection, and it implements this methodology to develop a German-language radical right populism detector. Second, it constructs a corpus of German state Landtag speeches, cleaned and annotated with detailed speaker and party information. Finally, it offers an empirical exploration of the effects of exogenous populist entry into legislatures on legislative speech, net of demand effects, which shows strong and consistently positive results.

One limitation worth exploring is the question of external validity. Is the populism detector externally valid, and to what other settings would it transfer? The methodology is transferable wholesale. Multilingual models perform well in non-English (or, with reference to my case, non-German) settings. Direct translation pairs (bi-texts) produce extremely similar spatial embeddings (Feng et al. 2022). Training corpora for the text models are, by dint of being derived from text on the internet, skewed towards English, but language representation is strong in Western European languages, Slavic languages, and Chinese, and weakest in African and South Asian languages (Joshi et al. 2020; Wenzek et al. 2020; Xu et al. 2025). Among comparably situated "high-resource" languages, performance varies only by a matter of degree: on the popular Mmteb multilingual benchmark, the non-English performance of the Gemini embeddings selected here on a hundred-point scale varied between 82 points for Arabic and 90–92 points for most Romance languages (Enevoldsen et al. 2025). Similar populism detectors could be trained in other settings using the same embedders and labelers. Transferring the *exact* German-trained model to other linguistic settings is also theoretically possible — earlier work by Licht (2023) suggests good "zero-shot" cross-lingual text classification, though transfer loss is a problem; thus, using auxiliary training corpora in the target language would be required to achieve comparable accuracy.

Second, do I expect the mainstream reaction to populist entry to hold across national settings? The present analysis depends on unexamined cultural and institutional factors: the contours of institutions such as campaign rules (Rahat and Hazan 2001; Scarrow 2007; Stevenson and Vavreck 2000), media environments (Hallin and Mancini 2004), the capacity for agenda control (Bräuninger and Debus 2009), coalition structures (Geys, Heyndels, and Vermeir 2006), incentives to cultivate a personal vote (Carey and Shugart 1995), cultural demography, and more. But taking the AfD as *primus inter pares*, or simply as a vanguard among RRP parties in Western democracies, I observe that discursive politics *can* be hijacked, even in a country with strong institutional, cultural, and legal norms against extremism. This hijacking affects even the speech of parties that ostensibly face no direct spatial competition with an RRP party, and even speech on issues not traditionally "owned" by RRP parties. My research suggests that political scientists should take parliamentary speech, and not just voting or electioneering, seriously as a venue for downstream RRP impact.

The findings complicate a narrative of uni-dimensional ideological adaptation in response to electoral threats. Standard voter spatial utility assumptions predict that only right-wing incumbents should be competing with the far-right for votes (irrespective of the underlying allocation of voter preferences).[23] The results here are consistent with RRP as a cross-cutting political force that risks destabilizing all incumbents, not just the right (as per Lucardie 2008; March 2017; Mudde 2007; Mudde and Kaltwasser 2013; Stanley 2008).

My results go further. Most of the signal from electoral threat is visible in advance, which is why much of the literature on RRP effects studies campaigns. Demand signals such as elections in neighboring constituencies and polls should cause parties and candidates to head off the insurgent RRP parties at the pass by adapting in advance. The findings do not bear this out in the legislatures. Pre-election adaptation is weak, the adaptation that does occur is not directionally consistent with post-election adaptation, and I observe no post-election spillover in untreated states that are the most vulnerable to spillover. The striking absence of expected findings suggests that the adaptation is being driven by activity in the legislature, not just electoral considerations.

Finally, more work needs to be done to unpack exactly what it means for speech to adapt. Traditional approaches measure the similarity between reference speeches (or parties) and comparison speeches, assuming that the reference text captures populism. My measure improves on these by isolating those characteristics of speech that are demonstrably populist. Still, speech has an ineffable quality. What does it mean for your speech to be 61% more populist than your baseline (or 3 percentage points more populist overall)? What does it mean that 65% of non-RRP speakers in top-tercile treated state-periods show their average

23. Under standard spatial assumptions and most non-degenerate distributions of voter utilities, left-wing incumbents could even move *left* in response to right-wing adaptation without loss of vote share and gain utility due to their own spatial preference.

speech populism increase? Is speech contagion just a memetic tic? Does it mean a coarsening of tone or demeanor, an increasing Manichean framing of political stakes, or real shifts in how legislators engage with issues? Will speech become unglued from voting and policy proposals, or will it be the harbinger of changes in those measures as well? Much more must be done to concretize the stakes, but in this article, I have demonstrated at minimum that my measure captures more than just tone and that the results are operative across policy areas.

The model identifies legislative speech adaptation and accommodation solely attributable to the *entry* of the AfD and is net of prior public demand. I find that this effect is positive, significant, moderately large, and visible across virtually all specifications of interest, including across parties and legislative themes. If there is any poison in the air, it arrives with the AfD.

References

Abou-Chadi, Tarik, and Werner Krause. 2020. "The Causal Effect of Radical Right Success on Mainstream Parties' Policy Positions: A Regression Discontinuity Approach." *British Journal of Political Science* 50 (3): 829–847. https://doi.org/10.1017/s0007123418000029.

Abou-Chadi, Tarik, and Thomas Kurer. 2021. "Economic Risk within the Household and Voting for the Radical Right." *World Politics* 73 (3): 482–511. https://doi.org/10.1017/s0043887121000046.

Akkerman, Agnes, Cas Mudde, and Andrej Zaslove. 2014. "How Populist Are the People? Measuring Populist Attitudes in Voters." *Comparative Political Studies* 47 (9): 1324–1353. https://doi.org/10.1177/0010414013512600.

Akkerman, Tjitske. 2012. "Comparing radical right parties in government: Immigration and integration policies in nine countries (1996–2010)." *West European Politics* 35 (3): 511–529. https://doi.org/10.1080/01402382.2012.665738.

Alonso, Sonia, and Sara Claro da Fonseca. 2012. "Immigration, left and right." *Party Politics* 18 (6): 865–884. https://doi.org/10.1177/1354068810393265.

Alrababah, Ala, Andreas Beerli, Dominik Hangartner, and Dalston Ward. 2025. "The Free Movement of People and the Success of Far-Right Parties: Evidence from Switzerland's Border Liberalization." *American Political Science Review* 119 (3): 1426–1445. https://doi.org/10.1017/s0003055424001151.

AlSumait, Loulwah, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. "Topic significance ranking of LDA generative models." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* 67–82.

Altemeyer, Bob. 1981. *Right-wing authoritarianism.* Univ. of Manitoba Press.

Andersen, Robert, and Anna Zimdars. 2003. "Class, education and extreme party support in Germany, 1991–98." *German Politics* 12 (2): 1–23. https://doi.org/10.1080/09644000412331307564.

Anduiza, Eva, and Guillem Rico. 2024. "Sexism and the Far-Right Vote: The Individual Dynamics of Gender Backlash." *American Journal of Political Science* 68 (2): 478–493. https://doi.org/10.1111/ajps.12759.

Araujo, Matheus, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. 2016. "An evaluation of machine translation for multilingual sentence-level sentiment analysis." In *Proceedings of the 31st annual ACM symposium on applied computing,* 1140–1145.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma. 2017. "A simple but tough-to-beat baseline for sentence embeddings." In *International Conference on Learning Representations.*

Art, David. 2007. "Reacting to the Radical Right: Lessons from Germany and Austria." *Party Politics* 13 (3): 331–349. https://doi.org/10.1177/1354068807075939.

Arzheimer, Kai. 2009. "Contextual Factors and the Extreme Right Vote in Western Europe, 1980–2002." *American Journal of Political Science* 53 (2): 259–275. https://doi.org/10.1111/j.1540-5907.2009.00369.x.

———. 2015. "The AfD: Finally a Successful Right-Wing Populist Eurosceptic Party for Germany?" *West European Politics* 38 (3): 535–556. https://doi.org/10.1080/01402382.2015.1004230.

Arzheimer, Kai, Carl Berning, Sarah de Lange, Jerome Dutozia, Jocelyn Evans, Myles Gould, Eelco Harteveld, et al. 2024. "How Local Context Affects Populist Radical Right Support: A Cross-National Investigation Into Mediated and Moderated Relationships." *British Journal of Political Science* 54 (4): 1133–1158. https://doi.org/10.1017/s0007123424000085.

Arzheimer, Kai, and Carl C. Berning. 2019. "How the Alternative for Germany (AfD) and their voters veered to the radical right, 2013–2017." *Electoral Studies* 60:102040. https://doi.org/10.1016/j.electstud.2019.04.004.

Atzpodien, Dana Siobhan. 2022. "Party competition in migration debates: The influence of the AfD on party positions in German state parliaments." *German Politics* 31 (3): 381–398. https://doi.org/10.1080/09644008.2020.1860211.

Baccini, Leonardo, and Thomas Sattler. 2025. "Austerity, economic vulnerability, and populism." *American Journal of Political Science* 69 (3): 899–914. https://doi.org/10.1111/ajps.12865.

Bakker, Bert N., Gijs Schumacher, and Matthijs Rooduijn. 2021. "The Populist Appeal: Personality and Antiestablishment Communication." *The Journal of Politics* 83 (2): 589–601. https://doi.org/10.1086/710014.

Barr, Robert R. 2009. "Populists, outsiders and anti-establishment politics." *Party Politics* 15 (1): 29–48.

Becher, Michael, Irene Menéndez González, and Daniel Stegmüller. 2023. "Proportional Representation and Right-Wing Populism: Evidence from Electoral System Change in Europe." *British Journal of Political Science* 53 (1): 261–268. https://doi.org/10.1017/s0007123421000703.

Bélanger, Marie-Ève, and Natasha Wunsch. 2022. "From cohesion to contagion? Populist radical right contestation of EU enlargement." *JCMS: Journal of Common Market Studies* 60 (3): 653–672. https://doi.org/10.1111/jcms.13280.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation learning: A review and new perspectives." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8): 1798–1828. https://doi.org/10.1109/tpami.2013.50.

Benoit, Kenneth, Scott De Marchi, Conor Laver, Michael Laver, and Jinshuai Ma. 2025. "Using large language models to analyze political texts through natural language understanding." *American Journal of Political Science,* https://doi.org/10.1111/ajps.70050.

Berbuir, Nicole, Marcel Lewandowsky, and Jasmin Siri. 2015. “The AfD and its Sympathisers: Finally a Right-Wing Populist Movement in Germany?” *German Politics* 24 (2): 154–178. https://doi.org/10.1080/09644008.2014.982546.

Berman, Sheri. 2021. “The Causes of Populism in the West.” *Annual Review of Political Science* 24 (1): 71–88. https://doi.org/10.1146/annurev-polisci-041719-102503.

Bestvater, Samuel E, and Burt L Monroe. 2023. “Sentiment is not stance: Target-aware opinion classification for political text analysis.” *Political Analysis* 31 (2): 235–256. https://doi.org/10.1017/pan.2022.10.

Betz, Hans-George. 1993. “The new politics of resentment: Radical right-wing populist parties in Western Europe.” *Comparative Politics,* 413–427. https://doi.org/10.2307/422034.

Bichay, Nicolas. 2020. “Public campaign financing and the rise of radical-right parties.” *Electoral Studies* 66:102159.

Blais, André. 1991. “The debate over electoral systems.” *International Political Science Review* 12 (3): 239–260.

Blätte, Andreas. 2020. *polmineR: Verbs and Nouns for Corpus Analysis.* V. v0.8.5. Computer software, September. https://doi.org/10.5281/zenodo.4042093.

Bol, Damien, and Pierre-Henri Bono. 2025. “Can ChatGPT accurately identify the position of parties? A validation study with an expert survey in France.” *Research & Politics* 12 (2): 20531680251335653. https://doi.org/10.1177/20531680251335653.

Bräuninger, Thomas, and Marc Debus. 2009. “Legislative agenda-setting in parliamentary democracies.” *European Journal of Political Research* 48 (6): 804–839.

Bustikova, Lenka. 2014. “Revenge of the Radical Right.” *Comparative Political Studies* 47 (12): 1738–1765. https://doi.org/10.1177/0010414013516069.

Calvo, Esther Mary L, Hanna Bäck, and Royce Carroll. 2024. “Debating the populist pariah: Changing party dynamics and elite rhetoric in the Swedish Riksdag.” *Political Research Quarterly* 77 (3): 950–961.

Campbell, John L, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. “Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement.” *Sociological Methods & Research* 42 (3): 294–320. https://doi.org/10.1177/0049124113500475.

Canovan, Margaret. 1999. “Trust the people! Populism and the two faces of democracy.” *Political Studies* 47 (1): 2–16. https://doi.org/10.1111/1467-9248.00184.

Carey, John M, and Matthew Soberg Shugart. 1995. “Incentives to cultivate a personal vote: A rank ordering of electoral formulas.” *Electoral Studies* 14 (4): 417–439. https://doi.org/10.1016/0261-3794(94)00035-2.

Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu. 2026. “Causal panel analysis under parallel trends: lessons from a large reanalysis study.” *American Political Science Review* 120 (1): 245–266. https://doi.org/10.1017/s0003055425000243.

Church, Kenneth Ward. 2017. “Word2Vec.” *Natural Language Engineering* 23 (1): 155–162. https://doi.org/10.1017/s1351324916000334.

Colantone, Italo, and Piero Stanig. 2018. “The Trade Origins of Economic Nationalism: Import Competition and Voting Behavior in Western Europe.” *American Journal of Political Science* 62 (4): 936–953. https://doi.org/10.1111/ajps.12358.

Dai, Andrew M, Christopher Olah, and Quoc V Le. 2015. “Document embedding with paragraph vectors.” *arXiv preprint arXiv:1507.07998,* https://doi.org/10.48550/arXiv.1507.07998.

Dancygier, Rafaela, Sirus H. Dehdari, David D. Laitin, Moritz Marbach, and Kåre Vernby. 2025. “Emigration and radical right populism.” *American Journal of Political Science* 69 (1): 252–267. https://doi.org/10.1111/ajps.12852.

De Chaisemartin, Clément, and Xavier d’Haultfoeuille. 2023. “Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey.” *The Econometrics Journal* 26 (3): C1–C30. https://doi.org/10.1093/ectj/utac017.

De Lange, Sarah L. 2012. “New alliances: why mainstream parties govern with radical right-wing populist parties.” *Political Studies* 60 (4): 899–918. https://doi.org/10.1111/j.1467-9248.2012.00947.x.

DeBell, Matthew. 2013. “Harder Than It Looks: Coding Political Knowledge on the ANES.” *Political Analysis* 21 (4). https://doi.org/10.1093/pan/mpt010.

Der Standard. 2016. “Deutsche AfD und FPÖ beschließen Zusammenarbeit” [in German]. Der Standard. Accessed May 23, 2026. https://www.derstandard.at/story/2000031396307/deutsche-afd-und-fpoe-beschliessen-zusammenarbeit.

Deutsche Welle. 2020. “AfD sacks former spokesman over comments on migrants.” Accessed: 2026-05-08. https://www.dw.com/en/germany-afd-sacks-former-spokesman-for-saying-migrants-could-be-gassed/a-55082170.

Di Leo, Riccardo, Chen Zeng, Elias Dinas, and Reda Tamtam. 2025. “Mapping (a) ideology: A taxonomy of european parties using generative llms as zero-shot learners.” *Political Analysis* 33 (4): 456–463. https://doi.org/10.1017/pan.2025.7.

Die Welt. 2017. “Albig hopes for continuation of Red-Green-Blue coalition.” Accessed: 2026-05-08. https://www.welt.de/regionales/hamburg/article161274626/Albig-hofft-auf-Fortsetzung-von-Rot-Gruen-Blau.html.

Dinas, Elias, Konstantinos Matakos, Dimitrios Xefteris, and Dominik Hangartner. 2019. “Waking Up the Golden Dawn: Does Exposure to the Refugee Crisis Increase Support for Extreme-Right Parties?” *Political Analysis* 27 (2): 244–254. https://doi.org/10.1017/pan.2018.48.

Döring, Holger, and Philip Manow. 2012. “Parliament and government composition database (ParlGov).” *An infrastructure for empirical information on parties, elections and governments in modern democracies. Version* 12 (10).

Downs, Anthony. 1957. *An Economic Theory of Democracy.* Harper / Row.

Downs, William. 2001. “Pariahs in their midst: Belgian and Norwegian parties react to extremist threats.” *West European Politics* 24 (3): 23–42. https://doi.org/10.1080/01402380108425451.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. “Improving factuality and reasoning in language models through multiagent debate.” In *Forty-First International Conference on Machine Learning.*

Enevoldsen, Kenneth, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, et al. 2025. *MMTEB: Massive Multilingual Text Embedding Benchmark.*

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. “Language-agnostic BERT sentence embedding.” In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 878–891.

Fiorina, Morris P, and Samuel J Abrams. 2008. “Political polarization in the American public.” *Annual Review of Political Science* 11 (1): 563–588.

Franzese Jr, Robert J. 2002. “Electoral and partisan cycles in economic policies and outcomes.” *Annual Review of Political Science* 5 (1): 369–421.

Gelman, Andrew, and Eric Loken. 2013. *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.* Discussion note. Department of Statistics, Columbia University.

Gest, Justin, Tyler Reny, and Jeremy Mayer. 2018. “Roots of the Radical Right: Nostalgic Deprivation in the United States and Britain.” *Comparative Political Studies* 51 (13): 1694–1719. https://doi.org/10.1177/0010414017720705.

Geys, Benny, Bruno Heyndels, and Jan Vermeir. 2006. “Explaining the formation of minimal coalitions: Anti-system parties and anti-pact rules.” *European Journal of Political Research* 45 (6): 957–984.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. “ChatGPT outperforms crowd workers for text-annotation tasks.” *Proceedings of the National Academy of Sciences* 120 (30): e2305016120. https://doi.org/10.1073/pnas.2305016120.

Goodman-Bacon, Andrew. 2021. “Difference-in-differences with variation in treatment timing.” *Journal of Econometrics* 225 (2): 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014.

Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. “Revisiting deep learning models for tabular data.” *Advances in Neural Information Processing Systems* 34:18932–18943. https://doi.org/10.48550/arXiv.2106.11959.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297. https://doi.org/10.1093/pan/mps028.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. "On calibration of modern neural networks." In *International Conference on Machine Learning,* 1321–1330.

Halikiopoulou, Daphne, and Tim Vlandas. 2020. "When economic and cultural interests align: the anti-immigration voter coalitions driving far right party success in Europe." *European Political Science Review* 12 (4): 427–448. https://doi.org/10.1017/s175577392000020x.

Hallin, Daniel C, and Paolo Mancini. 2004. *Comparing media systems: Three models of media and politics.* Cambridge University Press.

Hansen, Michael A., and Jonathan Olsen. 2019. "Flesh of the Same Flesh: A Study of Voters for the Alternative for Germany (AfD) in the 2017 Federal Election." *German Politics* 28 (1): 1–19. https://doi.org/10.1080/09644008.2018.1509312.

———. 2024. "The Alternative for Germany (AfD) as populist issue entrepreneur: Explaining the party and its voters in the 2021 German federal election." *German Politics* 33 (4): 643–667. https://doi.org/10.1080/09644008.2022.2087871.

Hawkins, Kirk A. 2009. "Is Chávez populist? Measuring populist discourse in comparative perspective." *Comparative Political Studies* 42 (8): 1040–1067.

He, Jianfeng, Linlin Yu, Changbin Li, Runing Yang, Fanglan Chen, Kangshuo Li, Min Zhang, et al. 2026. "Survey of uncertainty estimation in LLMs - Sources, methods, applications, and challenges." *Information Fusion* 130:104057. https://doi.org/10.1016/j.inffus.2025.104057.

Heinze, Anna-Sophie. 2022. "Dealing with the populist radical right in parliament: mainstream party responses toward the Alternative for Germany." *European Political Science Review* 14 (3): 333–350. https://doi.org/10.1017/s1755773922000108.

Heseltine, Michael, and Bernhard Clemm von Hohenberg. 2024. "Large language models as a substitute for human experts in annotating political text." *Research & Politics* 11 (1): 20531680241236239. https://doi.org/10.1177/20531680241236239.

Hix, Simon, Richard Whitaker, and Galina Zapryanova. 2024. "The political space in the European parliament: Measuring MEPs' preferences amid the rise of Euroscepticism." *European Journal of Political Research* 63 (1): 153–171.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." *ACM Transactions on Information Systems* 43 (2): 1–55. https://doi.org/10.1145/3703155.

Huang, Xin, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. 2020. "Tabtransformer: Tabular data modeling using contextual embeddings." *arXiv preprint arXiv:2012.06678,* https://doi.org/10.48550/arXiv.2012.06678.

Ignazi, Piero. 1992. “The silent counter-revolution: Hypotheses on the emergence of extreme right-wing parties in Europe.” *European Journal of Political Research* 22 (1): 3–34. https://doi.org/10.1111/j.1475-6765.1992.tb00303.x.

Imai, Kosuke, and In Song Kim. 2021. “On the use of two-way fixed effects regression models for causal inference with panel data.” *Political Analysis* 29 (3): 405–415. https://doi.org/10.1017/pan.2020.33.

Inglehart, Ronald. 1977. *The silent revolution: Changing values and political styles among Western publics.* Princeton University Press.

Jackman, Robert W, and Karin Volpert. 1996. “Conditions favouring parties of the extreme right in Western Europe.” *British Journal of Political Science* 26 (4): 501–521.

Jagers, Jan, and Stefaan Walgrave. 2007. “Populism as political communication style: An empirical study of political parties’ discourse in Belgium.” *European Journal of Political Research* 46 (3): 319–345.

Joshi, Aditya, Pushpak Bhattacharyya, and Mark J Carman. 2017. “Automatic sarcasm detection: A survey.” *ACM Computing Surveys (CSUR)* 50 (5): 1–22.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. “The state and fate of linguistic diversity and inclusion in the NLP world.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 6282–6293.

Jung, Yujin J, Eduardo Ryô Tamaki, Julia Chatterley, Grant Mitchell, Semir Dzebo, Cristóbal Sandoval, Levente Littvay, and Kirk A Hawkins. 2025. “Populism Meets AI: Advancing Populism Research with LLMs.” *arXiv preprint arXiv:2510.07458.*

Kirby, Paul. 2025. “AfD classified as extreme-right by German intelligence.” Accessed: 2026-05-08. https://www.bbc.com/news/articles/cwy6zk9wkrdo.

Klemmensen, Robert, Sara Binzer Hobolt, and Martin Ejnar Hansen. 2007. “Estimating policy positions using political texts: An evaluation of the Wordscores approach.” *Electoral Studies* 26 (4): 746–755.

Knight, Alan. 1998. “Populism and neo-populism in Latin America, especially Mexico.” *Journal of Latin American Studies* 30 (2): 223–248.

Lau, Jey Han, and Timothy Baldwin. 2016. “An empirical evaluation of doc2vec with practical insights into document embedding generation.” In *Proceedings of the 1st Workshop on Representation Learning for NLP,* 78–86.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. “Extracting policy positions from political texts using words as data.” *American Political Science Review* 97 (2): 311–331. https://doi.org/10.1017/s0003055403000698.

Laver, Michael, and Ian Budge. 1992. “Measuring policy distances and modelling coalition formation.” In *Party Policy and Government Coalitions,* 15–40. Springer.

Le Mens, Gaël, and Aina Gallego. 2025. "Positioning Political Texts with Large Language Models by Asking and Averaging." *Political Analysis* 33 (3): 274–282. https://doi.org/10.1017/pan.2024.29.

Lehmann, Pola, and Lisa Zehnter. 2024. "The Self-Proclaimed Defender of Freedom: The AfD and the Pandemic." *Government and Opposition* 59 (4): 1109–1127. https://doi.org/10.1017/gov.2022.5.

Leonhardt, Christoph, and Andreas Blätte. 2020. (AustroParl Corpus of Parliamentary Debates. V. 0.1.0). May. https://doi.org/10.5281/zenodo.3819505.

Lewandowsky, Marcel, Julia Schwanholz, Christoph Leonhardt, and Andreas Blätte. 2021. "New parties, populism, and parliamentary polarization: Evidence from plenary debates in the German Bundestag." In *The Palgrave handbook of populism,* 611–627. Springer.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. "Holistic evaluation of language models." *arXiv preprint arXiv:2211.09110.*

Licht, Hauke. 2023. "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings." *Political Analysis* 31 (3): 366–379. https://doi.org/10.1017/pan.2022.29.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2024. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *American Journal of Political Science* 68 (1): 160–176. https://doi.org/https://doi.org/10.1111/ajps.12723.

Lowe, Will. 2008. "Understanding wordscores." *Political Analysis* 16 (4): 356–371.

Lucardie, Paul. 2008. "The Netherlands: populism versus pillarization." In *Twenty-First Century Populism: The Spectre of Western European Democracy,* 151–165. Springer.

Mair, Peter. 2002. "Populist Democracy vs Party Democracy." In *Democracies and the Populist Challenge,* 81–98. Springer.

March, Luke. 2012. *Radical Left Parties in Europe.* Routledge.

———. 2017. "Left and right populism compared: The British case." *The British Journal of Politics and International Relations* 19 (2): 282–303. https://doi.org/10.1177/1369148117701753.

Mares, Isabela. 2026. "Defending Parliament Responses of Mainstream Parties to Parliamentary Erosion." *Comparative Political Studies,* 00104140251342933. https://doi.org/10.1177/00104140251342933.

Margalit, Yotam. 2011. "Costly Jobs: Trade-related Layoffs, Government Compensation, and Voting in U.S. Elections." *American Political Science Review* 105 (1): 166–188. https://doi.org/10.1017/s000305541000050x.

Marks, Gary, Liesbet Hooghe, and Arjan H Schakel. 2008. "Measuring regional authority." *Regional and Federal Studies* 18 (2-3): 111–121. https://doi.org/10.1080/13597560801979464.

Matsunaga, Miku, and Thomas Winzen. 2025. "Strengthening mainstream consensus? The effect of radical right populist parties on the defense policies of left parties." *Political Science Research and Methods* 13 (3): 629–644.

McAleese, Nat, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. "LLM critics help catch LLM bugs." *arXiv preprint arXiv:2407.00215.*

McCallum, Andrew, and Kamal Nigam. 1998. *A Comparison of Event Models for Naive Bayes Text Classification.* AAAI Technical Report WS-98-05.

McElroy, Gail, and Kenneth Benoit. 2007. "Party groups and policy positions in the European Parliament." *Party Politics* 13 (1): 5–28.

McInnes, Leland, John Healy, and James Melville. 2018. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426,* https://doi.org/10.48550/arXiv.1802.03426.

Meguid, Bonnie M. 2005. "Competition Between Unequals: The Role of Mainstream Party Strategy in Niche Party Success." *American Political Science Review* 99 (3): 347–359. https://doi.org/10.1017/s0003055405051701.

———. 2008. *Party Competition between Unequals Strategies and Electoral Fortunes in Western Europe.* Cambridge University Press.

Meijers, Maurits J, and Harmen Van der Veer. 2019. "Issue competition without electoral incentives? A study of issue emphasis in the European parliament." *The Journal of Politics* 81 (4): 1240–1253. https://doi.org/10.1086/704224.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale." *Research & Politics* 11 (1): 20531680241231468. https://doi.org/10.1177/20531680241231468.

Mény, Yves, and Yves Surel. 2002. "The constitutive ambiguity of populism." In *Democracies and the Populist Challenge,* 1–21. Springer.

Merrill, Samuel, and Bernard Grofman. 2019. "What are the effects of entry of new extremist parties on the policy platforms of mainstream parties?" *Journal of Theoretical Politics* 31 (3): 453–473.

Merz, Nicolas, Sven Regel, and Jirka Lewandowski. 2016. "The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis." *Research & Politics* 3 (2): 2053168016643346. https://doi.org/10.1177/2053168016643346.

Messick, Samuel. 1987. "Validity." *ETS Research Report Series* 1987 (2): i–208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x.

———. 1995. "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning." *American Psychologist* 50 (9): 741. https://doi.org/10.1037/0003-066x.50.9.741.

Minkenberg, Michael. 2000. “The renewal of the radical right: Between modernity and anti-modernity.” *Government and Opposition* 35 (2): 170–188.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. “Distant supervision for relation extraction without labeled data.” In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP,* 1003–1011.

Mosteller, Frederick, and David L Wallace. 1963. “Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers.” *Journal of the American Statistical Association* 58 (302): 275–309. https://doi.org/10.1080/01621459.1963.10500849.

Mudde, Cas. 2004. “The populist zeitgeist.” *Government and Opposition* 39 (4): 541–563. https://doi.org/10.1111/j.1477-7053.2004.00135.x.

———. 2007. *Populist Radical Right Parties in Europe.* Cambridge University Press.

Mudde, Cas, and Cristóbal Rovira Kaltwasser. 2013. “Exclusionary vs. inclusionary populism: Comparing contemporary Europe and Latin America.” *Government and Opposition* 48 (2): 147–174. https://doi.org/10.1017/gov.2012.11.

Muennighoff, Niklas, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. “MTEB: Massive Text Embedding Benchmark.” *arXiv preprint arXiv:2210.07316,* https://doi.org/10.48550/arxiv.2210.07316.

Norris, Pippa. 2005. *Radical Right: Voters and Parties in the Electoral Market.* Cambridge University Press.

Oesch, Daniel. 2008. “Explaining workers’ support for right-wing populist parties in Western Europe: Evidence from Austria, Belgium, France, Norway, and Switzerland.” *International Political Science Review* 29 (3): 349–373. https://doi.org/10.1177/0192512107088390.

Oesterreich, Detlef. 2005. “Flight into security: A new approach and measure of the authoritarian personality.” *Political Psychology* 26 (2): 275–298.

Ornstein, Joseph T, Elise N Blasingame, and Jake S Truscott. 2025. “How to train your stochastic parrot: Large language models for political texts.” *Political Science Research and Methods* 13 (2): 264–281. https://doi.org/10.1017/psrm.2024.64.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. “Thumbs up? Sentiment classification using machine learning techniques.” In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002),* 79–86.

Patton, David F. 2020. “Party-political responses to the Alternative for Germany in comparative perspective.” *German Politics and Society* 38 (1): 77–104.

Pérez, Laura Cabeza. 2025. “Subnational Roads to National Success? Decentralization and the Rise of Populist Parties in Europe.” *European Political Science Review* 17 (3): 432–448.

Pirro, Andrea LP, and Stijn Van Kessel. 2017. "United in opposition? The populist radical right's EU-pessimism in times of crisis." *Journal of European Integration* 39 (4): 405–420. https://doi.org/10.1080/07036337.2017.1281261.

Platt, John. 2000. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." In *Advances in Large Margin Classifiers,* 61–74. Cambridge, MA: MIT Press.

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." *Legislative Studies Quarterly* 44 (1): 97–131. https://doi.org/https://doi.org/10.1111/lsq.12218.

Pytlas, Bartek, and Jan Biehler. 2024. "The AfD within the AfD: Radical Right Intra-Party Competition and Ideational Change." *Government and Opposition* 59 (2): 322–340. https://doi.org/10.1017/gov.2023.13.

Qader, Wisam, Musa Ameen, and Bilal Ahmed. 2019. "An overview of bag of words; importance, implementation, applications, and challenges." In *2019 International Engineering Conference (IEC),* 200–204. IEEE.

Rahat, Gideon, and Reuven Y Hazan. 2001. "Candidate selection methods: An analytical framework." *Party Politics* 7 (3): 297–322.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3982–3992.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. "The structural topic model and applied social science." In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation,* 4:1–20. 1. Harrahs and Harveys, Lake Tahoe.

Rooduijn, Matthijs, Sarah L De Lange, and Wouter Van Der Brug. 2014. "A populist Zeitgeist? Programmatic contagion by populist parties in Western Europe." *Party Politics* 20 (4): 563–575.

Rooduijn, Matthijs, Andrea LP Pirro, Daphne Halikiopoulou, Caterina Froio, Stijn Van Kessel, Sarah L De Lange, Cas Mudde, and Paul Taggart. 2024. "The PopuList: A database of populist, far-left, and far-right parties using expert-informed qualitative comparative classification (EiQCC)." *British Journal of Political Science* 54 (3): 969–978. https://doi.org/10.1017/s0007123423000431.

Rovny, Jan, Jonathan Polk, Ryan Bakker, Liesbet Hooghe, Seth Jolly, Gary Marks, Marco Steenbergen, and Milada Anna Vachudova. 2025. "The 2024 Chapel Hill Expert Survey on political party positioning in Europe: Twenty-five years of party positional data." *Electoral Studies* 97:102981. https://doi.org/10.1016/j.electstud.2025.102981.

Rydgren, Jens. 2008. "Immigration sceptics, xenophobes or racists? Radical right-wing voting in six West European countries." *European Journal of Political Research* 47 (6): 737–765. https://doi.org/10.1111/j.1475-6765.2008.00784.x.

Sartori, Giovanni. 1976. *Parties and Party Systems: Volume 1: A Framework for Analysis.* Cambridge University Press.

Scarrow, Susan E. 2007. "Political finance in comparative perspective." *Annual Review of Political Science* 10 (1): 193–210.

Schmidt, Vivien A. 2008. "Discursive institutionalism: The explanatory power of ideas and discourse." *Annual Review of Political Science* 11 (1): 303–326.

Schwalbach, Jan. 2023. "Talking to the Populist Radical Right: A Comparative Analysis of Parliamentary Debates." *Legislative Studies Quarterly* 48 (2): 371–397. https://doi.org/10.1111/lsq.12397.

Scott, William A. 1955. "Reliability of content analysis: The case of nominal scale coding." *Public Opinion Quarterly,* 321–325. https://doi.org/10.1086/266577.

Shehaj, Albana, Adrian J Shin, and Ronald Inglehart. 2021. "Immigration and right-wing populism: An origin story." *Party Politics* 27 (2): 282–293. https://doi.org/10.1177/1354068819849888.

Shorinwa, Ola, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. "A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions." *ACM Comput. Surv.* (New York, NY, USA) 58 (3). https://doi.org/10.1145/3744238.

Sides, John, Michael Tesler, and Lynn Vavreck. 2019. *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America.* Princeton University Press.

Silva Filho, Telmo, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. 2023. "Classifier calibration: a survey on how to assess and improve predicted class probabilities." *Machine Learning* 112 (9): 3211–3260. https://doi.org/10.1007/s10994-023-06336-7.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22 (11): 1359–1366. https://doi.org/10.1177/0956797611417632.

Slapin, Jonathan B, and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52 (3): 705–722. https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Stanley, Ben. 2008. "The thin ideology of populism." *Journal of Political Ideologies* 13 (1): 95–110. https://doi.org/10.1080/13569310701822289.

Stevenson, Randolph T, and Lynn Vavreck. 2000. "Does campaign length matter? Testing for cross-national effects." *British Journal of Political Science* 30 (2): 217–235.

Stimson, James. 2018. *Public opinion in America: Moods, cycles, and swings.* Routledge.

Toney, Autumn, and Ryan Wails. 2025. "Certain but not Probable? Differentiating Certainty from Probability in LLM Token Outputs for Probabilistic Scenarios." In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025),* 51–60. Suzhou, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.uncertainlp-main.6.

Valentim, Vicente, and Tobias Widmann. 2023. "Does radical-right success make the political debate more negative? Evidence from emotional rhetoric in German state parliaments." *Political Behavior* 45 (1): 243–264. https://doi.org/10.1007/s11109-021-09697-8.

Van Atteveldt, Wouter, Mariken ACG Van der Velden, and Mark Boukes. 2021. "The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms." *Communication Methods and Measures* 15 (2): 121–140. https://doi.org/10.1080/19312458.2020.1869198.

Van Spanje, Joost, and Wouter Van Der Brug. 2007. "The party as pariah: The exclusion of anti-immigration parties and its effect on their ideological positions." *West European Politics* 30 (5): 1022–1040.

Wagner, Franziska, Dean Schafer, and Mehmet Yavuz. 2025. "Opposition to government and back: How illiberal parties shape immigration discourse and party competition." *Politics and Governance* 13. https://doi.org/10.17645/pag.9609.

Wang, Junlin, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. 2025. "Mixture-of-agents enhances large language model capabilities." In *International Conference on Learning Representations,* 2025:33944–33963.

Wang, Yu. 2023. "Topic classification for political texts with pretrained language models." *Political Analysis* 31 (4): 662–668. https://doi.org/10.1017/pan.2023.3.

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang. 2016. "A survey of transfer learning." *Journal of Big Data* 3 (1): 9. https://doi.org/10.1186/s40537-016-0043-6.

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. "CCNet: Extracting high quality monolingual datasets from web crawl data." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 4003–4012.

Westle, Bettina, and Oskar Niedermayer. 1992. "Contemporary right-wing extremism in West Germany." *European Journal of Political Research* 22 (1): 83–100. https://doi.org/10.1111/j.1475-6765.1992.tb00306.x.

Weyland, Kurt. 2001. "Clarifying a contested concept: Populism in the study of Latin American politics." *Comparative Politics,* 1–22. https://doi.org/10.2307/422412.

Wu, Lingfei, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnwan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J Witbrock. 2018. "Word mover's embedding: From word2vec to document embedding." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 4524–4534.

Wunsch, Natasha, and Marie-Eve Bélanger. 2024. “Radicalisation and discursive accommodation: responses to rising Euroscepticism in the European Parliament.” *West European Politics* 47 (6): 1223–1250. https://doi.org/10.1080/01402382.2023.2202031.

Xiong, Miao, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. “Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs.” In *International Conference on Learning Representations,* 2024:23650–23678.

Xu, Yuemei, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. “A survey on multilingual large language models: Corpora, alignment, and bias.” *Frontiers of Computer Science* 19 (11): 1911362. https://doi.org/10.1007/s11704-024-40579-4.

Ye, Fanghua, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. “Benchmarking LLMs via Uncertainty Quantification.” In *Advances in Neural Information Processing Systems,* 37:15356–15385. https://doi.org/10.52202/079017-0491.

Young, Lori, and Stuart Soroka. 2012. “Affective news: The automated coding of sentiment in political texts.” *Political Communication* 29 (2): 205–231. https://doi.org/10.1080/10584609.2012.671234.

Zaslove, Andrej, Robert A Huber, and Maurits J Meijers. 2024. “The state of populism: Introducing the 2023 wave of the Populism and political parties expert survey.” *Party Politics,* 13540688251361813.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. “Judging llm-as-a-judge with mt-bench and chatbot arena.” *Advances in Neural Information Processing Systems* 36:46595–46623. https://doi.org/10.52202/075280-2020.

A.1 Landtag Corpus

A.1.1 Landtag Text Extraction Flowchart

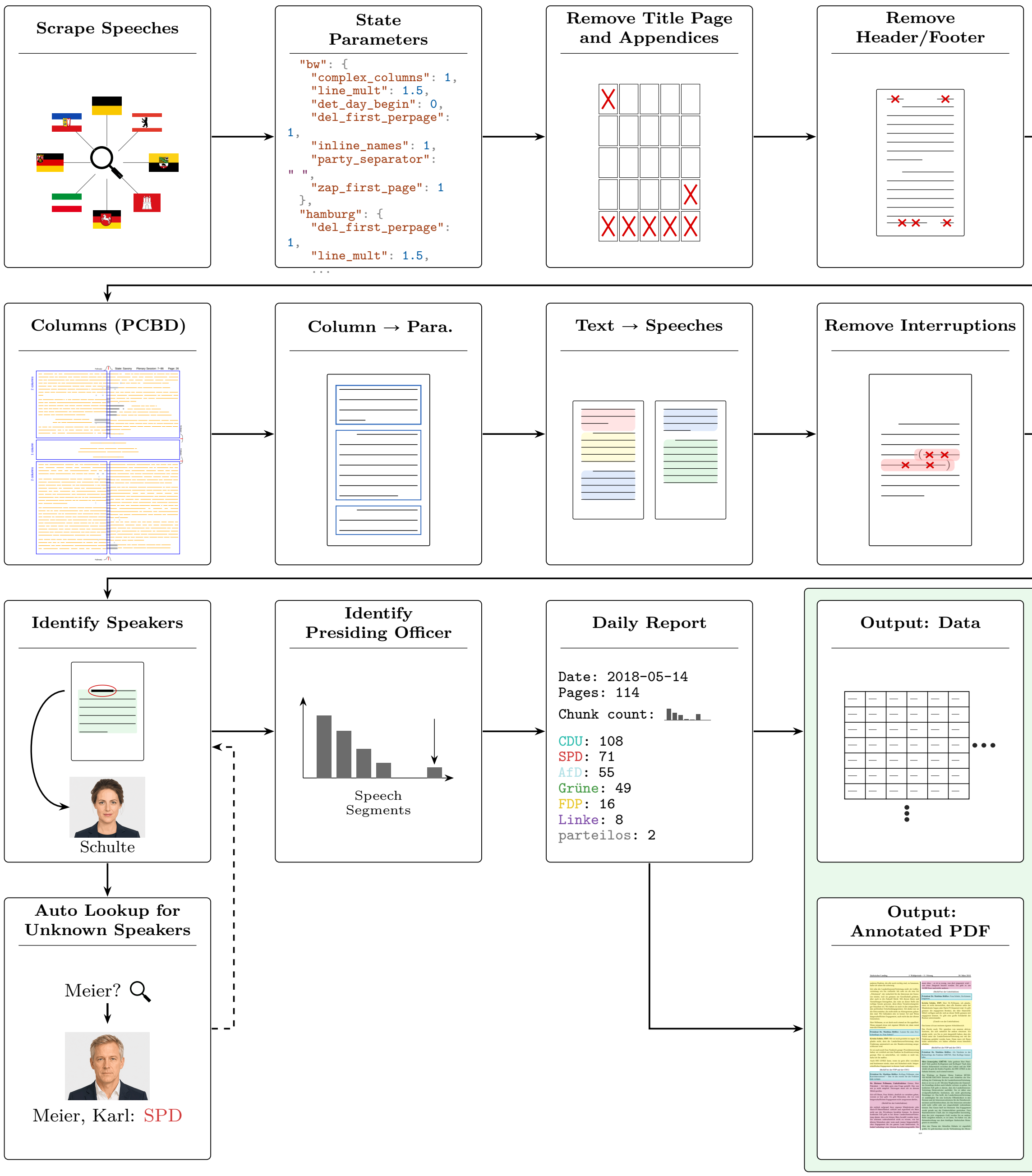


Figure A.1: Landtag Corpus Construction: This flowchart describes the steps taken to build the German Landtag corpus from gathering the data, linearizing columnar PDFs into text, and extracting speaker-specific information (including party), through to producing the tabular data outputs used for analysis in this paper. Politician photos and names are composites for demonstration purposes only.

A.1.2 Probabilistic Columnar Boundary Detection

Researchers who work with text-as-data often need to extract that text from published documents, many of which are distributed in the form of PDFs. PDFs store text natively as $(x, y, width, height, text)$ tuples on a per-symbol basis. As a result, extracting text from a PDF and linearizing it into chapters, paragraphs, or in my case, speeches, requires an algorithmic assumption about which way text is ordered. In single-column PDFs, this assumption follows the ordinary way that humans read the text: left to right, top to bottom. But government documents, including the plenary speech documents I work with, often contain more complex layouts. In specific, several states distribute plenary speeches as multi-columnar PDFs. Because plenary documents are typeset for print, recto and verso (left and right) pages typically differ in their layout, making it difficult to impose a fixed assumption about column boundaries; fonts and layouts also change over time, and differ between states. At the time this article was being prepared, artificial intelligence-based layout detection remained costly and performance was a concern. As a result, I develop *probabilistic columnar boundary detection* (PCBD) to extract text from PDFs.

The algorithm is simple: treat the page as a grid of points. Transform the bounding box of each word on the page into a series of filled points, and treat other points as empty. Sweeping across the page horizontally, calculate the probability density of filled points for each x value. If the document is preprocessed with headers and footers removed, column breaks should be zero-density; even if not, column breaks should be low density. Some layouts do not admit this algorithm because they feature shifts in column break within the page; this is especially true in my case for states that demarcate the end of debate on a particular subject with an intermediary header. In cases with varying layouts, instead sweep the y-axis. The ranges with no density are potential transition points between column layouts: run the original algorithm on each portion. Panel A of **Figure A.2** depicts the internal representation of the PCBD algorithm: the visual diagnostic output allows for easy human audit. I hand-audit 5% of all pages in the three states with the most complex layouts and observe 100% accuracy in column detection. I ease the audit process by visually annotating speech PDFs with boxes of speech boundaries and with speakers colored according to their party identification, as depicted in Panel B of **Figure A.2**.

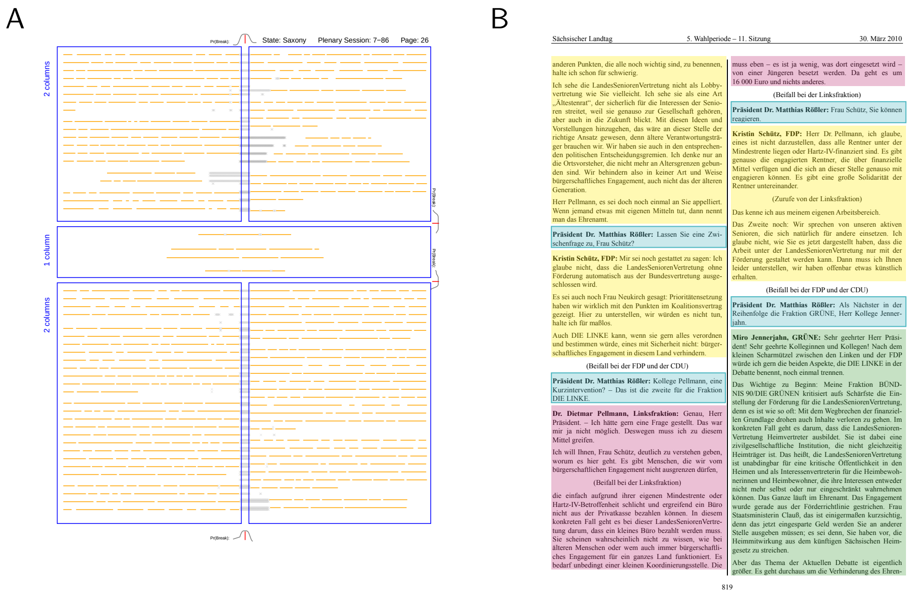


Figure A.2: Document Column and Speech Detection: This figure depicts the intermediate steps of extracting Landtag speech from plenary minute PDFs. **Panel A**: The probabilistic columnar boundary detection (PCBD) algorithm identifies the switch from two columns to one and back again, as well as both column breaks and both breaks between regions. **Panel B**: Speaker detection highlights the beginning and end of each speech on the page, and color annotation is used to provide a visual indication of the party actively speaking; the presiding officer is correctly assigned to the Christlich Demokratische Union (*CDU*) despite not being identified as such in-text.

A.1.3 Landtag Corpus Descriptive Statistics

State	From	To	n	Prop. AfD	Avg. Words	Mean $Pr(Y_i)$	Top Party
Baden-Wurttemberg	2006-06	2024-07	21,496	0.127	484	0.092	CDU
Bavaria	2008-01	2024-07	36,149	0.085	464	0.114	CSU
Berlin	2002-12	2024-09	34,735	0.077	375	0.092	Grune
Brandenburg	2009-10	2024-06	28,933	0.170	447	0.096	CDU
Bremen	2007-06	2024-08	14,908	0.013	485	0.036	CDU
Hamburg	2008-03	2024-06	25,484	0.119	476	0.094	SPD
Hesse	2008-04	2024-08	30,005	0.062	586	0.091	SPD
Lower Saxony	2008-02	2024-09	**44,502**	0.054	370	0.060	CDU
Mecklenburg-Vorpommern	2006-10	2024-03	31,460	0.133	635	**0.131**	Linke
North Rhine-Westphalia	2010-06	2024-09	38,594	0.099	490	0.077	SPD
Rhineland-Palatinate	2011-05	2024-07	17,683	0.154	495	0.070	CDU
Saarland	2009-09	2024-06	7,937	0.131	**820**	0.056	CDU
Saxony	2009-09	2024-06	34,827	0.150	429	0.103	CDU
Saxony-Anhalt	2006-06	2024-08	34,954	**0.172**	399	0.082	Linke
Schleswig-Holstein	2005-03	2024-07	34,691	0.029	529	0.029	SPD
Thuringia	2007-12	2024-06	32,864	0.110	570	0.083	Linke

Table A.1: Landtag Descriptive Statistics by State: This table reports basic descriptive statistics about the Landtag corpus. States with the highest numeric value in numeric columns are bolded to draw attention. I include the state's modal (top) party to help contextualize the state's politics.

A.2 Populism Detector: Document Embeddings

A.2.1 Selection of Embedding Models

The question of which embedding model is best is difficult to answer because the models themselves are trained on a proprietary data blend using proprietary methods. Moreover, there are many embedding providers, and it is infeasible to test all of them to determine subject-matter performance. The industry-standard benchmark for embedding, search, and retrieval tasks is the Multilingual Massive Text Embedding Benchmark (Mmteb) (Muennighoff et al. 2022). **Table A.2** reports a subset of the mid-2025 Mmteb leaderboard. Using this ranking, I select the top three open-source model families and the top three commercial model families. An alternative benchmark, the Retrieval Embedding Benchmark (Rteb), would select four of the same six models (KaLM and Cohere would be outranked).

Model	Family	Param. (B)	Max $\mathbb{R}^D$	Mmteb	Model Type
KaLM-Embedding-Gemma3-12B-2511	Tencent	11.8	3,840	0.723	Open Source
llama-embed-nemotron-8b	Llama	7.50	4,096	0.695	Open Source
Octen-Embedding-8B	Qwen	7.57	4,096	0.706	Open Source
~~Seed1.6-embedding-1215~~	~~Bytedance~~		~~2,048~~	~~0.703~~	~~Commercial~~
gemini-embedding-001	Google		3,072	0.684	Commercial
Linq-Embed-Mistral	Mistral	7.11	4,096	0.615	Open Source
multilingual-e5-large-instruct	RoBERTa	0.56	1,024	0.632	Open Source
GritLM-7B	GritLM	7.24	4,096	0.609	Open Source
Cohere-embed-multilingual-v3.0	Cohere		512	0.611	Commercial
Solon-embeddings-large-0.1	OrdalieTech	0.56	1,024	0.596	Open Source
voyage-3.5	Voyage AI		1,024	0.585	Commercial

Table A.2: Subset of Mmteb Leaderboard, Mid-2025: Using an industry-standard benchmark for emebdding tasks, I identify three commercial and three open-source model families. Struck-through rows were not technically viable at the time of evaluation; greyed rows are other ranked models that were not selected for comparison. Parameter counts are in billions and are unreported for commercial models.

A.2.2 Sensitivity to Embedding Choice

Dependent Variable:	Predicted Populism			
Embedder	Kalm	Nemotron	Octen	Voyage
Labeler	LLM-as-a-Judge			
Model:	(1)	(2)	(3)	(4)
Variables				
AfD Intensity	0.026	0.056***	0.065***	0.110***
	(0.008)	(0.010)	(0.010)	(0.006)
Controls	Yes	Yes	Yes	Yes
Fixed-effects				
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Fit statistics				
Observations	205,100	205,100	205,100	393,951

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.3: Robustness to Embedding Choice: This table depicts the primary analysis run on the 2nd–5th best performing embeddings (Kalm, Nemotron, Octen, and Voyage AI), combined with the LLM-as-a-Judge labeler, and the best-performing models. The results vary from somewhat smaller to somewhat larger than the primary model and all remain positive and significant.

A.3 Populism Detector: Labels

A.3.1 Selection of Labeling Models

Analogously to the embedding models above, I consult an industry-standard benchmark for general LLM intelligence, the Holistic Evaluation of Language Models (Helm) (Liang et al. 2022). **Table A.4** reports a subset of the mid-2025 Helm leaderboard. Using this ranking, I select the top three LLMs with a feasible cost structure and accessible APIs that support structured responses, with a weak preference for including at least one open-source model.

Model Name	Family	Helm	Mmlu	Gpqa	Ifeval	Model Type
(GPT) o4-mini	OpenAI	0.812	0.820	0.735	0.929	Commercial
~~Kimi K2 Instruct~~	~~Kimi~~	~~0.768~~	~~0.819~~	~~0.652~~	~~0.850~~	~~Open Source~~
Gemini 2.5 Pro (preview)	Gemini	0.745	0.863	0.749	0.840	Commercial
~~Grok 3 Beta~~	~~Grok~~	~~0.727~~	~~0.788~~	~~0.650~~	~~0.884~~	~~Commercial~~
Qwen3 235B A22B FP8	Qwen	0.726	0.817	0.623	0.816	Open Source
Llama 4 Maverick Instruct FP8	Llama	0.718	0.810	0.650	0.908	Open Source
Palmyra X5	Palmyra	0.696	0.804	0.661	0.823	Commercial
Claude 3.7 Sonnet	Claude	0.674	0.784	0.608	0.834	Commercial
DeepSeek v3	DeepSeek	0.665	0.723	0.538	0.832	Open Source
Amazon Nova Premier	Amazon	0.637	0.726	0.518	0.803	Commercial

Table A.4: Subset of Helm Leaderboard, Mid-2025: Using an industry-standard benchmark for general LLM intelligence, I identify three models with a feasible cost structure and accessible APIs supporting structured responses, with a weak preference for at least one open-source model. Struck-through rows were not technically viable at the time of evaluation; greyed rows are other ranked models that were not selected for comparison. Mmlu measures academic-style knowledge across domains; Gpqa measures complex reasoning; and Ifeval measures instruction-following.

A.3.2 LLM Prompting and Coding

In addition to the "distant supervision" labeler, I use three LLM-based codings: one using OpenAI's gpt-5.1-mini model, one using Google's gemini-2.5-flash model, and one using the open source Qwen3-30B-A3B-Instruct-2507 model. I supply each LLM with the following instructional prompt to classify the speech:

> The speech below is a speech given in the Austrian parliament in German. Your job is to decide whether the speech contains populist rhetoric. Populist rhetoric includes rhetoric that is anti-elite, anti-expert, anti-gay, anti-woman, anti-Semitic, Islamophobic, anti-vaccine, anti-immigrant, anti-democracy (or uses democracy as a rhetorical flourish to undermine the rule of law), or that appeals to a sense of Austrian nationality or nationalism.
>
> If a speech is responding to another speaker's nationalist rhetoric, it is not necessarily populist. Likewise, speech that merely acknowledges class (for example, criticism of the wealthy or calling for higher taxes) is not necessarily populist unless the speech is framed in a populist way.
>
> Code speeches as 1 (populist), 0 (not populist), or -1 (junk or boilerplate text, no substantive content). You might use -1 to refer to speeches that are procedural, like someone merely noting a break for lunch. Provide a very short, one-sentence explanation of anything coded 1/-1 — it suffices to say 'speech contains anti-immigrant rhetoric' or 'boilerplate text'. Explanations should be given in English and not include direct quotes[...][24]

Prompts for LLMs are comparable to rubrics for human coders. Thus, the specific examples included are not exhaustive contours for coding but merely illustrative examples. The LLM's own internal context and knowledge of populism as a concept also contribute to the coding decision. I ask the LLMs to return reasoning for responses coded as a "1" both because reasoning has been shown to improve output quality and because it enables easier human auditing of outputs. Examples include "Contains anti-elite populist rhetoric—accuses the government and banks of scheming/inaction, frames them as betraying the Austrian people and appeals to voters to punish them." or, more ambiguously, "Contains anti-elite/populist

24. The abridged prompt continues with LLM-specific instructions for structuring data outputs, none of which affect the substantive consideration of the issue.

rhetoric advocating a citizens' referendum and portraying the government/parliament as out of touch with the people." The LLM reasoning is part of the inputs supplied to the LLM-as-a-Judge labeler to reconcile the outputs.

A.3.3 LLM-as-a-Judge Reconciliation Process

The goal of this appendix is to illustrate the value of the LLM-as-a-Judge Reconciliation process. Put simply, why not choose a single labeler, or why not simply take the majority vote in any given coding decision and be done with it? **Figure A.3.3** illustrates that labeler disagreement is hardly uncommon, that the judge often chooses the "minority" viewpoint among models, and that the effect of this can move labels in both populist and non-populist directions.

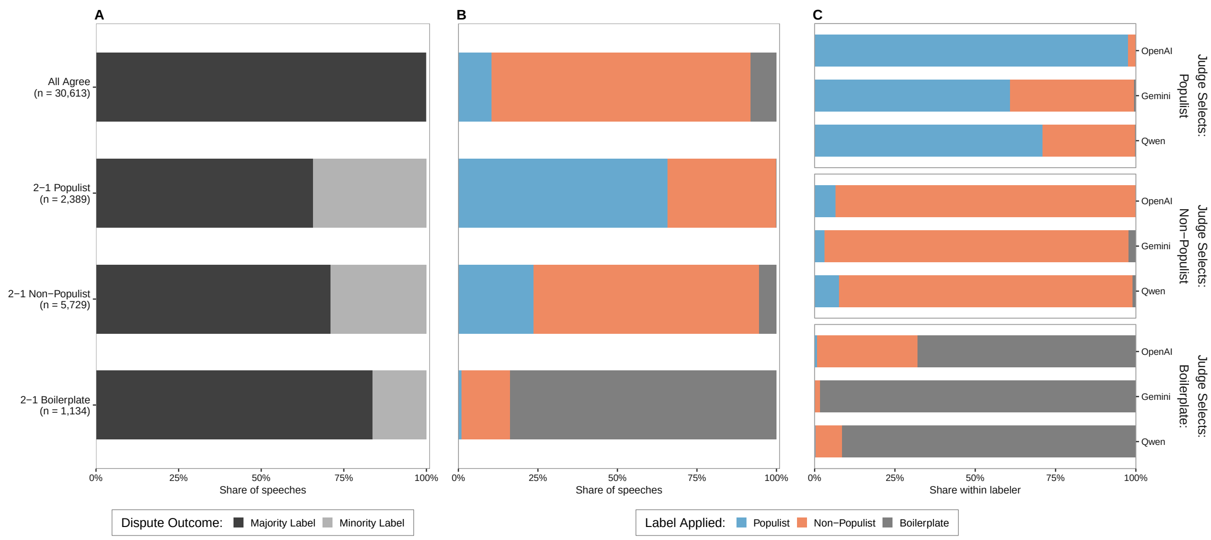


Figure A.3: LLM-as-a-Judge Coder Reconciliation: This figure presents statistics related to the LLM-as-a-Judge Reconciliation process. **Panel A:** Conditional on various arrangements of labeler disagreements, how does the Judge resolve them? In a non-trivial proportion of cases, the Judge selects the outlier LLM's label. **Panel B:** The effect of this can move labels in both populist and non-populist directions. **Panel C:** Conditional on the final selection made by the Judge, how did the agents do? OpenAI seems more aggressive in classifying speeches as populist, and also less willing to classify speeches as boilerplate. n=51 cases produced a three-way disagreement and are suppressed from these figures.

A.3.4 Sensitivity to Label Choice

Dependent Variable:	Predicted Populism			
Embedder	Gemini			
Labeler	Gemini	OpenAI	Qwen	Dist. Supervision
Model:	(1)	(2)	(3)	(4)
Variables				
AfD Intensity	0.158***	0.095***	0.295***	0.069***
	(0.007)	(0.006)	(0.010)	(0.007)
Controls	Yes	Yes	Yes	Yes
Fixed-effects				
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Fit statistics				
Observations	393,951	393,951	393,951	393,951

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.5: Robustness to Labelling Choice: This table depicts the primary analysis run on the Gemini embeddings, varying the labeler, and selecting the best-performing model per labeler. Results are comparable to the primary results (n.b. that with different labelers, there are different scales and base rates), and remain significant.

A.4 Populism Detector: Models

This appendix describes the results of the "model horse race": the process of identifying which model serves as the best predictor of populist speech in the corpus. I first list the models in **Appendix A.4.1**, then show the results of the models on training data in **Appendix A.4.2** (using the selected GEMINI embeddings and LLM-as-a-Judge labels). I re-estimate the study's main results using alternative models in **Appendix A.4.3** and find that my results are not sensitive to the choice of Neural Network 3.

A.4.1 Full Model Parameterization

Model	Notes / Hyperparameters
	Traditional Models
Linear Regression	Output clipped to $[0, 1]$.
Logistic Regression	
Linear Discriminant Analysis	
Naive Bayes	
	Machine Learning
AdaBoost	$n \in \{50, 100, 200\}$; $\eta \in [10^{-2}, 1]$
Elastic Net	$\ell_1/\ell_2 \in [0.1, 1]$; $C \in [10^{-3}, 10]$
Extra Trees	$n \in \{100, 200, 300\}$; $\mathit{depth} \in \{10, 20, \infty\}$; $m \in \{2, 5, 10\}$
LASSO Linear Regression	$\lambda \in [10^{-4}, 1]$
LightGBM	$n \in \{100, 200, 300\}$; $\mathit{depth} \in \{3, 5, 7, \infty\}$; $\mathit{leaves} \in \{31, 50, 100\}$, $\eta \in \{0.01, 0.1, 0.3\}$
Random Forest	$n \in \{100, 200, 300\}$; $m \in \{2, 5, 10\}$
SVM (Linear Kernel)	$C \in [10^{-2}, 100]$
SVM (Radial Kernel)	$\gamma \in [10^{-3}, 10^{-2}]$; $C \in [10^{-1}, 100]$
XGBoost	$n \in \{100, 200, 300\}$; $\mathit{depth} \in \{3, 5, 7\}$; $m \in \{2, 5, 10\}$, $\eta \in \{0.01, 0.1, 0.3\}$
	Neural Networks
Neural Network 1	$512 \to 256 \to 64 \to 1$
Neural Network 2	$512 \to 512 \to 256 \to 1$
Neural Network 3	$256 \to 128 \to 64 \to 1$
Neural Network 4	$64 \to 1$
Neural Network 5	$128 \to 1$
Neural Network 6	$32 \to 128 \to 1$ *(Bottleneck)*
Neural Network 7	$512 \to 1$
Neural Network 8	$\{128, 128\} \to 256 \to 1$ *(Split-Path)*
Neural Network 9	$128 \to 1$ *(Focal Loss Function)*
Neural Network 10	$128 \to 1$ *(Residual Connection)*
Attention Gated Network	$128 \to 128 \to 1$ *(Attention Gated)*

Table A.6: Model Notes and Parameterizations: This display lists all models fit to the training data. Models are broken into three groups: traditional statistical models (fit to the data one-shot, no penalization terms, no hyperparameters); machine learning models (cross-validation used to fit the data and learn model hyperparameters); and neural networks (trained by backpropagation).

A.4.2 Model Horse Race

This appendix presents the results of all the models that were evaluated (on the Gemini embeddings and LLM-as-a-Judge labels, which were the best performing combination). I present the results by model class, the metrics used for model selection and building the composite score, and other metrics of interest. Models perform well across the board, but neural network models generally outperform other model classes.

Model Name	Calibrator	Composite Score	PR-AUC	Accuracy (CB)	1 - Brier	Prediction Gap	Separation
Neural Networks							
Neural Network 3	Platt	**0.866**	0.816	**0.880**	0.954	**0.389**	**0.562**
Neural Network 4	Platt	0.865	**0.822**	0.868	**0.956**	0.392	0.561
Neural Network 2	Platt	0.860	0.803	0.880	0.953	0.418	0.530
Neural Network 7	Platt	0.859	0.807	0.873	0.953	0.403	0.547
Neural Network 8	Platt	0.859	0.802	0.877	0.953	0.404	0.547
Neural Network 1	Platt	0.857	0.802	0.875	0.953	0.419	0.527
Neural Network 10	Platt	0.857	0.802	0.873	0.953	0.404	0.548
Neural Network 5	Platt	0.857	0.805	0.870	0.953	0.410	0.543
Neural Network 6	Platt	0.847	0.794	0.855	0.951	0.427	0.521
Neural Network 9	Platt	0.846	0.786	0.863	0.951	0.433	0.514
Attention Gated Network	Platt	0.841	0.775	0.864	0.949	0.414	0.535
Machine Learning							
Linear Support Vector Machine	Platt	0.854	0.801	0.866	0.952	0.414	0.537
Elastic Net	Isotonic	0.845	0.777	0.870	0.952	0.403	0.549
LASSO Linear Regression	Isotonic	0.842	0.775	0.865	0.952	0.402	0.550
LightGBM	Isotonic	0.804	0.706	0.851	0.944	0.456	0.486
XGBoost	Isotonic	0.786	0.679	0.835	0.940	0.484	0.456
AdaBoost	Isotonic	0.750	0.615	0.819	0.933	0.527	0.406
Random Forest	Isotonic	0.704	0.532	0.797	0.926	0.593	0.329
Traditional Models							
Logistic Regression	Isotonic	0.844	0.776	0.871	0.952	0.406	0.545
Linear Discriminant Analysis	Isotonic	0.841	0.774	0.864	0.952	0.403	0.549
Linear Regression	Isotonic	0.841	0.774	0.864	0.952	0.403	0.549
Naive Bayes	Isotonic	0.624	0.389	0.761	0.913	0.691	0.221

Table A.7: Model Horse Race Results: All models are trained on AustroParl data and a training sample of the German Landtag data and tested and calibrated on held-out data. In addition to the composite score and the components therein, we report the prediction gap (underprediction of calibrated scores on true positives, lower is better) and the separation (difference in mean predicted probability between true positives and true negatives, higher is better). Rows are grouped by model family and sorted by composite score descending.

A.4.3 Sensitivity to Model Choice

Dependent Variable:	Predicted Populism			
Embedder	Gemini			
Classifier	Neural Net 2	Neural Net 4	Neural Net 7	Neural Net 8
Model:	(1)	(2)	(3)	(4)
Variables				
AfD Intensity	0.080***	0.077***	0.079***	0.077***
	(0.005)	(0.005)	(0.005)	(0.005)
Controls	Yes	Yes	Yes	Yes
Fixed-effects				
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Fit statistics				
Observations	393,951	393,951	393,951	393,951

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.8: Robustness to Model Choice: This table depicts the primary analysis run on the 2nd–5th best performing Populism Detector models (alternate Neural Network architectures) with the Gemini embeddings and LLM-as-a-Judge labelings. Results are comparable to the primary results and remain significant.

A.5 Populism Detector: Loss Function

As noted in the main article, the loss function I use to select the optimal models is the average of three benchmarks: PR-AUC, class-balanced accuracy, and Brier score. In this Appendix, I examine how the choice of optimal models depends on the weights assigned to each of those components using a ternary (simplex) plot, **Figure A.4**. The result shows that the selection of embedding, estimation model, and labeler are highly stable, and alternative selections are also quite similar to the primary selection.

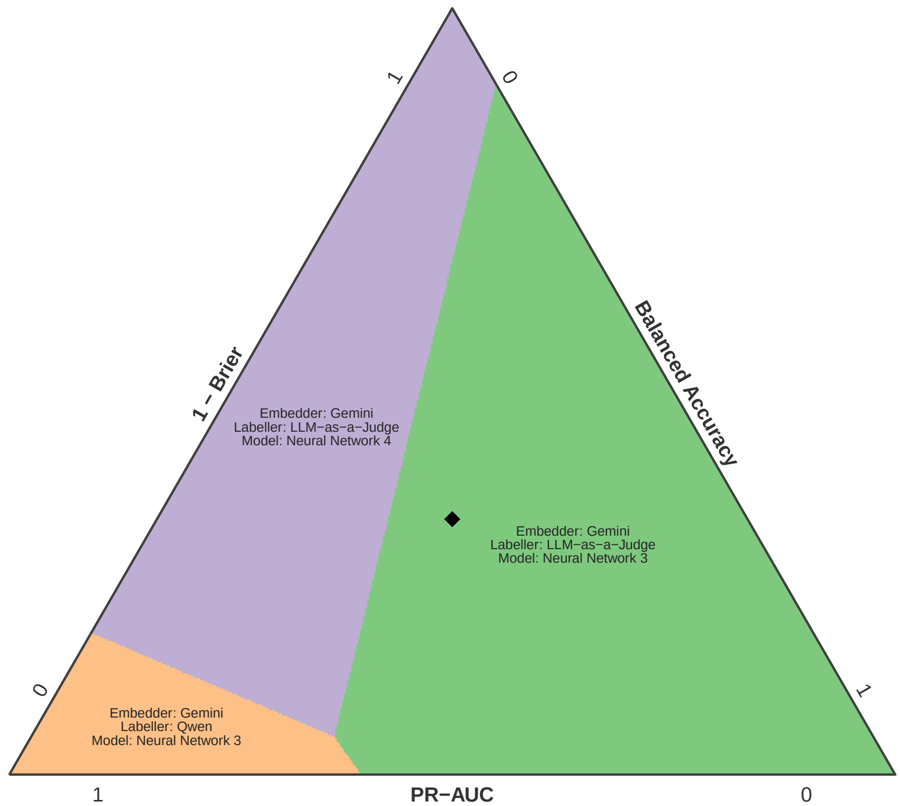


Figure A.4: Loss Function Sensitivity: This figure reports the winning embedder–labeler–model combination under every convex combination of the three constituent metrics of $\mathcal{L}$: PR-AUC, class-balanced accuracy, and $1 -$ Brier. The black diamond marks equal weights, the specification used in the main analysis. Across nearly the entire simplex the selected pipeline uses Gemini embeddings and the LLM-as-a-Judge labeler, with Neural Network 3 or Neural Network 4; a small region that heavily weights PR-AUC selects Qwen labels instead. The chosen pipeline is therefore similar under any convex combination of the constituent error functions, and the paper's substantive results hold under other "winners" (and, indeed, other "losers").

A.6 Populism Detector: Validation

A.6.1 Example High-Populism Speeches

In this appendix, I present abridged quotations from six speeches coded as high-populism, randomly sampled so that three of the selected speeches are by AfD legislators, two by CDU legislators, and one by a legislator from another party. I translate the speeches for the purposes of this appendix and annotate them in footnotes where required. All six reflect the themes and language expected of the RRP speech.

> Your migration policy can be summed up in a single sentence: 50% of eligible Hartz IV recipients are foreigners... this means: your entire migration policy is not producing skilled workers, but rather social welfare cases... Two-thirds of all Syrian asylum seekers still subsist on Hartz IV, and half of all German-Turks lack any vocational training... These people are turning our cities into slums... and are continuing to transform the entire country into one vast Illerkirchberg.[25]
> *Katrin Ebner-Steiner, AfD, December 13, 2022, Bavaria*

> Mr. Büttner, the people at the heart of this issue are criminals. They are criminals who exploit the plight of people from other countries—including EU nations. The reality is that these criminals seek to maximize their profits from dilapidated properties at the expense of the very people affected; for such dilapidated properties do indeed exist... This trend is just beginning in my hometown as well: entire residential quarters will ultimately undergo social decline. Eventually, no German will want to live there anymore. Moreover, it is undignified that foreigners are forced to live in such conditions... To drain this swamp, Mr. Büttner, we have now employed this specific [policy] instrument.[26] *Frank Scheurell, CDU, October 24, 2018, Saxony-Anhalt*

25. Hartz IV refers to a labor and benefit reform package adopted by Germany in 2005; Illerkirchberg is a small town in Baden-Württemberg, which the AfD has called attention to after two incidents of violent crime by perpetrators who entered Germany seeking asylum.

26. Büttner is an AfD politician; the speaker is adopting AfD rhetorical framing in many regards, but redirecting anger about slum housing towards property owners rather than migrants. The "drain the swamp" metaphor is present in the original German. The speaker's brother is an AfD politician.

In Berlin, over the past few days, we have witnessed further harbingers of the decay of the old parties. Merkel's decision to open the borders is acting like a catalyst for the decline of the former ”catch-all parties.” But how, indeed, can one claim to be a ”people's party” if one either rejects the very concept of ”the people” or seeks to redefine it?... Under pressure from the AfD, Horst Seehofer sought—at least in certain areas—to restore the rule of law and uphold existing statutes. Yet the worst Chancellor in post-war history—she alone—refused to allow this... Law, statutes, and order—these are simply no longer part of the CDU's political agenda... One frequently reads and hears of a detached political caste that treats itself differently than it treats those it is supposed to represent—the citizens, the people...[27] *Markus Wagner, AfD, July 13, 2018, North Rhine-Westphalia*

This “traffic-light” government is ruining our country at full throttle—at full throttle! For what is currently being done in Berlin can no longer be explained by common sense. Germany used to be an energy exporter. Now, under this “traffic-light” government, we have—entirely through our own doing—become an energy importer. What is the Federal Government attempting to do in the midst of this supply crunch? To ship coal from Ecuador... all the way around the world on dirty heavy-cargo vessels, only to burn it here at home. We are purchasing the most expensive fracked gas, shipping it from America to our shores on massive vessels—a process that entails a one-third loss in efficiency. This ”traffic-light” government couldn't care less—as long as the ideology is right!... Our own domestic capabilities for producing energy ourselves are being systematically destroyed. We have 10,000 biogas plant operators in Germany. Yet, this Federal Government is fighting against them with the utmost severity. Why?[28] *Steffen Vogel, CSU, February 7, 2024, Bavaria*

First the Dieselgate scandal, then the Diesel Summit — in both instances, politicians have unfortunately failed. Both events made it glaringly obvious how the automotive industry holds the political sphere in its grip and dictates its

27. Horst Seehofer is a prominent CSU politician who sparred with Angela Merkel on immigration and asylum issues; the speaker here is exploiting divides within the CDU/CSU coalition. The speech goes on to contrast the presence of security at the Landtag with the claimed insecurity of the border.

28. “Traffic-light” government refers to a coalition of the SPD, FDP, and Grüne parties. The speaker prevaricates between expressing climate change concern and also endorsing domestic energy sovereignty and nationalism, adopting a scattered approach in order to attack the ruling coalition.

actions—clearly demonstrating who truly possesses the power and leverage in this country. Ultimately, these developments underscore the disempowerment of all citizens and reveal that our parliamentary democracy is, in fact, in jeopardy. At the Diesel Summit, the automakers and their lobbyists once again succeeded in getting their way completely. Public health continues to be compromised, while owners of diesel vehicles are, in a sense, being dispossessed as their cars plummet drastically in value. Indeed, a great many jobs — both within the automotive industry itself and among its suppliers — continue to be put at risk, given the sheer volume of employment opportunities tied to these sectors. For many people, their very livelihoods depend on them.[29] *Andreas Höppner, Die Linke, August 25, 2017, Saxony-Anhalt*

In times of war and on the high seas, women and children come first. What we are witnessing in Germany, however, is the exact opposite: predominantly young men are arriving — as has been the case since 2016 — coming primarily from Africa and the Middle East. It is precisely this form of tourism, masquerading as a mass migration, that we must put an end to. Home should be a place of familiarity, where our children can grow up safe and carefree. That is why we conduct politics *of* the people and *for* the people—while you conduct politics *of* the system and *for* the system... How many more blatant insults—how many more metaphorical ”slaps in the face”—are we, as citizens, expected to tolerate from this government?... Pensioners from Ukraine receive basic old-age security here, despite having made zero contributions. This, too, is unacceptable. It constitutes a betrayal of the citizen and the taxpayer. *Oliver Kirchner, AfD, October 13, 2023, Saxony-Anhalt*

29. Dieselgate refers to an imbroglio wherein Volkswagen deliberately programmed the automotive computers of diesel vehicles to produce lower vehicle emissions *when being tested for emissions*, in order to cheat emissions standards. The speaker here is concerned about health and environmental impacts, but more concerned that the scandal will impact the common person and workers. The prompt to detect populism steers the labelers away from identifying mere labor politics as populism, but here, presumably, the conspiratorial framing (justified as it is by the facts) flags as populist.

A.6.2 Keyword and Topic Clustering of High-Populism Speeches

This appendix presents the top unigrams, bigrams, and trigrams from speeches classified as high populism ($\widehat{Pr(Y_i = 1)} > 0.9$). The selection criterion is the log-odds ratio between high populism and low populism speeches (whether the speech functions as a shibboleth or "wedge"). The result bodes well for concept validity: most major topics that one would expect (the EU; migration and refugee claims; Islam; "woke"; climate skepticism; COVID-19 skepticism; conspiratorial thought; bogeymen) are represented. Moreover, virtually none of the words reflect ordinary, non-populist speech, and many words reflect stance, not just subject or sentiment.

Unigram	Log-Odds	Bigram	Log-Odds	Trigram	Log-Odds
bürger [citizen]	28.31	milliarden euro [billions of euro]	16.51	rieten rieten grün [red red green] §	5.77
volk [the people]	23.53	öffentlich rechtlich [public law]	14.05	bürger unser lande [citizens of our land]	5.46
sogenannen ["so-called"]	15.61	etablieren politik [establish[ment] politics]	10.98	amadeu antonio stiftung ¶	5.32
illegal	15.51	illegal einwanderung [illegal migration]	10.98	hans wern sinn ¶	5.25
migranen	14.72	unser volk [our people]	10.62	euro pro tonnen [euros per ton]	4.96
merkel ¶	13.58	illegal migranen	10.47	franz josef straüssen ¶	4.96
steuerzahler [taxpayer]	13.36	sogenannen flüchtlingen ["so-called refugees"]	10.28	öffentlich rechtlich rundfunk [public broadcasting]	4.37
ausländer [foreigner]	13.34	sogenannen energiewend [... energy transition]	10.19	vollziehbar ausreisepflichtigen ausländer	4.17
altparteien [establishment party]	12.91	etablieren parteien [establishment parties]	9.74	freiheitlich demokratisch grundordnung †	4.12
angeblich [allegedly]	12.57	corona politik	9.12	sogenannen erneuerbaren energeinen [... renewables]	4.08
asylanen [asylum seeker]	12.46	politisch klasse	9.08	rieten grün gelb [red green yellow] §	3.71
ideologisch	12.35	europäisch union	9.05	black liv matter	3.61
massenmigration	12.30	offen grenzen [open borders]	8.84	caspar david friedrich ¶	3.61
corona	11.93	sächsisch bürger [Saxon people] ‡	8.78	hans jürgen papier ¶	3.51
gendersprachen [gender speech]	11.43	politisch korrektheit [political correctness]	8.37	minderjährigen unbegleiten flüchtlingen [minor refugees]	3.51
demokratie	11.42	euro steuergeld [euros of taxpayer money]	8.32	sogenannen demokratisch parteien	3.40
zuwanderung [immigration]	11.42	multikulturellen sellschafen [... societies]	8.32	sämtlich corona beschränkungen [COVID restrictions]	3.40
windindustrieanlagen [wind turbine]	11.34	politisch islommen [political islam]	8.26	öffentlich rechtlich fernsehen [public television]	3.40
islommen	11.19	politisch korreken [political correctness]	8.20	politisch medial komplex [politician-media complex]	3.29
steuergeld [tax revenue]	10.95	sogenannen klimaschutz [... climate protection]	8.07	schwarz rieten gold [black red yellow] §	3.29
multikulti	10.94	national opposition	8.01	sicheren drittstaeten einreisen [... safe third-countries]	3.29
grundgesetz [Basic Law] †	10.83	rechtlich rundfunk [public broadcasting]	7.63	allahu akbar rufen [shouting Allahu Akbar]	3.06
freiheit [freedom]	10.81	ausreisepflichtigen ausländer [deportable...]	7.37	katrin görehen eckarden ¶	3.06
kriminellen	10.45	rechtlich rundfunk [public broadcasting]	7.63		
scharia [sharia]	10.30	unkontrollieren masseneinwanderung	7.10		
grenzen [border]	10.26	demokratisch grundordnung [... basic law] †	7.05		
links [left]	10.13				
masken	9.95				
abschieben [deport]	9.69				

† Basic law / basic order here generally referring not to the constitution as a whole, but an underlying principle of the constitution which bans extremist left and right political parties; speakers oppose such designations for (plausibly justified) fear that they will be used to ban their party.
‡ Saxon people here meaning people from Saxony, lacking the racial valence typical in English.
§ Color patterns refer to political coalitions according to party traditional colors.
¶ Various public or historical figures, some referenced disparagingly, others approvingly; Friedrich is the painter of the same name.

Table A.9: Top N-Grams for High Populism Speeches: This table depicts the top unigrams, bigrams, and trigrams that characterize high-populism speeches and not low-populism speeches by log odds of the ratio of their use. Some terms are removed due to redundancy (e.g., "illegal migranen", "illegal zuwanderung", etc. all being suppressed from the bigram list for functional duplication of "illegal einwanderung"), or no meaning without broader context (e.g., "einzig politisch krafen", the only political force). Words were pre-processed by a German lemmatizer before calculation of log-odds ratios.

A.6.3 High-Populism Topic Model Analysis

Labeling topics discovered by a topic model is more of an art than a science: it represents a significant researcher degree of freedom. Since the validity of the populism detector is that the topics in high-populism speech reflect a complete understanding of (radical right) populism, the labels should unambiguously follow from the words and speeches here. In **Table A.10**, I list the top $n = 50$ words characteristic of each topic, along with the most characteristic speech of the topic. The table lists the words in descending order of distinctiveness for the topic and is morphologically grouped. For ease of interpretation (this choice has no bearing on the model estimation), I color-code words whose relevance to the topic is low in gray and words whose relevance is highest in red.

Topic Num	Prob.	Topic Label	Characteristic Data
1	0.11	Public Broadcasting, Funding, Propaganda	rundfunk (rundfunkbeitrag, rundfunkrat, rundfunkanstalten, rundfunksystem), sender (rbb, wdr, zdf, mdr), staatsfunk (staatsfern), beitrag (zwangsbeitrag), medienstaatsvertrag, haushalt (haushaltsentwurf, haushaltsplan, landeshaushalt, doppelhaushalt), rechnungshof (landesrechnungshof), einzelplan, intendant, radio, berichterstattung, finanzierung, gehälter, ausgabe, freistaat, staatskanzlei, thüringer, ramelow, hoff, awo, vetternwirtschaft, petitionsausschuß, transparenz, entwurf (gesetzentwurf), änderungsantrag, regierungsfraktion, kommunal, finanzausschuß, monitor, beratung, rechtlich, landesprogramm, gez "Das Vertrauen der Bürger in den öffentlich-rechtlichen Rundfunk ist weitgehend erodiert; eine bloße schrittweise Reformierung reicht nicht aus, weil die bestehenden Gremien und Anstalten den Auftrag verfehlen." *André Barth, AfD, Saxony, February 09, 2024*

Topic Num	Prob.	Topic Label	Characteristic Data
2	0.12	Asylum, Migration,nand Border	asyl (asylantrag, asylverfahren, asylbewerber, asylrecht, asylpolitik, asylbewerberleistungsgesetz), einwanderung (einwanderungsland, einreise, einreisen, zuwanderung, zuzug, zustrom), ausreise (ausreisepflicht, ausreisepflichtig, ausreisepflichtige), abschiebung (rückführung), flüchtling (kriegsflüchtling, flüchtlingspolitik), migration (migranten), unterbringung (unterkunft), aufenthalt (aufenthaltsrecht), herkunftsstaat (herkunftsländer), drittstaat, heimatland, familiennachzug, bleiberecht, unbegleitet, ausländerbehörde, bamf, ukrainer, illegal, fachkraft, sozialsystem, sozialleistung (sachleistung, geldleistung), pakt, bürgergeld, mittelmeer, vollziehbar, abgelehnt "Die Ausländerpolitik in Deutschland und Niedersachsen ist gescheitert; seit Jahren werden offensichtliche Fehlsteuerungen nicht korrigiert, obwohl Aufnahme- und Betreuungskapazitäten begrenzt sind und die Interessen der Bürger geschützt werden müssen." *Stephan Bothe, AfD, Lower Saxony, October 27, 2023*
3	0.07	COVID-19 and anti-Lockdown	impfung (impfen, impfpflicht, impfstoff, impfzwang), corona (coronavirus, covid, coronamaßnahmen, coronapolitik), lockdown (lockdowns, lockerung), maske (maskenpflicht), inzidenz (inzidenzwert), test (testen, getestet, pcr), infektion (infektionsgeschehen, infiziert, infektionsschutzgesetz), ungeimpft (ungeimpfte), geimpft (geimpfte), rki, virus, pandemie, patient, risikogruppe, intensivbette (intensivstation), klinik (krankenhaus), erkrankung, übersterblichkeit, nebenwirkung, einrichtungsbezogen, pflegeheim, arzt, lauterbach, söder, spahn, grippe, gastronomie, grundrechtseinschränkung "Laut WHO kann ein PCR-Test allein keine Infektion feststellen; bei positiven Ergebnissen ohne Symptome sei eine Wiederholung nötig. Trotzdem wird politische Kontrolle über Impfungen und Maßnahmen weiter ausgebaut." *Robert Farle, AfD, Saxony Anhalt, February 05, 2021*
4	0.13	Energy, Nuclear Phase-out Anti-Green Energy	energie (energiepreise, energiekost, energieversorgung, energiepolitik, energiepolitisch, energieträger), strom (strompreis, strompreise, kilowattstunde), kernkraft (kernkraftwerk, kernenergie, atomkraftwerk, reaktor), kohle (kohlekraftwerk, kohleausstieg), energiewende, gas (gaskraftwerk), eeg, erneuerbaren, fossil, ausstieg, heizung, wärmepumpen, mobilität, verkehrswende, verbrennungsmotor, elektroauto, auto, automobilindustrie, ausstoß, industrie, bezahlbar, inflation, cent, klimapolitik, planwirtschaft, umlage, thyssenkrupp, pinkwart, china, miete, günstig, kraftwerk "Nach Abschaltung der Kernkraftwerke stieg der CO2-Ausstoß pro Kilowattstunde stark; die zwanzigjährige Klimapolitik habe Energie verteuert und Versorgungssicherheit gefährdet, statt bezahlbaren und verlässlichen Strom zu liefern." *Christian Loose, AfD, Nrw, December 08, 2022*

Topic Num	Prob.	Topic Label	Characteristic Data
5	0.08	Violent Crime and Policing	straftat (straftäter), täter (tatverdächtig, tatverdächtige, gewalttäter), polizei (polizeibeamte, polizeibeamter, polizist, polizeilich), kriminalität (kriminalstatistik, kriminell, kriminelle, ausländerkriminalität, clankriminalität), gewalt (gewalttat, gewalttätig), delikt, übergriff, körperverletzung, vergewaltigung (gruppenvergewaltigung, vergewaltigen), messer (messerangriff, messerattacken), terror (terroranschlag, terrorist), anschlag, gefährder, mord, opfer, justiz, staatsanwalt, sicherheitslage, clan (clans), reul, cottbus, mannheim, silvesternacht, linksextremist, linksextremismus, verüben, begehen, poliziste "Angriffe auf Einsatzkräfte sind keine Einzelfälle mehr; allein in Hessen wurden 2019 über 4.000 Polizeibeamte Opfer. Rechnet man Feuerwehr, Rettungsdienste und Zoll hinzu, steigt die Zahl auf über 7.000." *Uwe Junge, AfD, Rhineland, March 22, 2018*
6	0.11	Anti-EU, Bailouts, and Fiscal Sovereignty	europa (europäisch), eu (brüssel, brüsseler, mitgliedsstaat), währung (mark), bank (landesbank, sparkasse, zentralbank, ezb), souveränität (nationalstaat, national), esm, griechenland, brexit, lissabon, pleitestaat, kapitalismus, finanzmarkt, vertrag, union, juncker, zypern, arbeitnehmer, löhne, mindestlohn, rentenversicherung, arbeitslosengeld, altersarmut, sparer, empfänger, versichert, vereinigt, sellering, holter, nieszery, schäuble, tillich, brd, nato, bundeswehr, amerikaner, amerikanisch, brite, polnisch, hartz, rent "Die Wiedereinführung nationaler Währungen werde bereits ernsthaft diskutiert; wenn der Euro kollabiert, bleibe kaum eine Alternative. Deutsche Steuergelder dürften nicht länger eingesetzt werden, um insolvente Eurostaaten zu stützen." *Stefan Köster, NPD, Mecklenburg, June 10, 2010*
7	0.16	Elections, Anti-Establishment, Extremism Designation	wahl (wählen, wähler, wahlergebnis, wahlkampf), demokratie (demokrat, demokratisch, undemokratisch, demokratieverständnis), opposition (oppositionspartei), partei (bürgerlich), meinungsfreiheit, kandidat, parlamentarisch, parlament, diktatur, protest, diskreditieren, gegner, bürgermeister, fraktionsvorsitzende, sed, stasi, ddr, nazi, striegel, correctiv, huber, mauer, leute, draußen, hetz, kriegen, schämen, gucken, hund, dumm, zuruf, merken, meinung, arroganz, anstand, gesagen, blöd, bißchen, hinstellen, demonstrieren "Das Wahlergebnis zeigt laut Rede eine klare Deklassierung der SPD; die Regierung habe die Bürger moralisch unter Druck gesetzt, um Wahlentscheidungen zu steuern, und verliere damit demokratische Legitimation." *Dirk Nockemann, AfD, Hamburg, June 12, 2024*

Topic Num	Prob.	Topic Label	Characteristic Data
8	0.10	'Gender Ideology' Poliical Correctness (Schools)	gender (gend, gendern, gendergerecht, genderideologie, gendersprache), schule (schüler, schülerin, schulsystem, grundschule), bildung (bildungspolitik, bildungssystem, bildungspolitisch), lehrer (lehrkraft, lehrermangel, lehrplan, unterrichten), hochschule (studium, student, studierende, studiengang, studiengebühr, universität), geschlecht, sprache (sprachlich), erziehung (erziehen), gleichstellung, mainstreaming, gymnasium, unterricht, kita, familienpolitik, frühkindlich, inklusion, kind, eltern, maskulinum, generisch, bologna, pisa, kunst, mutter, abitur, pädagogisch "Die Gendersprache habe den Alltag verändert und nehme immer groteskere Züge an; unter dem Vorwand der Nichtdiskriminierung werde Sprache politisch umgebaut und als pädagogisches Instrument zur gesellschaftlichen Umerziehung genutzt." *Anne Cyron, AfD, Bavaria, July 09, 2020*
9	0.08	Islam and Muslims	islam (islamisch, islamismus, islamisierung), muslime (muslimische, muslimischen, moslems), moschee, kopftuch (vollverschleierung, burka), koran (scharia, allah), ditib, izh, imame, minarette (minarett), religion (religionsfreiheit, religionsunterricht, religionsgemeinschaft, religiös), antisemitismus (antisemitisch), jude (jüdisch), türkisch (türke), erdogan, leitkultur, grundgesetz, verfassung (verfassungsänderung), grundordnung, christlich (christ, kirche), israel, hamas, unterdrückung, ungläubige, volksverhetzung, multikulturell, holocaust, flagge, symbol, iran "‚Solange wir in der Minderheit sind, akzeptieren wir eure Rechtsordnung': Diese Haltung belege, dass islamistische Akteure den Koran über das Grundgesetz stellen und damit zentrale Prinzipien freiheitlicher Ordnung infrage stellen." *Arne Schimmer, NPD, Saxony, October 17, 2012*
10	0.06	Agricultural Policy and GMOs	landwirtschaft (landwirtschaftlich), landwirt, agrarpolitik, gentechnik (gentechnisch), anbau, bauer (bäuerlich), acker, saatgut, futtermittel, hektar, fläche, natur, umwelt (umweltschutz), ökologisch, naturschutz, biodiversität, artenschutz, insekt, tier (tierhaltung), ländlich, kulturlandschaft, lebensmittel, windkraft (windkraftanlage, windenergie, windenergieanlage, windpark, windrad, windindustrieanlage), fahrverbot, meter, bvvg, verändert, monsanto, grenzwert, amflora, infraschall, milch, wolf (wölfe), anlage, backhaus, wald, landschaft "Das angestrebte Anbauverbot für gentechnisch veränderte Kartoffeln wird als notwendiger Schutz von Landwirtschaft, Umwelt und Verbrauchern dargestellt; agrarpolitische Entscheidungen dürften nicht Konzerninteressen überlassen werden." *Raimund Frank Borrmann, NPD, Mecklenburg, April 29, 2010*

Table A.10: Topic Model Details: This table lists the topics found by the high-populism speech topic model, the most characteristic words, and quotes an example speech that loads high on the topic. Words are arranged in descending order of distinctness. Morphologically related words are parenthesized. Red words establish the core meaning of the topic label, gray words are less relevant.

A.6.4 Full Corpus Speech Examples

In this appendix, I present abridged quotations from three speeches on the "Economy, Taxes, and Banks" topic and three speeches on the "Energy, Climate" topic as representative examples of what such speeches look like. I first subset to speeches whose modal assigned topic (discussed above in the *Speech Affected* section of the main paper analysis) are these topics. I then sample randomly to produce a stratified sample of low, medium, and high populism speeches in those topics. I excerpt the speech's core point but provide the necessary speech metadata for readers to read the full speech. I translate the speeches in this appendix.

Speeches with Modal Topic "Economy, Taxes, and Banks"

> Against the backdrop of the financial crisis we have just experienced—during which we saw that size does not necessarily guarantee success — I would like to ask you, as someone who is not an expert on economic policy, to explain in detail the advantages of an insurance company that is significantly larger than the current structure: where do you see the margin benefits, and how might this increase market share within the service area? *Stefan Wenzel, Grüne, November 24, 2010, Lower Saxony*; **Topic:** Economy, Taxes, Banks (p=0.414); **Populism:** Low (p=0.004)

> The *DDR* [East Germany] had 12 billion euros in debt... They left behind a country that, in many respects, can only be described as rotten... Then, Mr. Kuschel, there is the myth that tax cuts ruin public finances... It is not the tax rate that is the decisive factor, but the actual tax revenue... We want to pursue a tax policy that... eases the burden on low and middle incomes in this country. *Uwe Barth, FDP, April 26, 2013, Thuringia*; **Topic:** Economy, Taxes, Banks (p=0.131); **Populism:** Medium (p=0.486)

> When one in three euros paid out in pensions today is funded by taxes rather than contributions, it is entirely appropriate to speak of a crisis... Western societies

are societies of death, and the book title ... ”Germany Abolishes Itself” captures this precisely... We demand the exclusion of foreigners from the German social security system... Germans pay for Germans. *Andreas Storr, NPD, September 1, 2010, Saxony*; **Topic:** Economy, Taxes, Banks (p=0.267); **Populism:** High (p=0.975)

‘**Speeches with Modal Topic “Energy, Climate”**

Two and a half years ago, the question of where to locate a final nuclear waste repository was completely open, and a 36-year extension of operating life was planned for the three sites in Schleswig-Holstein alone... You have to view the current stance — ”We don’t want this shit” — in that context... taking comprehensive responsibility... whether to shut down and establish a plan for the future, or to proceed with expansion without having anything in place for the future. *Heiner Rickers, CDU, April 24, 2013, Schleswig-Holstein*; **Topic:** Energy, Climate (p=0.231); **Populism:** Low (p=0.002)

These measures [hydrogen fuel-cell buses and national charging infrastructure] remind us of a statement by the former President of the European Commission...: “We decide on something, put it out there, and wait a while to see what happens. If there is no major outcry” — and there is very little outcry in Germany — “and no uprisings” — even fewer of those; for who takes to the streets these days? Because most people do not even grasp what has been decided, “then we proceed... until there is no turning back...” [W]e cannot consent to far-reaching measures regarding energy generation and distribution as long as the fundamental question remains unresolved... Current alternative energy sources, such as wind and solar, are insufficient. *Erich Heidkamp, AfD, March 16, 2021, Hesse*; **Topic:** Energy, Climate (p=0.486); **Populism:** Medium (p=0.332)

Colleges and universities across the country – and naturally in Hesse as well — have largely degenerated into ideological projects for the leftist elite. Gender studies, restrooms for 186 imaginary genders, and so-called “Anti-Colonial Days” are turning universities into ideological hubs for cultural Marxists. In your motion... that the ”energy and heating transition in higher education” be accelerated in light of the climate and energy crisis.... You demand... that university buildings be equipped with heat pumps. First of all: for reasons of heritage preservation alone, this proposal is utter madness. That is precisely your modus operandi: testing just how far the majority of society is willing to play along with your Green propaganda projects. Then, when the house of cards collapses, you backtrack and sheepishly admit that it was "only a test." The good news is that fewer and fewer people — and especially fewer within your own ranks—are falling for this ploy. *Jochen K. Roos, AfD, July 9, 2024, Hesse*; **Topic:** Energy, Climate (p=0.194); **Populism:** Medium (p=0.977)

A.7 Confounding Variables

A.7.1 Literature on Demand for RRP Parties

The demand for RRP is a well-studied, though contested, subject (Berman 2021). I need not resolve the debate here. Rather, I need only observe that the factors identified are probable confounders of the relationship between legislative entry and RRP speech and adjust for them accordingly.

In earlier decades, RRP support was viewed to be a type of "protest vote" (Westle and Niedermayer 1992). Today, the majority of research indicates that support for RRP is, on some level, sincere and emerges under economic threat, including trade-related de-industrialization (Colantone and Stanig 2018; Margalit 2011), austerity in response to economic shock (Baccini and Sattler 2025), supply shocks to the local labor force due to immigration (Alrababah et al. 2025), household precarity (Abou-Chadi and Kurer 2021), low income (Andersen and Zimdars 2003), and a sense of local decline (Arzheimer et al. 2024). Others emphasize the role of more acute crises — whether refugee (Dinas et al. 2019) or pandemic (Lehmann and Zehnter 2024) — and how RRPs make a strategic choice to lean into the crisis in response (Pirro and Van Kessel 2017).

The primary dissenting views come from those who see the roots of support as fundamentally attitudinal (Hansen and Olsen 2019) or post-material in the sense of Inglehart (1977). These include authors who characterize RRP support as driven by sexism (Anduiza and Rico 2024), group identity threat (Sides, Tesler, and Vavreck 2019), anti-immigration attitudes (Hansen and Olsen 2024; Rydgren 2008), resentment toward the cultural accommodation of minorities (Bustikova 2014; Shehaj, Shin, and Inglehart 2021), nostalgic deprivation (Gest, Reny, and Mayer 2018), cultural protectionism (Oesch 2008), social atomization (Betz 1993), or big-five personality psychology traits (Bakker, Schumacher, and Rooduijn 2021). The answer may even be that the underpinnings of support are either highly parochial (Arzheimer 2009) or all of the above (Halikiopoulou and Vlandas 2020).

Institutional features, including population and demographic sorting (Dancygier et al. 2025), electoral system design (Becher, Menéndez González, and Stegmüller 2023; Jackman and Volpert 1996), public financing of elections (Bichay 2020), federalism (Pérez 2025), and the spatial positioning of parties (Meguid 2005), magnify baseline electoral support for RRPs (Norris 2005).

A.7.2 Confounding Variable Sourcing

Variable	Source	Years
	Population	
population	Eurostat `demo_r_pjanaggr3`	2000–2025
	Economic Indicators	
gdppc	Eurostat `nama_10r_2gdp` (GDP per capita)	2000–2024
unemployment ∗	Eurostat `lfst_r_lfu3rt`	2000–2024
poverty_rate †	Statistische Ämter des Bundes und der Länder (Sozialberichterstattung)	2005–2024
	Sectoral Employment Rate	
employment_primary	Eurostat `lfst_r_lfe2en2` (NACE A)	2008–2024
employment_secondary	Eurostat `lfst_r_lfe2en2` (NACE B–E)	2008–2024
employment_service	Eurostat `lfst_r_lfe2en2` (NACE other)	2008–2024
	Educational Attainment Rate	
edu_secondary	Eurostat `edat_lfse_04` (ISCED 3–4, 25–64, %)	2000–2024
edu_tertiary	Eurostat `edat_lfse_04` (ISCED 5–8, 25–64, %)	2000–2024
	Demography	
pct_foreign_born	Eurostat `lfst_r_lfsd2pwc` (country of birth, %)	2000–2024
pct_noneu	Eurostat `lfst_r_lfsd2pwc` (non-EU27 born, %)	2017–2024
	Crime	
murder_rate ‡	Bundeskriminalamt PKS (pre-2012, 2020–2022, 2024); Eurostat `crim_gen_reg` ICCS0101 (2012-2019, 2023)	2000–2024
serious_assault ‡	Bundeskriminalamt PKS (pre-2012, 2020–2022, 2024); Eurostat `crim_gen_reg` ICCS02011 (2012–2019, 2023)	2000–2024

∗ **Unemployment**: Two states are missing from 2020 in Eurostat. I source data from the *Bundesagentur für Arbeit*, create a crosswalk between the two sources, and impute the missing states.

† **Poverty**: Methodology for calculating this variable changed in 2020 (pre-2020 uses a national benchmark to calculate local poverty, post-2020 uses a local benchmark).

‡ **Crime:** Pre-2012 data, as well as 2020, 2021, 2022, and 2024 are from the Polizeiliche Kriminalstatistik, BKA (Mord und Totschlag; Gefährliche und schwere Körperverletzung); remaining years from Eurostat `crim_gen_reg` (ICCS0101, ICCS02011). Data sources are bridged by calculating region-specific ratios based on 2012 data.

Table A.11: Codebook: This table provides sourcing and construction details for the state-year level confounders included in model specifications.

A.7.3 Confounder Inclusion and Missingness

Dependent Variable:	Predicted Populism			
Sample	All Parties		CDU/CSU	
Controls	None	Imputed	None	Imputed
Model:	(1)	(2)	(3)	(4)
Variables				
AfD Intensity	0.115***	0.097***	0.073***	0.068***
	(0.004)	(0.006)	(0.009)	(0.011)
Controls		Yes		Yes
Fixed-effects				
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Fit statistics				
Observations	393,951	393,951	96,646	96,646

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.12: Results without Covariates or with Multiply-Imputed Covariates: These models establish that the estimates are not strongly sensitive to the included controls, or to handling control variable missingness by linear interpolation. Columns 2 and 4 use multiple imputation to fill in missing control variables.

A.8 Counterfactual Estimators

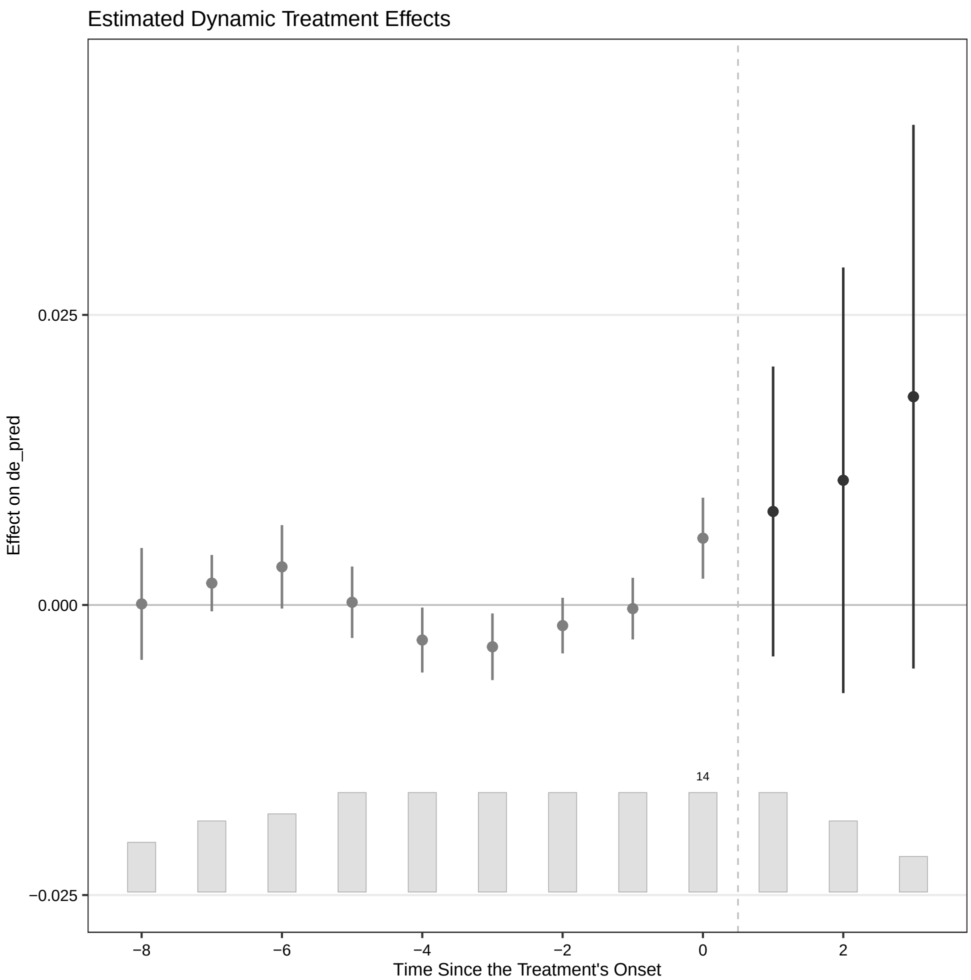


Figure A.5: Counterfactual Estimator: This effect plot shows the result of running estimating the treatment effect using counterfactual estimators, which use more conservative (lower-power). This estimator class requires a binarized treatment variable (rescaling the treatment coefficient). As in the TWFE event study, the model finds small negative anticipation effects, and consistently positive treatment effects. The counterfactual estimator relies on comparing treated units to strictly untreated units; because all units are treated within 4 years, the rightmost effect estimates are underpowered and non-significant. Under quarterly time cadence, the same results obtain.

A.9 Robustness to Functional Form

A.9.1 Binary or Binned Treatment

Dependent Variable:	Predicted Populism			
Sample	All Parties		CDU/CSU	
Treatment	Binary	Binned	Binary	Binned
Model:	(1)	(2)	(3)	(4)
Variables				
AfD Presence	0.010***		0.011***	
	(0.001)		(0.002)	
AfD Intensity (Tercile)		0.008***		0.008***
		(0.000)		(0.001)
Controls	Yes	Yes	Yes	Yes
Fixed-effects				
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Fit statistics				
Observations	393,951	393,951	96,646	96,646

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.13: Results with Binned Treatment: This table repeats the main analysis with a binned treatment variable. The primary model assumes that treatment is additive in the degree of AfD presence in the Landtag. These models establish that the effect does not depend on the functional form of the treatment. Columns 1 and 3 binarize the treatment (a state is treated if any AfD members serve in the Landtag). Columns 2 and 4 bin the treatment into terciles (No Treatment, Low, Medium, High). Results suggest that AfD entry causes a positive shock to speech populism regardless of how treatment is measured.

A.9.2 Binary Dependent Variable

Dependent Variable:	High Populism?			
Sample	All Parties		CDU/CSU	
Threshold	75th %ile	Fixed: 0.2	75th %ile	Fixed: 0.2
Model:	(1)	(2)	(3)	(4)
Variables				
AfD Intensity	0.345***	0.170***	0.260***	0.126***
	(0.018)	(0.011)	(0.036)	(0.023)
Controls	Yes	Yes	Yes	Yes
Fixed-effects				
State FEs	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes
Fit statistics				
Observations	393,951	393,951	96,646	96,646

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.14: Results with Dichotomized Dependent Variable: This table repeats the main analysis with a dichotomized dependent variable. To the extent that the Populism Detector produces invalid predicted probabilities but still consistently discriminates between high- and low-populism speeches, this alternative model should confirm the robustness of the primary finding. Columns 1 and 3 consider a speech to be high-populism if it is above the 75th percentile of predicted populism. Columns 2 and 4 consider a speech to be high-populism if it is above a fixed threshold of $\widehat{Pr(Y_i)} = 0.2$. In all cases, the treatment effect remains positive and significant.

A.9.3 Per-State Sensitivity

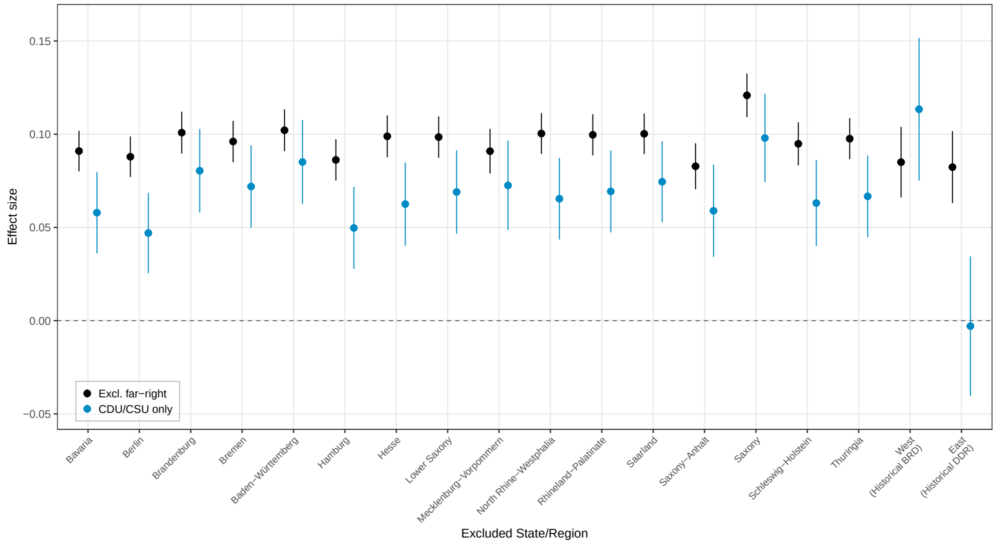


Figure A.6: Per-State and Per-Region Sensitivity: This plot depicts the results of running the main model, sub-sampling by excluding each state in turn (leave-one-out), and then excluding each region in turn. The all-parties results are robust to any such exclusions. The CDU-CSU effect is being driven by increased populism in the former-*DDR* (East German) states and is non-significant and near 0 without those states included as a group.

A.9.4 Alternative Fixed Effect Structures

Dependent Variable:	Predicted Populism All Parties						
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables							
AfD Intensity	0.098***	0.100***	0.101***	0.078***	0.085***	0.086***	0.082***
	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects							
State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-quarter FEs	Yes						
Year-week FEs		Yes					
Year FEs			Yes	Yes	Yes	Yes	Yes
Party FEs			Yes				
Speaker FEs				Yes			
Region (East/West) × Year FEs						Yes	
Varying Slopes							
Year FEs (State FEs)					Yes		
Year FEs (Region (East/West))							Yes
Fit statistics							
Observations	393,951	393,951	393,951	393,750	393,951	393,951	393,951

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.001*

Table A.15: Sensitivity to Fixed Effects Structure (All Parties): This table repeats the primary sample while varying the fixed-effect structure. Column (1) uses state and year-quarter fixed effects; column (2) uses state and year-week fixed effects; column (3) uses state, year, and party fixed effects (to rule out compositional party effects); column (4) uses state, year, and speaker fixed effects; column (5) adds a state-specific linear time trend in calendar year; column (6) adds region (East/West) × year fixed effects; and column (7) adds region-specific linear time trends in year. The primary results are not strongly impacted by the form of the fixed effects.