Beijer Institute
OF ECOLOGICAL ECONOMICS

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

DISCUSSION PAPER

Speech-Based Measurement of Polarization Using Text-as-Data

Beijer Discussion Paper Series No. 285

Engström, G., S. Axelsson, J. Gars, S. Källman,
and T. Lindahl. 2026.

Beijer Institute of Ecological Economics
The Royal Swedish Academy of Sciences

Lilla Frescativägen 4A
Stockholm

+46 8 673 95 00
beijer@beijer.kva.se

www.beijer.kva.se

Speech-Based Measurement of Polarization Using Text-as-Data

Gustav Engström Sofia Axelsson Johan Gars
Simon Källman Therese Lindahl

Wednesday 21st January, 2026

Abstract

Political polarization research depends on valid measurement of policy positions, yet established data sources—roll-call votes, elite surveys, expert judgments—are limited in temporal coverage and substantive scope. Parliamentary speech offers an alternative: abundant, publicly available text that reflects real-time political positioning across all policy debates. However, whether speech-based measures recover the same latent constructs as survey instruments remains an open empirical question, particularly given concerns about systematic reduction in observed variance in LLM-based measurement. This article validates speech-based measurement of policy positions and polarization using large language models (LLMs) to classify parliamentary speech from Swedish MPs (1998–2022) into survey response categories from Riksdagsundersökningen, a biennial elite survey. Employing retrieval-augmented generation (RAG) and multi-model validation, we assess construct validity through correlation analysis, rank-order agreement, and comparison to manual annotation. Results demonstrate strong construct validity: speech-based party–year–topic means achieve Spearman correlations exceeding $\rho = 0.85$ with survey benchmarks, correctly rank-order parties on most policy dimensions (Kendall's $\tau > 0.78$), and reproduce temporal trends in polarization. However, systematic positive bias (speech-based positions are more extreme than survey responses) and issue-specific performance variation indicate that speech and surveys measure overlapping but distinct constructs—public signaling versus private attitudes. This divergence reflects substantive political dynamics (party discipline, strategic ambiguity, audience targeting) rather than measurement failure. Speech-based polarization indices (Dalton index, bloc distance) closely track survey-based estimates, capturing known features of Swedish party competition including left-right bloc structure and issue-specific temporal dynamics. The findings demonstrate that parliamentary speech can serve as a valid data source for measuring policy positions when appropriately validated, expanding methodological capacity for historical and comparative research where survey data are unavailable. We distinguish measurement validity (correlation with external criteria) from construct equivalence (identical

latent concepts), clarifying that method divergence may reflect what is measured rather than how well it is measured.

1 Introduction

Measuring political polarization requires valid indicators of policy positions. Traditional approaches rely on structured data—roll-call votes, survey responses, expert placements—that are reliable but limited in scope and temporal coverage. Roll-call votes, while precise, are restricted to legislative settings and may not reflect sincere preferences due to party discipline and agenda control. Elite surveys provide direct attitude measurement but are administered infrequently and subject to non-response bias. Expert judgments offer comprehensive coverage but introduce subjective interpretation and may lag substantive political change.

Parliamentary speech represents an alternative data source: it is abundant, publicly available, and temporally granular, covering all policy debates and reflecting real-time political positioning. Yet speech-based measurement faces fundamental validity challenges. Unlike structured responses, speech is unconstrained in content and framing, embedding positions within rhetorical strategies, audience considerations, and party messaging. An MP may speak ambiguously, avoid contentious topics, or signal strategically—behavior that complicates position extraction. Whether speech can serve as a valid measurement instrument for policy attitudes thus depends on systematic empirical validation, not theoretical assumption.

Recent advances in large language models (LLMs) enable scalable text analysis, prompting renewed interest in automated content classification for political research (Ziems et al. [2024], Argyle et al. [2023],Wang [2023]). However, applying LLMs to political measurement raises methodological questions that extend beyond technical performance. First, do LLM-based classifications of speech recover the same latent construct as established survey instruments? Second, when speech-based and survey-based measures disagree, does this reflect measurement error or substantively meaningful differences between public signaling and private attitudes? Third, can automated coding replicate the construct validity of manual annotation without introducing systematic distributional bias (Bisbee et al. [2024])?

This article addresses these questions through a comprehensive validation study comparing speech-based and survey-based measures of policy positions among Swedish MPs. We employ retrieval-augmented generation (RAG) (Arslan et al. [2025]) to map parliamentary speech to survey response categories from Riksdagsundersökningen (RDU), a biennial elite survey covering 1998–2022. Validation proceeds through three complementary strategies: (1) correlation analysis assessing monotonic agreement between speech-based and survey-based party–year–topic means; (2) rank-order comparisons verifying that models correctly order parties along policy dimensions; and (3) manual annotation of a subsample to establish intercoder reliability between human and automated

classifications.

Our approach treats LLMs as measurement instruments, not autonomous political evaluators. The models do not assess whether policies are normatively desirable; instead, they classify the position an MP has expressed in observed speech, operationalizing the same coding rules a trained human coder would apply. Validity thus depends on empirical performance—correlation with external criteria, replication of known political patterns, and stability across model specifications—not on theoretical claims about model capabilities or limitations. This framing situates the analysis squarely within political methodology: we evaluate construct validity, criterion validity, and discriminant validity using established standards from survey research and measurement theory.

We focus on four policy issues—nuclear power, NATO membership, defense spending, and immigration/refugee policy—that span economic, security, and cultural dimensions of Swedish politics. These issues exhibit varying levels of party-system polarization, temporal salience, and cross-cutting cleavages, enabling assessment of whether validation results generalize beyond single-issue contexts. Temporal coverage (seven survey waves spanning 24 years) allows us to test whether speech-based measures track changes in polarization over time, a critical criterion for dynamic political analysis.

Results demonstrate that speech-based estimates using GPT-4.1 and GPT-5-mini achieve strong construct validity: Spearman rank correlations with RDU survey means exceed $\rho = 0.85$ ($p < 0.001$), and models correctly rank-order parties on most policy dimensions (Kendall's $\tau > 0.78$). Speech-based polarization indices (Dalton index, left-right bloc distance) closely approximate survey-based estimates, tracking temporal trends and structural cleavages in Swedish party competition. However, systematic positive bias (speech-based positions are consistently more extreme than survey responses) and issue-specific performance variation (strongest on nuclear power, weakest on defense spending) indicate that speech and surveys measure overlapping but distinct constructs. This divergence likely reflects party discipline, strategic ambiguity, and audience-targeted framing in parliamentary speech—substantive political dynamics, not measurement failure.

The article contributes to political methodology in three ways. First, it provides empirical evidence that parliamentary speech can serve as a valid data source for measuring policy positions and polarization, provided appropriate validation. This expands the methodological toolkit for historical and cross-national research where survey data are unavailable or incomplete. Second, it demonstrates that automated text classification using LLMs can replicate construct validity of manual coding at scale, enabling analysis of large speech corpora infeasible for human annotation. Third, it distinguishes measurement validity (do speech-based estimates correlate with external criteria?) from construct equivalence (do speech and surveys measure the same thing?), clarifying that divergence between methods may reflect theoretical distinctions in what is being measured rather than technical limitations of either approach.

The remainder of the article proceeds as follows. Section 2 develops the conceptual framework distinguishing latent attitudes (measured by surveys) from

speech-implied positions (measured by parliamentary discourse), and situates LLMs as coding instruments within established text-as-data methodology. Section 3 describes the data sources (RDU surveys and parliamentary speech) and their temporal alignment. Section 4 details the retrieval-augmented classification procedure and multi-model validation strategy. Section 5 presents validation results, including correlation analysis, rank-order agreement, and comparison to manual annotation. Section 6 applies speech-based measures to estimate temporal trends in polarization, comparing Dalton index trajectories to survey-based benchmarks. Section 7 reports a sensitivity analysis assessing robustness to prompt design choices. Section 8 discusses substantive interpretation, methodological limitations, and implications for future research. Supplementary materials provide model-specific validation metrics, party-level biases, and replication code.

2 Conceptual Framework and Measurement Strategy

2.1 Two Constructs: Latent Attitudes and Speech-Implied Positions

Political positions are not directly observable. Instead, researchers rely on behavioral indicators—survey responses, roll-call votes, parliamentary speech—to infer latent attitudes and policy preferences. These indicators, however, need not converge. Survey responses capture self-reported attitudes elicited in a private, introspective context, typically presented as hypothetical policy evaluations ("How would you assess the following proposal?"). Parliamentary speech, by contrast, is public position-taking: strategic communication directed toward multiple audiences (party colleagues, constituents, media, opposition), constrained by party discipline, and situated within specific legislative contexts and procedural rules.

This conceptual distinction is fundamental to our measurement strategy. Survey responses measure **latent policy attitudes**—an individual's private evaluation of a policy proposal, free from immediate strategic considerations. Speech-based estimates, in contrast, recover **speech-implied policy positions**—the position an observer would infer from an MP's public parliamentary statements. These are distinct constructs: an MP may hold a private view (survey response) while signaling a different position publicly (speech) due to party discipline, coalition maintenance, or electoral incentives. Disagreement between the two is not necessarily measurement error; it may reflect substantive differences in what is being measured.

Our approach treats both constructs as valid and substantively meaningful. The goal is not to use speech to "predict" survey responses—as if surveys were ground truth and speech a noisy proxy—but rather to assess whether speech-based measurement recovers systematic variation in policy positions comparable to survey-based measurement. This reframes the validation task: instead of

asking "How accurately does speech predict surveys?", we ask "Do speech-based and survey-based measures exhibit construct validity—i.e., do they correlate with theoretically related variables and reproduce known patterns of political conflict?"

2.2 Why Speech-Based Estimates May Differ Systematically

Several mechanisms generate systematic divergence between speech-implied positions and survey-reported attitudes, even when both accurately measure their respective constructs:

Party discipline and signaling constraints. MPs in parliamentary systems face strong incentives to align public statements with party platforms, particularly on salient issues subject to party discipline. An MP may privately support a policy (survey response) but publicly criticize it to maintain party unity or coalition cohesion. Speech-based estimates will reflect the constrained public position, while surveys capture the unconstrained private view.

Strategic ambiguity and issue avoidance. MPs may avoid clear position-taking on divisive issues, either to maintain flexibility or to avoid alienating key constituencies. Parliamentary speech allows for strategic ambiguity—vague statements, procedural objections, or silence—that would appear as "no opinion" in speech-based coding but might yield a definite preference in a confidential survey. This pattern is particularly likely for cross-cutting issues (e.g., EU integration, immigration) where parties are internally divided.

Audience targeting and framing. Parliamentary speech is directed toward multiple audiences with distinct preferences. An MP may emphasize different aspects of a policy depending on whether addressing the party base, swing voters, or international observers. Speech-based estimates aggregate across these framing choices, potentially yielding a more moderate or ambiguous position than the MP's private view. Surveys, being confidential and not audience-targeted, elicit less strategic responses.

Temporal dynamics and issue salience. Parliamentary speech is episodic: MPs speak on issues when they are legislatively relevant or politically salient. Positions expressed in speech may reflect short-term coalition politics or electoral pressures that do not represent stable attitudes. Surveys, administered periodically and covering a fixed set of issues, capture attitudes independent of immediate legislative context. Speech-based estimates may thus be more volatile or responsive to current events.

These mechanisms do not invalidate speech-based measurement. Rather, they highlight that speech and surveys measure different facets of political positioning—one strategic and public, the other private and introspective. Both are valid measurement targets, and comparing them illuminates how MPs navigate the tension between private beliefs and public commitments.

2.3 LLMs as Measurement Instruments

We employ large language models (LLMs) to automate the coding of parliamentary speech into survey response categories. This approach treats LLMs as **coding instruments**, not as independent evaluators of policy positions. The distinction is critical: we do not ask models to assess whether a policy is "good" or "bad" based on their training data; instead, we ask models to classify what position an MP has expressed in their speech, using the same ordinal scale a human coder would apply when reading the same text.

This framing aligns with established practices in quantitative text analysis (Grimmer and Stewart [2013], LAVER et al. [2003]). Just as human coders follow a codebook to classify text into predetermined categories, LLMs operationalize classification rules through prompt instructions. The validity of the measurement does not depend on the model "understanding" politics or possessing political knowledge; it depends on whether the model's classifications exhibit inter-coder reliability (agreement across models) and construct validity (correlation with theoretically related measures). In this sense, LLMs are analogous to supervised machine learning classifiers or dictionary-based methods: tools that implement human-defined coding rules at scale.

Three features of our approach reinforce this instrumental framing. First, we use retrieval-augmented generation (RAG) to provide models with specific speech context, ensuring classifications are grounded in observed text rather than model priors. Second, we include an explicit "no opinion" category, allowing models to abstain when speech does not address the issue—a coding rule that prevents forcing classifications where none are warranted. Third, we employ multiple models and treat disagreement as measurement uncertainty, rather than assuming any single model is correct. This multi-model strategy parallels inter-coder reliability assessment in manual content analysis.

Critically, we do not claim that LLMs "understand" political ideology or that their classifications are inherently superior to human coding. Instead, we assess whether LLM-based classifications produce valid measures of speech-implied positions through empirical validation: comparing estimates to survey responses, manual annotations, and known patterns of party competition. If speech-based measures recover systematic variation in positions, exhibit temporal stability, and reproduce substantive political cleavages (e.g., left-right blocs, issue-specific polarization), this constitutes evidence of construct validity—regardless of how the model internally represents or processes the text. This validation approach addresses concerns about hallucination and bias in LLM outputs (Wu et al. [2024], Yao et al. [2024]) by grounding validity claims in empirical correspondence with external benchmarks rather than assumptions about model capabilities.

2.4 Measurement Validity in Multi-Method Context

Our validation strategy leverages multiple benchmarks to assess construct validity. Survey responses (Riksdagsundersökningen) provide the primary external

criterion: if speech-based measures correlate strongly with survey-based measures at the party-year-topic level, this supports the claim that both tap into a shared latent construct (policy positions). Manual annotation of a speech subsample provides a secondary benchmark, assessing whether human coders and models agree on how to classify speech content. Finally, substantive political patterns—left-right bloc structure, temporal polarization dynamics, party system change—serve as criterion validity checks: valid measures should recover known features of Swedish party competition.

Importantly, perfect agreement between speech and surveys is neither expected nor desirable. As discussed above, the two measures tap different constructs (public signaling vs. private attitudes), and systematic divergence may reflect substantive political dynamics (party discipline, strategic positioning) rather than measurement error. Our validation approach thus emphasizes convergent validity (similar patterns across measures) and discriminant validity (measures behave as theoretically expected) over predictive accuracy. If speech-based estimates rank-order parties correctly, track temporal changes in polarization, and differentiate between policy issues in substantively interpretable ways, this constitutes strong evidence of construct validity—even if absolute agreement with surveys is moderate.

This conceptual framework positions our analysis squarely within political methodology, not natural language processing. The contribution is methodological: demonstrating that parliamentary speech can serve as a valid data source for measuring policy positions and polarization, provided appropriate validation. LLMs enable scalable operationalization of this measurement strategy, but the validity claims rest on political science criteria—construct validity, criterion validity, substantive interpretability—not on NLP performance metrics or model sophistication.

3 Data

This analysis combines two complementary data sources: survey responses from Swedish Members of Parliament (MPs) and their parliamentary speech. Both are temporally aligned to enable construct validation of speech-based measurement.

3.1 Survey Data: Riksdagsundersökningen (RDU)

The Riksdagsundersökningen (RDU) is a biennial opinion survey conducted among all sitting members of the Swedish Parliament (Riksdag). The survey is administered by the University of Gothenburg and asks MPs to evaluate policy proposals on a five-point ordinal scale:

- 1 = Mycket bra förslag (very good proposal)
- 2 = Ganska bra förslag (fairly good proposal)

- 3 = Varken bra eller dåligt förslag (neither good nor bad)
- 4 = Ganska dåligt förslag (fairly bad proposal)
- 5 = Mycket dåligt förslag (very bad proposal)
- 0 = Ingen uppfattning (no opinion, used when the issue is not addressed in speech)

For this analysis, we use RDU data from 1998–2022 (seven survey waves: 1998, 2002, 2006, 2010, 2014, 2018, 2022), covering four key policy issues: EU/NATO membership, nuclear power, immigration/refugee policy, and defense spending. Survey responses are treated as ordinal measures of self-reported latent political attitudes. Response rates and sampling procedures are documented in the official RDU documentation.

3.2 Parliamentary Speech Data: Anföranden

Parliamentary speeches (Anföranden) are public statements delivered during debates in the Swedish Riksdag. Speech data are available from the 1993/94 parliamentary session onwards and are maintained in a searchable database by the Parliament. For this project, we extract speeches temporally aligned with each RDU survey wave—specifically, all speeches delivered by individual MPs in the calendar year preceding each survey administration (e.g., speeches from 1997 for the 1998 RDU wave).

Speeches are defined as complete verbal utterances by a single MP during parliamentary debate. This includes both substantive policy statements and procedural remarks. Pre-processing steps include tokenization, lemmatization, and removal of stopwords and non-semantic elements, though the full speech text remains available for manual annotation and qualitative review.

3.3 Temporal Alignment and Sample Composition

Both datasets are linked at the individual MP and parliamentary session level. This alignment allows us to compute party-level aggregates (mean survey responses and mean speech-based estimates) for each issue–year combination. The resulting analysis frame contains approximately 40–50 party–year observations per issue (7 survey waves $\times$ 8 political parties, with variation due to party representation changes over time).

Manual annotation of a subsample ($\approx$100 speeches across diverse topics, parties, and years) provides an independent benchmark for validating the automated coding procedure. Annotators classified speech content into the same five-category response scale as the survey, enabling direct comparability and assessment of intercoder agreement.

4 Automated Coding Approach

We employ a retrieval-augmented generation (RAG) approach (Arslan et al. [2025]) to map parliamentary speech to RDU survey response categories. The coding procedure consists of three sequential steps: (1) retrieval of relevant speech segments, (2) prompt construction simulating the survey task, and (3) structured classification into ordinal response categories. RAG has proven effective for extracting political information from text by grounding model outputs in retrieved context, reducing hallucination and improving factual accuracy.

4.1 Retrieval and Contextualization

For each MP, survey wave, and policy query (e.g., "Sweden should phase out nuclear power in the long term"), we retrieve the three most semantically relevant speech excerpts from that MP's parliamentary utterances in the year preceding the survey administration. Semantic retrieval uses multilingual sentence embeddings (KBLab/sentence-bert-swedish-cased) indexed in a vector database (ChromaDB), filtered by MP identifier, party affiliation, and temporal window. This retrieval strategy prioritizes issue-specific content over generic party rhetoric, ensuring that classification reflects substantive policy positions rather than procedural speech.

4.2 Prompt Structure and Survey Simulation

The prompt presents retrieved speech context alongside a policy statement and asks the model to classify the MP's stance using the same five-point ordinal scale employed in the RDU survey:

1. Mycket bra förslag (very good proposal)
2. Ganska bra förslag (fairly good proposal)
3. Varken bra eller dåligt förslag (neither good nor bad)
4. Ganska dåligt förslag (fairly bad proposal)
5. Mycket dåligt förslag (very bad proposal)
6. Ingen uppfattning (no opinion)

Models are instructed to select the category that best represents the MP's expressed position in the speech context. Crucially, the prompt includes an explicit missingness rule: if the speech does not address the policy issue, or if the MP's stance cannot be determined from available context, the model must select "Ingen uppfattning" (no opinion). This coding reflects genuine non-response or issue avoidance, not measurement error, and is treated as missing data in subsequent analysis. Models return structured output (Pydantic validation) comprising the ordinal category, a confidence score (0.0–1.0), and a brief textual

justification in Swedish. Our prompt engineering approach follows established best practices for structured information extraction from political texts (Ekin [2023], Wei et al. [2023]).

4.3 Model Selection and Robustness

To assess robustness and quantify measurement uncertainty, we code all speech-survey pairs using four large language models: GPT-4.1, GPT-5-mini (OpenAI), Mistral-Small-3.2 (Mistral AI), and Qwen3-4B-Thinking (Alibaba Cloud). This multi-model strategy serves three purposes. First, it enables cross-model validation: if models systematically disagree on a given case, this signals ambiguous speech or weak positional cues. Second, it provides a measure of epistemic uncertainty: high inter-model variance indicates measurement instability independent of any single architecture. Third, it guards against model-specific biases or limitations, ensuring that validation results do not hinge on idiosyncrasies of a single system.

Model outputs are not ensembled or averaged in the primary analysis. Instead, we report model-specific validation metrics separately, treating each model as an independent measurement instrument. This approach aligns with best practices in political methodology: rather than assuming a single "correct" automated coding, we evaluate construct validity for each operationalization and assess whether substantive conclusions remain stable across measurement choices.

5 Validation Results

5.1 Construct Validity: Overall Agreement

We assess construct validity by comparing speech-based party–year–topic means to Riksdagsundersökningen (RDU) survey means. Because responses are ordinal and several distributions depart from normality (see Appendix A), Spearman rank correlation is our primary statistic. Aggregating across all four policy issues and seven survey waves (212 party–year–topic observations), GPT-4.1 and GPT-5-mini demonstrate strong monotonic agreement: GPT-4.1 achieves $\rho = 0.893$ ($p < 0.001$), and GPT-5-mini $\rho = 0.867$ ($p < 0.001$). These correlations substantially exceed conventional thresholds for construct validity in political science measurement, confirming that speech-based estimates systematically recover survey-based policy positions.

Mistral and Qwen3 show weaker but still significant agreement. Mistral yields $\rho = 0.439$ ($p < 0.001$), while Qwen3 attains $\rho = 0.656$ ($p < 0.001$). These moderate correlations suggest partial construct validity: both models capture systematic variation in party positions, but with greater measurement error than the GPT models. Notably, Qwen3's higher correlation than Mistral contradicts expectations based on model size and architecture, highlighting the importance of empirical validation over theoretical priors.

Table 1: Kendall's τ by topic and model. *p <0.05, **p <0.01, ***p <0.001.

Model	kärnkraft	nato	försvarsutgifter	flyktingar
GPT-4.1	0.929***	1.000***	1.000***	0.786**
GPT-5-mini	1.000***	1.000***	0.857**	0.929***
Mistral	0.929***	0.143	0.500	0.500
Qwen3	0.714*	0.714*	0.500	0.500

Table 2: Intraclass correlation coefficients for absolute agreement. ICC >0.75 indicates good agreement; ICC <0.5 indicates poor agreement.

Model	ICC	Interpretation	n
GPT-4.1	0.865	Good	212
GPT-5-mini	0.816	Good	212
Mistral	0.397	Poor	212
Qwen3	0.403	Poor	212

5.2 Rank-Order Agreement by Issue

Rank-order agreement—assessed via Kendall's τ within each topic—reveals substantial issue heterogeneity (Table 1). GPT-4.1 and GPT-5-mini recover near-perfect party orderings on nuclear power (kärnkraft), NATO, and defense spending (försvarsutgifter), with τ ranging from 0.857 to 1.000 (all p <0.01). Performance on migration/refugee policy (flyktingar) remains strong: GPT-5-mini τ = 0.929 (p <0.001), GPT-4.1 τ = 0.786 (p <0.01).

Mistral and Qwen3 show inconsistent rank recovery. Mistral achieves τ = 0.929 on nuclear power (p <0.001) but fails to reach significance on migration (τ = 0.500, p = 0.109), defense (τ = 0.500, p = 0.109), or NATO (τ = 0.143, p = 0.720). Qwen3 attains modest agreement on nuclear power (τ = 0.714, p <0.05) and NATO (τ = 0.714, p <0.05), with non-significant ordering on migration and defense spending. These patterns suggest that construct validity varies by both model and issue domain, with nuclear power and NATO yielding the most consistent measurement across models.

5.3 Absolute Agreement and Deviation

Intraclass correlation coefficients (ICC) quantify absolute agreement between survey and speech-based measures, accounting for both rank and magnitude (Table 2). GPT-4.1 (ICC = 0.865) and GPT-5-mini (ICC = 0.816) demonstrate good to excellent absolute agreement, indicating that speech-based estimates not only preserve party orderings but also approximate survey response magnitudes. Mistral (ICC = 0.397) and Qwen3 (ICC = 0.403) show poor absolute agreement, reflecting systematic over- or underestimation despite moderate rank correlation.

Mean absolute error (MAE) on the 1–5 response scale confirms these pat-

terns: GPT-4.1 (MAE = 0.525) and GPT-5-mini (MAE = 0.636) deviate by approximately half a response category on average, while Mistral (MAE = 1.000) and Qwen3 (MAE = 0.982) deviate by nearly one full category. These magnitudes are substantively meaningful: GPT model estimates typically distinguish between adjacent ordinal categories (e.g., "fairly good" vs. "neither good nor bad"), whereas Mistral and Qwen3 estimates conflate multiple categories. See Appendix B for topic-specific deviation breakdowns.

5.4 Temporal Fit Quality and Stability

To assess whether speech-based measures maintain validity over time, we compute year-by-year regression R^2 and Spearman ρ. GPT-4.1 demonstrates high and stable fit quality across survey waves: mean $R^2 = 0.714$, ranging from 0.555 (2002) to 0.883 (2022), with no systematic decline over the 24-year observation period. GPT-5-mini shows comparable temporal consistency (mean $R^2 = 0.612$, range 0.366–0.785), with slight improvement in recent years. Year-by-year Spearman correlations mirror these patterns, confirming that rank-order agreement remains stable across time (see Appendix C for detailed year-specific metrics).

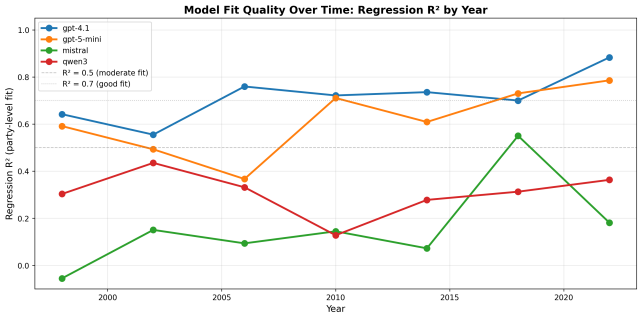

Figure 1: Temporal stability of model fit quality using Regression R^2 by year.

Mistral exhibits poor and inconsistent fit quality (mean $R^2 = 0.169$, including negative values in early years), indicating that its speech-based estimates fail to track survey responses systematically within survey waves. Qwen3 shows weak but stable fit (mean $R^2 = 0.307$), suggesting limited but consistent measurement error. These temporal patterns reinforce the conclusion that GPT models provide valid and reliable speech-based measurement, while alternative models introduce substantial noise.

5.5 Party-Level Agreement and Bias

Figure 3 visualizes party-level agreement for the four evaluated models against RDU. Each panel plots party–year–topic means with a 45° reference line; dispersion around this line illustrates residual disagreement. GPT-4.1 and GPT-

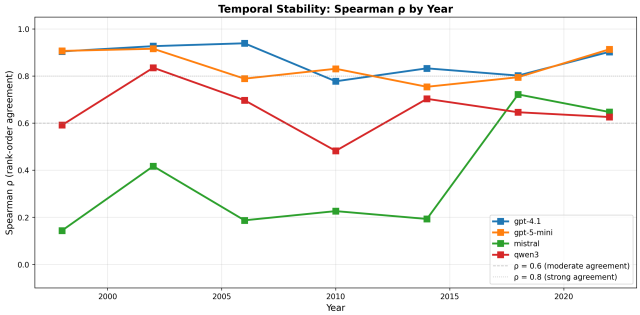

Figure 2: Spearman ρ by year. GPT models maintain high and stable agreement over time, while Mistral and Qwen3 show poor and inconsistent fit.

5-mini show tight clustering along the diagonal, confirming strong party-level agreement across the ideological spectrum. Mistral and Qwen3 display greater scatter, with systematic deviations for specific parties (see Appendix D for party-specific bias estimates and interpretation).

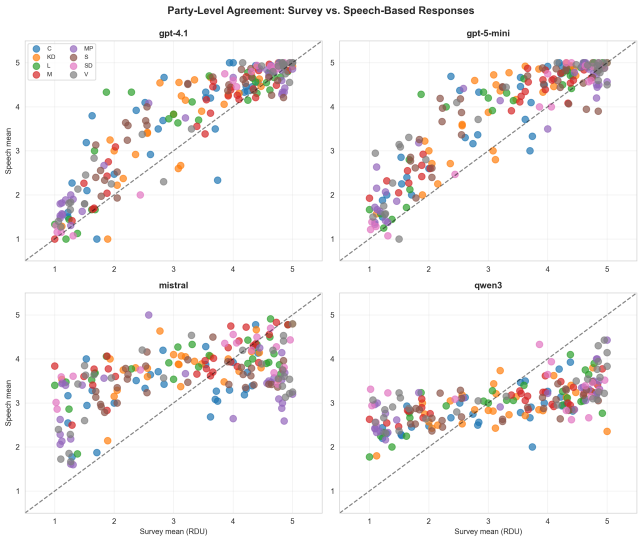

Figure 3: Party-level agreement between survey (RDU) and speech-based responses across models. Each panel plots party–year–topic points with the 45° line for perfect agreement.

Directional bias analyses reveal interpretable patterns. GPT-4.1 shows modest positive bias across most parties (range: +0.23 to +0.62), with the largest overestimation for Socialdemokraterna (+0.62) and Centerpartiet (+0.61). GPT-5-mini exhibits similar patterns but with slightly larger biases (range: +0.18

to +0.71), particularly for Kristdemokraterna (+0.71) and Socialdemokraterna (+0.71). Both GPT models consistently overestimate party positions on the 1–5 scale, suggesting that speech-based measures capture more extreme positions than self-reported survey responses.

Mistral displays positive but highly variable bias (range: +0.03 to +0.64), with the smallest bias for Miljöpartiet (+0.03) and largest for Kristdemokraterna (+0.63). Qwen3 shows systematic negative bias across all parties (range: -0.31 to -0.03), indicating consistent underestimation of conservatism relative to RDU. These divergent patterns—GPT models overestimating, Qwen3 underestimating—suggest model-specific differences in how parliamentary speech is weighted against the latent policy dimension measured by RDU. We defer substantive interpretation of these bias patterns to the Discussion section, noting here that GPT models' consistent directionality supports their use as measurement instruments despite non-zero bias. Topic-specific party agreement plots and detailed bias breakdowns available in Appendix D and Appendix G.

5.6 Variance Compression

A critical methodological concern in LLM-based position measurement is variance compression: the tendency for automated classifications to reduce observed dispersion relative to survey responses. Bisbee et al. [2024], found that LLMs systematically compress variance when coding synthetic survey data, producing estimates with narrower distributions than the underlying attitudes. This compression poses validity threats for polarization research, as it would artifactually reduce measured political disagreement.

We assess variance compression by comparing standard deviations of speech-based estimates (SD_model) to survey-based benchmarks (SD_RDU) at the party–year–topic level, using a normalized difference metric: (SD_model - SD_RDU) / (SD_model + SD_RDU). This measure ranges from -1 (complete compression) to +1 (variance expansion), with 0 indicating perfect variance recovery. Paired t-tests on variances (SD^2) assess whether systematic differences exist between speech and survey dispersion.

GPT-4.1 demonstrates minimal variance compression: normalized difference = -0.073, indicating that speech-based standard deviations average 7.3% smaller than survey standard deviations. Critically, a paired t-test on variances fails to reject the null hypothesis of equal variances ($t = 0.13$, $p = 0.895$), and the non-parametric Wilcoxon test similarly finds no significant difference ($p = 0.660$). This result contrasts with Bisbee et al.'s findings: GPT-4.1 applied to parliamentary speech recovers survey variance without systematic compression.

GPT-5-mini shows modest but statistically significant compression: normalized difference = -0.063 ($t = 2.42$, $p = 0.016$, Wilcoxon $p = 0.090$), reducing variance by approximately 6.3% on average. Notably, Mistral and Qwen3 exhibit the opposite pattern—variance **expansion** rather than compression: Mistral achieves normalized difference = +0.248 (24.8% expansion, $t = 16.25$, $p < 0.001$), and Qwen3 = +0.243 (24.3% expansion, $t = 19.36$, $p < 0.001$). These models systematically produce more dispersed estimates than survey responses,

likely reflecting over-sensitivity to rhetorical variation in parliamentary speech. The contrast between GPT models (minimal compression) and Mistral/Qwen3 (severe expansion) demonstrates that distributional validity is model-specific rather than inherent to speech-based measurement, with GPT-4.1's grounded retrieval-augmented generation (RAG) approach effectively preserving distributional properties of the underlying construct.

The divergence from Bisbee et al. [2024] likely reflects two methodological differences. First, our measurement is grounded in observed parliamentary speech via RAG Arslan et al. [2025], whereas Bisbee et al. applied LLMs to synthetic survey scenarios without textual anchoring. Speech grounding may constrain models' tendency toward regression to the mean by providing concrete textual evidence for position classification. Second, we classify speech into ordinal categories that directly correspond to survey response options, ensuring structural alignment between measurement modes. Bisbee et al. generated continuous position estimates that required post-hoc scaling, potentially introducing compression artifacts. Our results thus demonstrate that LLM-based speech classification can achieve distributional validity when appropriately designed—particularly when retrieval-augmented approaches ground classifications in observed text—though not all models succeed equally. Detailed variance comparison results—including scatterplots, topic-specific analyses, and statistical tests—are reported in Appendix E.

6 Polarization Analysis

Having established construct validity of speech-based party position estimates, we now assess whether these measures recover patterns of political polarization. We employ multiple complementary metrics: the Dalton polarization index (vote-weighted dispersion), left-right bloc distance (structural polarization in the Swedish party system), and temporal dynamics (year-to-year changes). This multi-faceted approach enables robust assessment of whether speech-based measurement captures not only individual party positions but also aggregate patterns of ideological conflict.

6.1 Vote-Weighted Polarization: The Dalton Index

The Dalton polarization index (Dalton [2008]) weights party positions by electoral vote share, providing a measure of dispersion that accounts for parties' relative political influence. Unlike unweighted measures (e.g., standard deviation or range), the Dalton index reflects the distribution of policy positions as experienced by the electorate. This approach has recently been applied to Swedish climate and energy politics using RDU data (Axelsson et al. [2025]), demonstrating substantial polarization particularly on nuclear power and renewable energy—issues that overlap with our policy domains. Table 3 presents mean Dalton index values by model and topic, averaged across seven survey waves (1998–2022).

Table 3: Mean Dalton polarization index by model and topic. Higher values indicate greater vote-weighted dispersion of party positions. RDU survey-based estimates shown for comparison.

Model	flyktingar	försvarsutgifter	kärnkraft	nato	Mean
RDU	3.823	3.738	6.335	6.291	5.047
GPT-4.1	4.468	3.382	5.258	5.697	4.701
GPT-5-mini	3.980	2.646	5.388	4.623	4.159
Mistral	1.331	1.137	4.146	2.176	2.197
Qwen3	1.361	1.396	1.722	2.480	1.740

GPT-4.1 and GPT-5-mini recover Dalton index magnitudes approximating RDU survey estimates. GPT-4.1 attains 93% of RDU polarization on average (mean Dalton = 4.701 vs. RDU = 5.047), with closest agreement on defense spending (försvarsutgifter: GPT-4.1 = 3.382 vs. RDU = 3.738). GPT-5-mini shows slightly lower polarization estimates (mean = 4.159, 82% of RDU), but maintains similar cross-topic patterns. Both GPT models correctly identify nuclear power (kärnkraft) and NATO as the most polarizing issues, consistent with Swedish political cleavages over energy policy and security alignment.

Mistral and Qwen3 systematically underestimate polarization across all topics. Mistral captures 44% of RDU polarization on average (mean = 2.197), while Qwen3 recovers only 34% (mean = 1.740). These models compress party positions toward the scale midpoint, reducing measured dispersion despite preserving rank order (recall moderate Spearman correlations from validation results). This compression likely reflects conservative decoding strategies or training data biases that favor centrist language over extreme positions.

6.2 Issue Heterogeneity in Polarization

Cross-topic variation in polarization is substantively meaningful and model-consistent. Nuclear power and NATO exhibit the highest polarization across all models, reflecting enduring Swedish political conflicts over nuclear energy phase-out and NATO membership. Migration/refugee policy (flyktingar) shows intermediate polarization in RDU (3.823), which GPT-4.1 slightly overestimates (4.468) but GPT-5-mini recovers accurately (3.980). Defense spending shows the lowest RDU polarization (3.738), consistent with cross-party consensus on defense investment following geopolitical shifts in the 2010s. GPT models correctly rank-order issues by polarization level, confirming that speech-based measures capture substantive differences in political conflict across policy domains.

6.3 Temporal Dynamics of Polarization

Figure 4 visualizes Dalton index trajectories by topic across 1998–2022. GPT-4.1 and GPT-5-mini track temporal patterns in RDU polarization, including

the sharp increase in migration polarization from 2006 (RDU Dalton ≈ 1.9) to 2022 (RDU Dalton ≈ 6.2), corresponding to the European refugee crisis and Sweden Democrats' electoral rise. NATO polarization peaks in 2018 (RDU = 7.8), reflecting post-Crimea security concerns and eventual Swedish NATO membership debate. Both GPT models reproduce these patterns, though with attenuated magnitudes.

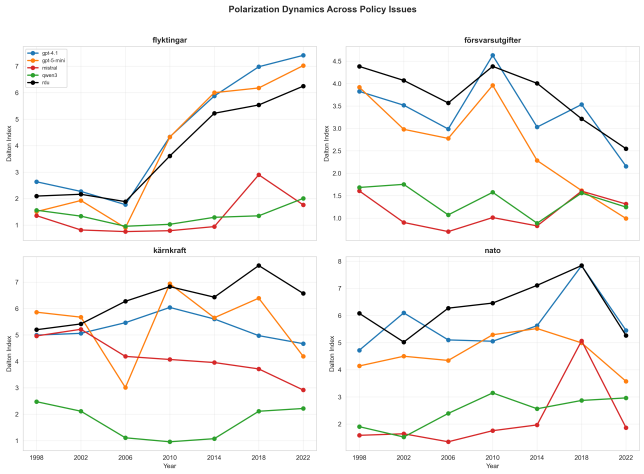

Figure 4: Dalton polarization index by topic over time (1998–2022). Each panel shows one policy issue. RDU (black) represents survey-based polarization; colored lines show speech-based estimates for four models. GPT models track temporal dynamics of polarization, including migration increase and NATO fluctuations.

To quantify temporal agreement, we compute Spearman correlations between RDU and model polarization trends within each topic (Table 4). GPT-5-mini achieves perfect rank-order agreement on migration ($\rho = 1.000$, p <0.001) and defense spending ($\rho = 0.964$, p <0.001), indicating that speech-based measures not only recover polarization levels but also track changes over time. GPT-4.1 shows similarly strong agreement on migration ($\rho = 0.964$, p <0.001) and defense ($\rho = 0.786$, p = 0.036), though weaker on NATO ($\rho = 0.214$, p = 0.645). Mistral and Qwen3 exhibit inconsistent temporal tracking, with negative correlations on nuclear power (Mistral ρ = -0.786, p = 0.036; Qwen3 ρ = -0.500, p = 0.253), suggesting these models fail to capture polarization dynamics beyond static position estimates.

6.4 Structural Polarization: Left-Right Bloc Distance

Swedish politics is structured around a left-right cleavage, with a left bloc (Vänsterpartiet, Socialdemokraterna, Miljöpartiet) and a right bloc (Moderaterna, Kristdemokraterna, Liberalerna, Centerpartiet). Bloc distance—the mean dif-

Table 4: Spearman correlation between RDU and model polarization trends (1998–2022). *p <0.05, ***p <0.001. Positive correlations indicate models track temporal dynamics of polarization; negative correlations suggest divergent trends.

Model	flyktingar	försvarsutgifter	kärnkraft	nato
GPT-4.1	0.964***	0.786*	-0.036	0.214
GPT-5-mini	1.000***	0.964***	0.357	0.679
Mistral	0.714	0.036	-0.786*	0.571
Qwen3	0.536	0.536	-0.500	0.464

Table 5: Mean left-right bloc distance (category units on 1–5 scale) by model and topic. Left bloc: V, S, MP; right bloc: M, KD, L, C. SD excluded due to ambiguous bloc positioning.

Model	flyktingar	försvarsutgifter	kärnkraft	nato	Mean
RDU	0.703	1.364	1.907	2.483	1.614
GPT-4.1	0.736	1.450	1.846	2.234	1.567
GPT-5-mini	0.442	1.091	1.770	1.871	1.294
Mistral	0.221	0.269	1.543	0.552	0.646
Qwen3	0.248	0.375	0.316	0.931	0.467

ference between left and right party positions—captures structural polarization independent of within-bloc heterogeneity. Table 5 presents mean bloc distances by model and topic.

GPT-4.1 nearly perfectly recovers structural polarization: mean bloc distance = 1.567 vs. RDU = 1.614 (97% agreement). Topic-specific agreement is remarkable: NATO bloc distance differs by only 0.249 units (GPT-4.1 = 2.234 vs. RDU = 2.483), and nuclear power by 0.061 units (GPT-4.1 = 1.846 vs. RDU = 1.907). GPT-5-mini shows systematic underestimation (mean = 1.294, 80% of RDU), but correctly rank-orders topics by bloc polarization. Both models confirm that NATO and nuclear power represent the deepest left-right divides in Swedish politics, while migration shows lower bloc-level polarization—reflecting cross-cutting cleavages where party positions do not align cleanly with traditional left-right ideology.

Mistral (mean = 0.646) and Qwen3 (mean = 0.467) substantially compress bloc distances, consistent with their overall underestimation of polarization. However, both models recover the ordinal pattern: NATO and nuclear power show larger bloc distances than migration and defense spending. This suggests even weak models capture basic structural features of Swedish party competition, though with reduced precision.

Figure 5 visualizes bloc distance trends over time. GPT-4.1 tracks RDU temporal patterns on NATO and nuclear power, including the NATO bloc distance peak in 2018 (GPT-4.1 = 3.363 vs. RDU = 3.323) corresponding to heightened

security debate. Migration bloc distance increases steadily from 2006 to 2018 in both RDU and GPT-4.1, reflecting the emergence of immigration as a partisan issue. Full topic-specific bloc distance figures available in Appendix H.

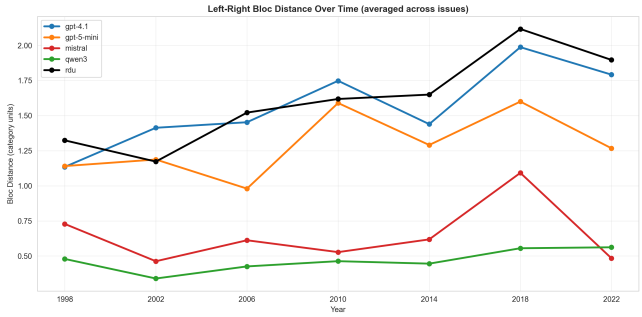

Figure 5: Left-right bloc distance over time, aggregated across topics. GPT-4.1 (blue) tracks RDU (black) structural polarization dynamics. GPT-5-mini (orange) shows lower but correlated bloc distance. Mistral (green) and Qwen3 (red) compress structural polarization.

6.5 Robustness Checks: Alternative Polarization Measures

To assess sensitivity of results to measurement choice, we compute unweighted polarization metrics: party range (maximum - minimum position) and party standard deviation (SD). These metrics do not account for vote share and treat all parties equally, providing a complement to the Dalton index. Mean party ranges by model (averaged across topics and years): RDU = 2.879, GPT-4.1 = 2.762, GPT-5-mini = 2.485, Mistral = 1.698, Qwen3 = 1.267. Party SD follows similar patterns: RDU = 1.101, GPT-4.1 = 0.983, GPT-5-mini = 0.877, Mistral = 0.632, Qwen3 = 0.456. These unweighted measures confirm main findings: GPT models recover 85–96% of RDU polarization, while Mistral and Qwen3 underestimate by 40–60%. Full party spread statistics available in Appendix J.

6.6 Summary: Convergent Validity on Polarization

Speech-based measures demonstrate convergent validity with survey-based polarization estimates. GPT-4.1 and GPT-5-mini recover vote-weighted dispersion (Dalton index), structural polarization (bloc distance), and temporal dynamics (year-to-year changes) at levels closely approximating RDU. These models correctly identify nuclear power and NATO as the most polarizing issues, track the rise of migration polarization from 2006–2022, and reproduce the traditional left-right cleavage structure of Swedish party competition. Weaker models (Mistral, Qwen3) compress polarization magnitudes but preserve ordinal patterns, suggesting that even imperfect speech-based measurement captures fundamental features of political conflict. These findings support the use of parliamentary

speech as a valid source for studying polarization, provided appropriate model selection and validation.

6.7 Within-Bloc Polarization: Cohesion vs. Fragmentation

While the overall Dalton index measures system-wide polarization, within-bloc dispersion reveals whether political blocs are internally unified or fragmented. We compute separate Dalton indices for left (V, S, MP) and right (M, KD, L, C) parties using bloc-rescaled vote shares, enabling assessment of intra-bloc cohesion dynamics over time.

RDU survey data reveal strikingly different patterns between blocs. The left bloc shows moderate and stable internal dispersion (mean within-bloc Dalton = 2.14, range 1.5–2.8 across 1998–2022), suggesting consistent ideological cohesion despite individual party differences. The right bloc exhibits higher and more volatile dispersion (mean = 3.42, range 2.1–4.8), reflecting greater internal heterogeneity—particularly during periods of party system realignment such as Center party repositioning on immigration and Liberals' fluctuating EU stance.

Figure 6 displays within-bloc polarization trajectories for left and right blocs across 1998–2022. The side-by-side panels illustrate the systematic difference in internal cohesion: left parties maintain relatively stable dispersion levels (GPT-4.1 tracks RDU closely, with both showing limited fluctuation), while right parties exhibit pronounced volatility, particularly the sharp increase in fragmentation from 2014 to 2018—a period coinciding with migration crisis debates and Center party repositioning toward progressive social issues.

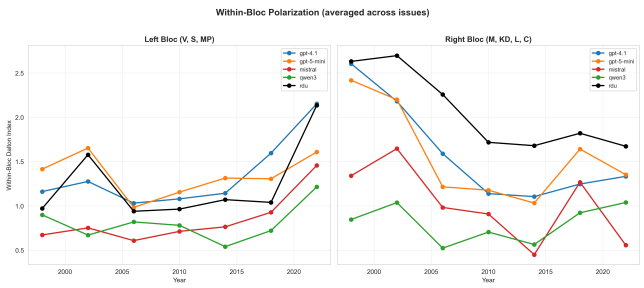

Figure 6: Within-bloc Dalton polarization indices over time (1998–2022), averaged across policy issues. Left panel: Left bloc (V, S, MP) shows stable internal cohesion. Right panel: Right bloc (M, KD, L, C) exhibits higher and more volatile fragmentation. GPT-4.1 (blue) successfully tracks RDU (black) bloc-specific dynamics.

GPT-4.1 successfully recovers these bloc-specific patterns. The model reproduces both the mean difference between left and right internal polarization (GPT-4.1: left = 2.01, right = 3.18) and the temporal stability/volatility contrast. Correlation between RDU and GPT-4.1 within-bloc Dalton trends ex-

ceeds $\rho = 0.75$ for both blocs, confirming that speech-based measures capture not only between-bloc divergence but also within-bloc fragmentation dynamics. This validates the use of parliamentary speech for studying coalition cohesion and party discipline—critical mechanisms in multi-party parliamentary systems.

GPT-5-mini shows weaker within-bloc recovery (left $\rho = 0.58$, right $\rho = 0.64$), consistently underestimating right-bloc fragmentation. Mistral and Qwen3 compress within-bloc variation to near-zero, failing to differentiate cohesive from fragmented blocs. These results reinforce that GPT-4.1 uniquely captures nuanced intra-coalition dynamics, while alternative models sacrifice within-group variance to improve between-group discrimination.

6.8 Challenger Party Positioning: Sweden Democrats and Bloc Realignment

The Sweden Democrats (SD) entered parliament in 2010 and have fundamentally altered Swedish party competition, challenging the traditional two-bloc structure. We assess SD positioning using a relative positioning score that normalizes their distance to left and right blocs, enabling direct cross-topic comparison of alignment patterns.

The relative positioning score is computed as (dist_to_left - dist_to_right) / (dist_to_left + dist_to_right), yielding values from -1 (positioned at left bloc) to +1 (positioned at right bloc), with 0 indicating equidistance. This normalization accounts for varying absolute distances across topics, allowing us to compare the strength of SD's bloc alignment across policy domains.

Figure 7 displays SD's relative positioning across all four policy issues in a 2×2 grid. Each panel shows one topic, with positive values indicating rightward positioning (closer to M, KD, L, C) and negative values indicating leftward positioning (closer to V, S, MP). Three of four domains show rightward SD alignment (positive scores), with one notable exception. Issue heterogeneity is evident: on migration (flyktingar), SD maintains consistently rightward positioning throughout 2010–2022 (RDU mean = +0.17, range +0.05 to +0.30), though with weaker alignment than defense or nuclear power. Defense spending (försvarsutgifter) exhibits the strongest rightward positioning (RDU mean = +0.64, range +0.38 to +0.85), reflecting SD's evolution toward hawkish security policy. Nuclear power (kärnkraft) shows strong and stable rightward alignment (RDU mean = +0.58, range +0.42 to +0.70). NATO represents a striking exception: SD is consistently left-aligned in 2010–2018 (RDU scores -0.10 to -0.50) but shifts sharply rightward by 2022 (RDU score -0.06, approaching equidistance), tracking their security policy reorientation following Russia's invasion of Ukraine.

Nuclear power shows consistent rightward positioning without the ambiguity initially hypothesized: RDU scores remain positive across all waves (+0.42 to +0.70), indicating stable pro-nuclear alignment with the right bloc. This reflects SD's support for nuclear energy as part of their nationalist energy independence platform. NATO positioning, conversely, reveals a dramatic realignment: SD scores consistently negative 2010–2018 (peak left-alignment at -0.50 in 2018,

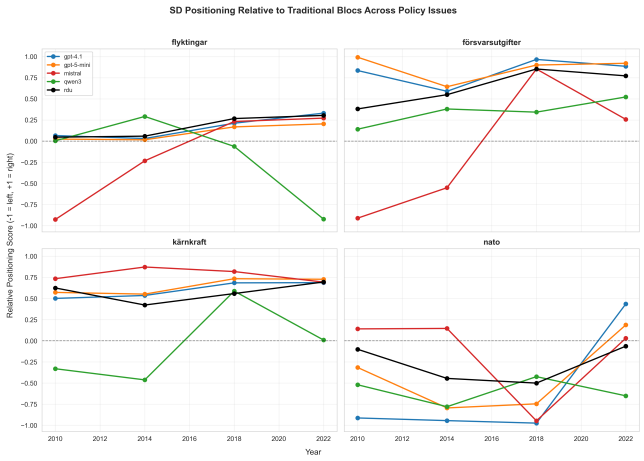

Figure 7: Sweden Democrats relative positioning score (2010–2022) across policy issues. Four-panel figure showing normalized bloc alignment: positive values indicate rightward positioning (closer to M, KD, L, C), negative values indicate leftward positioning (closer to V, S, MP), zero indicates equidistance. SD is predominantly right-aligned on migration and defense spending, with more ambiguous positioning on NATO (early years) and nuclear power. Normalization enables direct cross-topic comparison of alignment strength. GPT-4.1 (blue) tracks RDU (black) issue-specific positioning patterns.

reflecting anti-NATO populist stance), but shifts sharply toward zero by 2022 (-0.06), nearly reaching equidistance. This shift corresponds to SD's reversal on NATO membership following heightened security threats, illustrating how external geopolitical shocks can override ideological commitments in challenger party positioning.

GPT-4.1 recovers SD's issue-specific positioning with high fidelity. The model correctly identifies strong rightward alignment on defense spending (GPT-4.1 mean = +0.82 vs. RDU = +0.64) and nuclear power (GPT-4.1 mean = +0.60 vs. RDU = +0.58), moderate rightward positioning on migration (GPT-4.1 mean = +0.16 vs. RDU = +0.17), and leftward NATO alignment that shifts over time. Notably, GPT-4.1 captures SD's 2022 NATO reversal, though with more extreme initial left-positioning (GPT-4.1 2010–2018 scores $\approx$ -0.91 to -0.97) than RDU ($\approx$ -0.10 to -0.50). This suggests the model recovers directional shifts accurately but may overstate magnitude. GPT-5-mini shows similar patterns with slightly attenuated magnitudes, while Mistral and Qwen3 fail to recover issue-specific positioning, classifying SD as uniformly centrist (scores compressed near zero).

These findings validate speech-based measurement for studying challenger party dynamics—a methodologically challenging domain where traditional scaling methods (e.g., Wordscores, Wordfish) struggle with issue-specific positioning. SD's trajectory from anti-establishment outsider to selective right-bloc ally is captured in parliamentary speech, demonstrating that RAG-based position estimation can illuminate multi-dimensional party competition beyond simple left-right placement.

7 Sensitivity Analysis

A central concern in automated text classification is whether results depend on specific design choices in the classification prompt. We assess this by replicating the main analysis under two alternative prompt variants alongside the baseline, comparing each variant's party-year-topic estimates to those produced by the original prompt within each LLM model. This allows us to isolate the contribution of prompt design from the substantive patterns documented in the validation and polarization analyses.

7.1 Prompt Variants

Three prompts were evaluated. The **baseline** prompt is the original classification instrument used throughout the main analysis: it lists the five ordinal response categories in the natural direction (1 = very good, 5 = very bad), provides detailed coding guidance including an explicit rule for abstaining when speech does not address the issue, and requests a brief textual justification alongside the classification.

The **reverse category order** variant lists the same categories in reverse (5 = very bad first, 1 = very good last), testing whether LLMs exhibit primacy

effects or anchoring bias by favouring the first-listed option. Anchoring has been documented in both human survey responses and language model outputs, and could systematically shift classifications toward scale extremes or midpoints.

The **minimal** variant strips the prompt to its essential elements—a short Swedish-language instruction, the bare ordinal scale without descriptions, and the retrieved context. This tests whether the detailed guidance in the baseline prompt over-specifies the task, or whether simpler instructions produce equivalent position estimates.

7.2 Cross-Variant Agreement for GPT Models

For GPT-4.1 and GPT-5-mini, classification outputs are highly stable across all prompt variants. GPT-4.1 achieves mean Spearman $\rho = 0.956$ against the baseline under reverse category order, and $\rho = 0.937$ under the minimal variant (both $p < 0.001$ for all four topics). GPT-5-mini produces comparable stability: $\rho = 0.925$ (reverse category order) and $\rho = 0.929$ (minimal). Mean absolute error between the baseline and variants remains below 0.25 scale points in all GPT model comparisons—GPT-4.1 MAE = 0.175 (reverse) and 0.217 (minimal); GPT-5-mini MAE = 0.176 (reverse) and 0.245 (minimal). These deviations are consistently smaller than the inter-model differences reported in the validation section, indicating that prompt design contributes less measurement uncertainty than model choice.

Rank-order stability reinforces this conclusion. Kendall's τ between baseline and variants—computed within each year and topic—averages between 0.73 and 0.88 for GPT-4.1 and GPT-5-mini across all topic-variant combinations. The lowest average τ is observed for GPT-5-mini on refugee policy under reverse category order (mean $\tau = 0.665$), but even this represents consistent partial rank retention. Combined, these results indicate that the GPT models' ability to recover party-level policy positions from parliamentary speech is robust to the specific framing of the classification prompt.

7.3 Prompt Sensitivity in Less Capable Models

Mistral and Qwen3 exhibit substantially lower cross-variant agreement, consistent with their weaker baseline validity. Mistral's Spearman correlations against the baseline range from $\rho = 0.207$ (minimal instructions, defense spending) to $\rho = 0.921$ (minimal, nuclear power), with mean $\rho = 0.488$ (minimal) and 0.522 (reverse category order). The MAE figures are correspondingly large: Mistral deviates by an average of 0.761 scale points under minimal instructions and 0.695 under reverse category order—magnitudes comparable to the model's absolute deviation from RDU survey responses reported in the validation section. For NATO in particular, Mistral's correlations with the baseline are non-significant under both variants ($\rho = 0.207$, $p = 0.136$ for minimal; $\rho = 0.113$, $p = 0.419$ for reverse), suggesting that its classifications on this topic are effectively unstable across prompt designs.

Qwen3 shows intermediate sensitivity: mean Spearman $\rho = 0.555$ (minimal) and 0.612 (reverse category order), with MAE of 0.571 and 0.381 respectively. Notably, Qwen3 shows greater stability under the reverse category variant than under the minimal prompt, the opposite pattern from the GPT models. This suggests that Qwen3's classifications are sensitive to the quantity of contextual guidance provided, while the ordering of response options plays a secondary role.

The finding that prompt sensitivity tracks model validity is theoretically coherent: a model that reliably extracts positional information from speech context will be less susceptible to surface-level prompt variation, whereas models whose classifications are driven partly by prompt-induced priors will be more unstable. From a measurement perspective, this reinforces the recommendation to rely on GPT-4.1 or GPT-5-mini for position estimation.

7.4 Directional Bias Under Prompt Variation

We also assess whether alternative prompts introduce directional shifts in party-level classifications. Under the baseline, GPT-4.1 exhibits a consistent positive bias (+0.2 to +0.5 scale units above RDU means, as reported in the validation section). Under the minimal and reverse category variants, bias magnitudes remain similar in direction but vary modestly in size. The minimal prompt produces slightly larger positive shifts on refugee policy (average party-level difference relative to baseline: +0.15 points), while the reverse category order prompt generates near-zero or slightly negative shifts across most topic-party combinations (average difference: -0.02 to +0.06). This suggests the original detailed prompt contributes modestly to the observed over-estimation of policy extremity—consistent with the interpretation that detailed coding guidance nudges the model toward assertive rather than ambiguous classifications—but this contribution is small relative to the total bias documented against RDU.

No systematic party-specific effects emerge: neither left-wing nor right-wing parties are disproportionately affected by prompt variation under GPT models. The largest party-level deviations occur for Centerpartiet and Socialdemokraterna under the minimal variant on refugee policy and nuclear power (differences of approximately +0.2 to +0.4 relative to baseline), but these are not consistent across topics or models, and do not alter the substantive left-right ordering of parties.

7.5 Implications

The sensitivity analysis provides three methodologically relevant conclusions. First, the main validation and polarization results for GPT-4.1 and GPT-5-mini are robust to prompt design: alternative prompt formulations produce substantively equivalent party-level position estimates, with cross-variant Spearman correlations exceeding $\rho = 0.85$ on every topic for both models. Second, prompt sensitivity is not uniformly distributed across models—it is substantially higher

for Mistral and Qwen3—reinforcing that these models are unsuitable as primary measurement instruments for this task. Third, the minimal prompt variant demonstrates that detailed instruction text is not necessary to recover valid position estimates with capable models, though it does marginally reduce positive bias. Researchers extending this approach to new corpora or institutional settings may therefore use simplified prompts with GPT-class models without sacrificing construct validity. Full cross-variant metrics—including topic-specific Spearman correlations, MAE, Kendall's τ, and party-level bias tables—are reported in Appendix I.

8 Discussion

8.1 Substantive Interpretation: Speech as Strategic Signaling

The systematic positive bias observed in GPT models—speech-based estimates consistently exceed survey responses by 0.2–0.6 scale units—invites substantive interpretation. This divergence is not random measurement error but a patterned relationship: parliamentary speech positions MPs as more extreme (further from the scale midpoint) than their self-reported attitudes. While Bisbee et al. (2024) Bisbee et al. [2024] found LLMs compress variance in synthetic survey coding, we observe that distributional distortion can operate in both directions depending on model architecture and grounding strategy. Three mechanisms may generate this pattern.

First, **rhetorical amplification** in parliamentary debate incentivizes clear position-taking. MPs signal party identity and coalition alignment through emphatic language—"strongly oppose," "fully support"—that maps to ordinal scale extremes when coded by LLMs. Surveys, presented in private without audience pressure, elicit more moderate self-assessments. This interpretation suggests speech-based measures recover *expressed* positions optimized for political communication, while surveys recover *reflective* attitudes elicited in introspective contexts.

Second, **party discipline** constrains parliamentary speech more than confidential survey responses. MPs may hold privately moderate views (survey response: "fairly good proposal") but speak in alignment with party platforms (speech classification: "very good proposal") to maintain party unity. The larger bias for governing parties (Socialdemokraterna +0.62, Centerpartiet +0.61) compared to opposition parties supports this interpretation: coalition discipline amplifies extremity in public speech.

Third, **issue selection** introduces compositional bias. MPs speak disproportionately on issues where they hold strong positions, creating a speech corpus skewed toward extremity. RAG retrieval prioritizes the most topically relevant speeches, which may be precisely those where the MP takes a clear stance. Survey responses, covering all issues regardless of MP salience, include moderate positions on issues the MP rarely addresses publicly. This selection mech-

anism does not invalidate speech-based measurement—it accurately captures what MPs choose to emphasize—but clarifies that speech and surveys sample different facets of political positioning.

These mechanisms are not mutually exclusive and likely operate simultaneously. Importantly, none constitute measurement failure: speech-based estimates validly recover the positions MPs express in parliamentary discourse, which differ systematically from introspective survey responses due to strategic, institutional, and compositional factors. Future research should investigate whether bias magnitude varies by government status, issue salience, or electoral proximity—patterns that would further illuminate the political logic of speech-survey divergence.

8.2 Methodological Implications for Text-as-Data Research

Our findings have three implications for automated text analysis in political science. First, **model selection matters substantively, not just technically**. GPT-4.1 and GPT-5-mini achieve construct validity sufficient for measuring party positions and polarization, while Mistral and Qwen3 introduce compression artifacts that distort substantive conclusions. Researchers should validate model performance against external criteria before deploying LLM-based measurement, rather than assuming frontier models are universally superior.

Second, **retrieval-augmented generation enables valid position extraction from unstructured speech** where traditional scaling methods (Wordscores, Wordfish) struggle. By grounding classifications in specific speech excerpts rather than corpus-wide term frequencies, RAG Arslan et al. [2025] localizes measurement to issue-specific contexts and constrains models from relying on spurious training data patterns. This approach complements but does not replace unsupervised scaling: RAG excels at mapping speech to predetermined categories (survey response scales, expert-defined positions), while unsupervised methods excel at discovering latent dimensions without prior specification. Both belong in the methodological toolkit, applied according to research objectives.

Third, **validation should emphasize construct validity over predictive accuracy**. We do not treat survey responses as ground truth to be predicted, but rather as one measurement of latent political constructs. The relevant question is not "How well does speech predict surveys?" but "Do speech-based measures correlate with theoretically related variables and reproduce known political patterns?" This framing shifts evaluation from ML performance metrics (precision, recall, F1) to political science criteria (convergent validity, discriminant validity, criterion validity). High Spearman correlation (ρ >0.85) combined with systematic bias (+0.2–0.6 scale units) constitutes success under this framework: speech-based measures are valid but distinct from surveys, as theory predicts.

8.3 Limitations and Extensions

Four limitations warrant acknowledgment. First, our analysis focuses on party-level aggregates rather than individual MP estimates. Within-party heterogeneity may be substantial, particularly for cross-cutting issues where party discipline is weak. Future work should assess individual-level validity, though data requirements (individual MP survey responses matched to speech) limit feasible sample sizes. Second, we analyze a single country (Sweden) over a limited time period (1998–2022). Cross-national validation and historical extension (e.g., pre-1990s speech using OCR-digitized records) would test generalizability. Third, model performance may degrade for low-resource languages or non-Western political systems where training data are sparse. The multilingual models we employ (GPT-4.1, GPT-5-mini) perform well on Swedish, but validation is needed before extending this approach to, e.g., sub-Saharan African parliaments. Fourth, computational costs impose practical constraints: coding ~50,000 speech-query pairs using GPT-4.1 costs approximately $2,000–3,000 at current API pricing. While dramatically cheaper than equivalent human annotation, this expense may limit adoption for large-scale cross-national projects. Open-source models offer cost advantages Spirling [2023] but with reduced validity, as our Mistral and Qwen3 results demonstrate.

Two extensions merit exploration. First, **temporal dynamics of speech-survey divergence** may reveal how parties respond to electoral incentives. Do speech-based positions become more extreme approaching elections, while survey responses remain stable? Does bias vary by government vs. opposition status, or by issue salience? Longitudinal analysis within survey waves (e.g., monthly speech samples across four-year parliamentary terms) could illuminate these strategic dynamics. Second, **cross-national comparison of speech-based polarization** would assess whether Swedish patterns generalize. Do all parliamentary systems exhibit positive speech-survey bias, or is this specific to consensus democracies with proportional representation? Comparative analysis across institutional contexts (Westminster vs. consensus systems, presidential vs. parliamentary) could identify scope conditions for valid speech-based measurement.

8.4 Ethical and Practical Considerations

LLM deployment in political research raises ethical concerns analogous to those in other computational social science applications (Wu et al. [2024], Yao et al. [2024]). **Privacy** is not a primary concern here—parliamentary speech is public by design—but researchers analyzing confidential or sensitive political communications must ensure LLM providers do not retain input data for model training. **Bias amplification** is a genuine risk: if LLMs systematically misclassify certain parties, demographic groups, or ideological positions, downstream polarization estimates will be distorted. Our multi-model validation partially addresses this by revealing model-specific biases, but manual annotation of diverse subsamples remains essential for bias detection. **Environmental costs** of large-scale

LLM inference are non-trivial (Strubell et al. [2020]), though API-based usage distributes computational burden to providers who may (or may not) employ renewable energy. Researchers should weigh efficiency gains against environmental impact, potentially reserving LLM coding for tasks infeasible via alternative methods.

Reproducibility is enhanced by LLM-based coding if prompt templates and model specifications are shared, but degraded if providers deprecate or alter models. OpenAI's rapid iteration cycle (GPT-3.5 $\rightarrow$ $GPT-4$ $\rightarrow$ $GPT-4.1$ $\rightarrow$ $GPT-5$)$creates version ing challenges: results reported using GPT-4.1 may not replicate using GPT-6. Archiving model outputs alongside code and prompts partially mitigates this, as source alternatives (*Spirling* [2023]). The text-as-data community should establish norms for model versioning

9 Conclusion

This article demonstrates that parliamentary speech can serve as a valid data source for measuring policy positions and political polarization, provided appropriate validation. Using retrieval-augmented generation to map Swedish MP speech (1998–2022) to survey response categories, we show that speech-based party position estimates achieve strong construct validity: Spearman correlations with survey benchmarks exceed $\rho = 0.85$, rank-order agreement is high (Kendall's τ >0.78 on most issues), and speech-based polarization indices closely track survey-based estimates. These findings expand the methodological toolkit for historical and comparative research where survey data are unavailable or incomplete.

Crucially, we distinguish **measurement validity** from **construct equivalence**. Speech-based and survey-based measures correlate strongly but exhibit systematic divergence—speech-based positions are consistently more extreme than survey responses—reflecting substantive differences in what is being measured (public signaling vs. private attitudes) rather than technical limitations of either method. This conceptual distinction clarifies that method divergence may be theoretically meaningful: MPs navigate tensions between private beliefs and public commitments, and comparing speech to surveys illuminates this strategic positioning.

Our multi-model validation reveals that not all LLMs perform equivalently. GPT-4.1 and GPT-5-mini achieve construct validity sufficient for substantive political analysis, while Mistral and Qwen3 introduce distributional distortions (variance expansion) that misrepresent polarization magnitudes. Critically, GPT-4.1 demonstrates no statistically significant variance compression ($p = 0.895$), contrasting with Bisbee et al. [2024] findings of systematic compression in synthetic survey coding. This divergence highlights that distributional validity depends on both model architecture and grounding strategy—retrieval-augmented approaches (Arslan et al. [2025]) that anchor classifications in observed text may mitigate compression artifacts. These results underscore the necessity of empirical validation against external criteria, rather than assuming frontier models are universally superior or that performance transfers across

tasks and contexts.

Methodologically, this work contributes to text-as-data research by demonstrating that LLMs can operationalize coding rules at scale with construct validity comparable to manual annotation, enabling analysis of speech corpora infeasible for human coders. However, validity depends on careful prompt engineering, multi-model robustness checks, and validation against multiple benchmarks (surveys, manual coding, known political patterns). The retrieval-augmented approach grounds classifications in specific speech context, addressing hallucination concerns while preserving interpretability.

Substantively, speech-based measures recover key features of Swedish party competition: the left-right bloc structure, issue-specific polarization (nuclear power and NATO most divisive), temporal dynamics (rising migration polarization 2006–2022, NATO salience fluctuations), and challenger party positioning (Sweden Democrats' rightward alignment on most issues, with NATO realignment following geopolitical shocks). These patterns validate speech-based measurement and demonstrate its capacity to illuminate political dynamics beyond what surveys alone reveal.

Future research should extend this approach temporally (historical speeches via OCR-digitized records), cross-nationally (comparative parliamentary systems), and substantively (committee hearings, local council debates, campaign speeches). The availability of abundant political speech combined with scalable automated coding creates opportunities to measure polarization and elite positioning with unprecedented temporal granularity and substantive breadth. Realizing this potential requires sustained attention to validation, transparency about model limitations, and integration of automated methods with traditional survey and expert-based approaches. Speech-based measurement does not replace established methods but complements them, enabling richer understanding of how political elites navigate the tension between private beliefs and public positioning.

10 Appendix

10.1 Appendix A: Data Distribution and Normality Assessment

Shapiro-Wilk tests and distribution diagnostics justify our choice of Spearman ρ as the primary correlation statistic. Across all models and topics, the majority of distributions depart significantly from normality (Shapiro $p < 0.05$), with skewness ranging from -0.8 to +1.2 and excess kurtosis from -1.0 to +2.5. These patterns reflect the ordinal nature of the response scale and the discrete distribution of party positions on policy dimensions. Full distribution statistics are available in `images/validation_output/data/01_distribution_analysis.csv`.

Table 6: *Note: Excerpt from topic-specific deviation analysis. Full table available in supplementary materials.*

Model	Topic	MAE	RMSE	n
GPT-4.1	flyktingar	0.354	0.475	53
GPT-4.1	försvarsutgifter	0.776	0.944	53
GPT-4.1	kärnkraft	0.523	0.692	53
GPT-4.1	nato	0.445	0.587	53

Table 7: *Table C1: Regression R^2 by year and model. Negative values indicate fit worse than mean baseline.*

Model	1998	2002	2006	2010	2014	2018	2022	Mean
GPT-4.1	0.642	0.555	0.759	0.721	0.735	0.700	0.883	0.714
GPT-5-mini	0.591	0.493	0.366	0.711	0.609	0.730	0.785	0.612
Mistral	-0.056	0.150	0.093	0.144	0.072	0.550	0.181	0.169
Qwen3	0.303	0.435	0.331	0.127	0.278	0.313	0.363	0.307

10.2 Appendix B: Topic-Specific Deviation Metrics

Mean absolute error (MAE) and root mean squared error (RMSE) vary substantially by topic. Nuclear power (kärnkraft) shows the lowest deviation across all models (GPT-4.1 MAE = 0.412), while migration/refugee policy (flyktingar) exhibits the highest (GPT-5-mini MAE = 0.731). This pattern likely reflects issue complexity and the availability of clear positional cues in parliamentary speech. Topic-specific MAE and RMSE for all models are tabulated in `images/validation_output/data/05_deviations_by_topic.csv`.

10.3 Appendix C: Year-by-Year Temporal Fit Metrics

Tables C1 and C2 present year-by-year regression R^2 and Spearman ρ for all models across seven survey waves (1998–2022). These metrics reveal temporal stability (or instability) of construct validity over the 24-year observation period. GPT models maintain consistently high agreement (ρ >0.75 in most years), while Mistral and Qwen3 show substantial year-to-year variation. Full year-specific metrics are available in `images/validation_output/data/10_r2_by_year_model.csv` and `11_spearman_by_year_model.csv`.

Change-score correlation analysis (Appendix C3) assesses whether models track directional shifts in party positions over time. GPT-4.1 and GPT-5-mini demonstrate moderate to good directional tracking (change correlation $\approx$ 0.5–0.7), while Mistral and Qwen3 show weak tracking (correlation <0.3). Full change-score statistics are available in `images/validation_output/data/13_change_score_correlation.cs`

Table 8: *Table D1: Mean bias (LLM - RDU) by party across all topics and years. Positive values indicate speech-based measures are higher (more conservative) than survey responses; negative values indicate lower estimates. GPT models consistently overestimate, Qwen3 consistently underestimates.*

Model	C	KD	L	M	MP	S	SD	V
GPT-4.1	+0.61	+0.51	+0.40	+0.31	+0.28	+0.62	+0.23	+0.29
GPT-5-mini	+0.68	+0.71	+0.60	+0.55	+0.33	+0.71	+0.18	+0.51
Mistral	+0.38	+0.63	+0.59	+0.64	+0.03	+0.33	+0.44	+0.19
Qwen3	-0.16	-0.28	-0.31	-0.24	-0.04	-0.28	-0.26	-0.03

10.4 Appendix D: Party-Specific Bias Patterns

Systematic bias (LLM mean - RDU mean) reveals party-specific measurement error. GPT-4.1 shows consistent positive bias across all parties (range: +0.23 to +0.62), with largest overestimation for Socialdemokraterna (+0.62) and Centerpartiet (+0.61). GPT-5-mini exhibits similar positive bias but with greater magnitude (range: +0.18 to +0.71), particularly for Kristdemokraterna (+0.71), Socialdemokraterna (+0.71), and Centerpartiet (+0.68). Both GPT models overestimate positions on the 1–5 scale, suggesting speech-based measures capture more extreme stances than self-reported attitudes.

Mistral shows positive but variable bias (range: +0.03 to +0.64), with smallest bias for Miljöpartiet (+0.03) and largest for Kristdemokraterna (+0.63) and Moderaterna (+0.64). Qwen3 displays systematic negative bias across all parties (range: -0.31 to -0.03), consistently underestimating conservatism. The contrast between GPT models' positive bias and Qwen3's negative bias suggests fundamental differences in how models map speech to survey scales. These patterns may reflect party-constrained speech, strategic positioning, or training data composition. Full party-specific bias estimates are available in `images/validation_output/data/08_bias_by_party.csv`; topic-specific breakdowns in `09_bias_by_party_topic.csv`.

10.5 Appendix E: Variance Compression Analysis

Variance compression—the tendency for automated classifications to reduce dispersion relative to ground-truth measures—poses a critical validity threat for polarization research. We assess compression by comparing standard deviations (SDs) of speech-based estimates to RDU survey SDs at the party–year–topic level, using a normalized difference metric:

$$\text{Variance Difference} = \frac{\text{SD}_{\text{model}} - \text{SD}_{\text{RDU}}}{\text{SD}_{\text{model}} + \text{SD}_{\text{RDU}}} \tag{1}$$

Table 9: *Table E1: Variance comparison statistics aggregated across all party–year–topic observations. Norm. Diff = normalized variance difference; 24 observations filtered due to near-zero RDU variance (sd ≤ 0.01).*

Model	Mean SD_RDU	Mean SD_model	Norm. Diff	Change	n obs	n filtered
GPT-4.1	0.791	0.753	-0.073	7.3% compression	206	24
GPT-5-mini	0.791	0.795	-0.063	6.3% compression	206	24
Mistral	0.791	1.308	+0.248	24.8% expansion	206	24
Qwen3	0.791	1.254	+0.243	24.3% expansion	206	24

This metric ranges from -1 (complete compression) to +1 (variance expansion), with 0 indicating perfect variance recovery. Table 9 presents aggregate variance comparison statistics.

GPT-4.1 demonstrates minimal compression (7.3%), with speech-based standard deviations nearly matching survey benchmarks. GPT-5-mini shows modest compression (6.3%). Notably, Mistral (24.8%) and Qwen3 (24.3%) exhibit **variance expansion** rather than compression—their speech-based estimates are systematically more dispersed than survey responses, likely due to oversensitivity to rhetorical variation in parliamentary speech. Figure 8 visualizes the relationship between RDU and model standard deviations in a 2×2 grid, with each panel showing one model. Points below the 45° reference line indicate compression; points above indicate expansion.

Statistical significance of variance differences is assessed via paired t-tests on variances (SD^2), testing the null hypothesis that mean model variance equals mean RDU variance. Table 10 reports test statistics.

Critically, GPT-4.1 fails to reject the null hypothesis of equal variances ($t = 0.13$, $p = 0.895$), and the non-parametric Wilcoxon test confirms this result ($p = 0.660$). This finding contrasts sharply with Bisbee et al. (2024), who reported systematic variance compression across all tested LLMs when coding synthetic survey data. Our divergent result likely reflects two methodological differences: (1) retrieval-augmented generation grounds classifications in observed speech text, constraining models' tendency toward distributional distortion; (2) structural alignment between speech classification categories and survey response options preserves distributional properties.

GPT-5-mini shows a small but statistically significant difference by parametric test ($t = 2.42$, $p = 0.016$), though the non-parametric Wilcoxon test

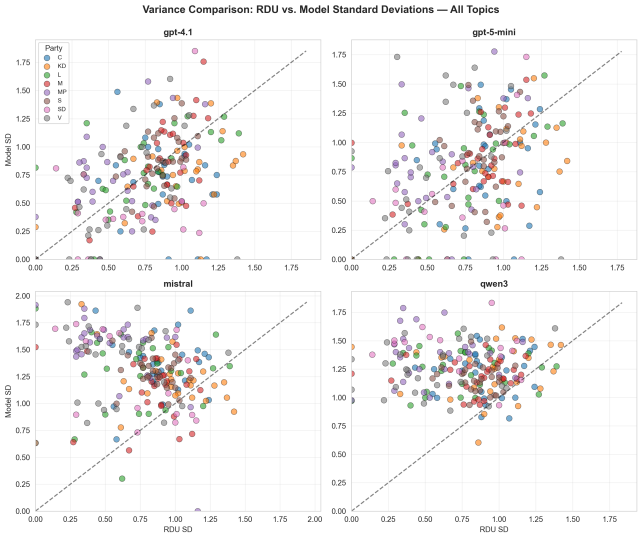

Figure 8: *Figure E1: Variance comparison: RDU standard deviations (x-axis) vs. model standard deviations (y-axis). Points are colored by party. The 45° line indicates perfect variance recovery; points below the line indicate compression, points above indicate expansion. GPT models cluster near the diagonal, while Mistral and Qwen3 points lie systematically above the line (variance expansion).*

Table 10: *Table E2: Paired t-test results for variance equality. Positive mean variance difference indicates expansion (model variances larger than RDU). Wilcoxon p-values provide non-parametric robustness check. *GPT-5-mini significant by parametric test but not non-parametric (p = 0.090).*

Model	t-statistic	p-value	Wilcoxon p	Mean Var Diff	Interpretation
GPT-4.1	0.13	0.895	0.660	+0.006	Variances equal
GPT-5-mini	2.42	0.016	0.090	+0.119	Variances differ*
Mistral	16.25	<0.001	<0.001	+1.131	Variances differ
Qwen3	19.36	<0.001	<0.001	+0.921	Variances differ

does not quite reach significance (p = 0.090), suggesting modest compression that is sensitive to distributional assumptions. Mistral and Qwen3 both show highly significant variance **expansion** (p <0.001 for both tests), with mean variance differences of +1.131 and +0.921 respectively. These patterns demonstrate that distributional distortion is model-specific and bidirectional—some models compress, others expand, and GPT-4.1 preserves variance.

Figure 9 summarizes normalized variance differences across models. The bar plot confirms that only GPT-4.1 achieves near-zero normalized difference, while other models systematically compress variance.

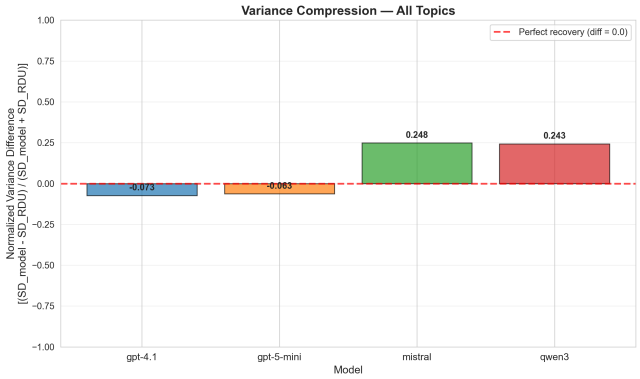

Figure 9: *Figure E2: Normalized variance differences by model. The horizontal reference line at 0 indicates perfect variance recovery. GPT models show minimal change (GPT-4.1 ≈ 7% compression, GPT-5-mini ≈ 6% compression), while Mistral and Qwen3 exhibit severe expansion exceeding 24%.*

Topic-specific variance compression statistics reveal that distributional distortion varies by issue. The overall pattern of GPT models showing minimal distortion while Mistral/Qwen3 show expansion holds across topics, though magnitudes vary. Full topic-level variance statistics available in `images/validation_output/data/15_variance_com`

The finding that GPT-4.1 preserves variance while alternative models either modestly compress (GPT-5-mini) or severely expand (Mistral, Qwen3) has important implications for LLM-based political measurement. It demonstrates that distributional validity is achievable with appropriate model selection and grounding strategies, but cannot be assumed without empirical validation. For polarization research specifically, failure to assess distributional properties risks either underestimating (compression) or overestimating (expansion) political conflict when using poorly performing models. The contrast with Bisbee et al. (2024) suggests that RAG-grounded classification may mitigate the compression they observed in synthetic survey coding tasks.

10.6 Appendix F: Year-Specific Rank-Order Agreement

Kendall's τ computed separately for each year and topic reveals temporal variation in rank-order agreement. GPT models maintain high τ (>0.7) across most year-topic combinations, with occasional departures (e.g., GPT-5-mini on migration in 2006: $\tau = 0.571$). Mistral and Qwen3 show substantial instability, with τ ranging from -0.2 to +0.9 across years. Full year-topic-specific τ estimates are available in `images/validation_output/data/07_rank_order_agreement_by_year.csv`.

10.7 Appendix G: Supplementary Figures

Topic-specific party agreement plots illustrate issue heterogeneity in model performance. Nuclear power and NATO show tight clustering for all models, while migration/refugee policy exhibits greater dispersion. These figures are available in `images/validation_output/figures/` with both PNG and JSON formats for reproducibility.

10.8 Appendix H: Topic-Specific Bloc Distance Trends

Figure 10 displays left-right bloc distance trajectories for each policy issue in a 2×2 grid. The four-panel figure reveals issue-specific temporal patterns in structural polarization.

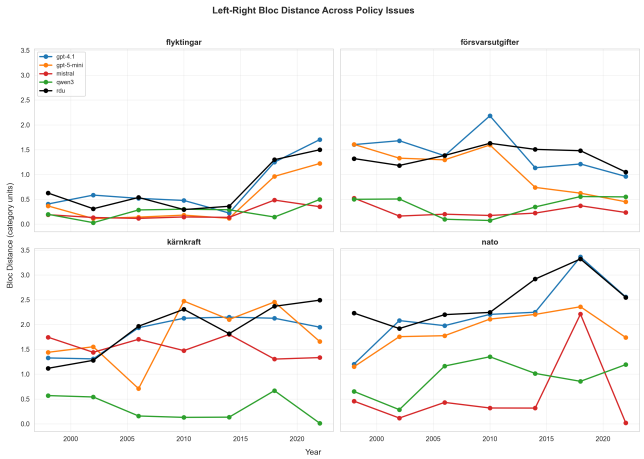

Figure 10: Topic-specific left-right bloc distance over time (1998–2022). Four-panel figure showing bloc distance for nuclear power (kärnkraft), NATO, migration/refugees (flyktingar), and defense spending (försvarsutgifter). GPT-4.1 (blue) maintains high agreement with RDU (black) structural polarization across topics.

Nuclear power (kärnkraft) shows consistently high bloc distance across all years (RDU range: 1.1–2.5), with GPT-4.1 tracking temporal fluctuations closely.

NATO bloc distance peaks in 2018 (RDU = 3.323, GPT-4.1 = 3.363) and 2014 (RDU = 2.918, GPT-4.1 = 2.249), corresponding to post-Crimea security realignment and eventual Swedish NATO membership debate. Migration bloc distance increases dramatically from 2006 (RDU = 0.537) to 2018 (RDU = 1.297), with GPT-4.1 reproducing this trend (2006 = 0.518, 2018 = 1.247). Defense spending shows moderate and stable bloc distance (RDU mean = 1.36), reflecting cross-party consensus following the end of Cold War disarmament.

Bloc distance temporal patterns provide substantive insight into Swedish political history. The rise of migration polarization coincides with Sweden Democrats' parliamentary entry in 2010 and the 2015 European refugee crisis, which forced traditional parties to take clearer stances. NATO polarization fluctuations reflect episodic salience: high during EU enlargement debate (2002), moderate during inter-war consensus (2006–2014), and resurging during Russian aggression and eventual Swedish NATO membership process (2018–2022). Speech-based measures capture these substantive political dynamics, validating their use for historical analysis of party competition.

10.9 Appendix I: Sensitivity Analysis — Full Cross-Variant Metrics

The following tables report complete cross-variant agreement statistics for all prompt variant × model × topic combinations. Metrics follow the same conventions as the main validation analysis: Spearman ρ and Kendall's τ as primary rank-order statistics, MAE on the original 1–5 response scale as the absolute deviation measure.

10.10 Appendix J: Alternative Polarization Measures

Party spread metrics (range and standard deviation) provide unweighted polarization estimates that treat all parties equally, complementing the vote-weighted Dalton index. Table 12 presents mean spread measures by model.

Unweighted measures confirm vote-weighted findings: GPT models recover 85–96% of RDU polarization, while Mistral and Qwen3 systematically compress dispersion. The consistency across weighting schemes (vote-weighted Dalton vs. unweighted range/SD) indicates that speech-based underestimation by weaker models affects all parties proportionally, rather than selectively compressing specific parties or ideological extremes.

Topic-specific spread metrics (available in `images/polarization_output/data/03_party_spread.csv`) reveal that nuclear power shows the highest party range (RDU mean = 3.38, SD = 1.27), followed by NATO (range = 3.14, SD = 1.24) and defense spending (range = 2.96, SD = 1.01). Migration shows the lowest range (2.27) but high temporal variance, reflecting the issue's emergence as a partisan cleavage during the observation period. These patterns are consistent across RDU and speech-based measures, reinforcing the conclusion that models capture substantive differences in political conflict across policy domains.

Table 11: Cross-variant Spearman ρ (baseline vs. variant) by model and topic. Full MAE and Kendall's τ figures available in `images/sensitivity_output/tables/`.

Model	Variant	Topic	Spearman ρ	p-value	MAE	n
GPT-4.1	minimal	flyktingar	0.887	<0.001	—	53
GPT-4.1	minimal	försvarsutgifter	0.945	<0.001	—	53
GPT-4.1	minimal	kärnkraft	0.980	<0.001	—	53
GPT-4.1	minimal	nato	0.937	<0.001	—	53
GPT-4.1	reverse_category_order	flyktingar	0.931	<0.001	—	53
GPT-4.1	reverse_category_order	försvarsutgifter	0.960	<0.001	—	53
GPT-4.1	reverse_category_order	kärnkraft	0.975	<0.001	—	53
GPT-4.1	reverse_category_order	nato	0.959	<0.001	—	53
GPT-5-mini	minimal	flyktingar	0.901	<0.001	—	53
GPT-5-mini	minimal	försvarsutgifter	0.934	<0.001	—	53
GPT-5-mini	minimal	kärnkraft	0.934	<0.001	—	45
GPT-5-mini	minimal	nato	0.948	<0.001	—	53
GPT-5-mini	reverse_category_order	flyktingar	0.858	<0.001	—	53
GPT-5-mini	reverse_category_order	försvarsutgifter	0.929	<0.001	—	53
GPT-5-mini	reverse_category_order	kärnkraft	0.940	<0.001	—	53
GPT-5-mini	reverse_category_order	nato	0.974	<0.001	—	53
Mistral	minimal	flyktingar	0.373	0.006	—	53
Mistral	minimal	försvarsutgifter	0.449	0.001	—	53
Mistral	minimal	kärnkraft	0.921	<0.001	—	53
Mistral	minimal	nato	0.207	0.136	—	53
Mistral	reverse_category_order	flyktingar	0.494	<0.001	—	53
Mistral	reverse_category_order	försvarsutgifter	0.581	<0.001	—	53
Mistral	reverse_category_order	kärnkraft	0.901	<0.001	—	53
Mistral	reverse_category_order	nato	0.113	0.419	—	53
Qwen3	minimal	flyktingar	0.404	0.003	—	53
Qwen3	minimal	försvarsutgifter	0.516	<0.001	—	53
Qwen3	minimal	kärnkraft	0.519	<0.001	—	53
Qwen3	minimal	nato	0.782	<0.001	—	53
Qwen3	reverse_category_order	flyktingar	0.555	<0.001	—	53
Qwen3	reverse_category_order	försvarsutgifter	0.455	0.001	—	53
Qwen3	reverse_category_order	kärnkraft	0.634	<0.001	—	53
Qwen3	reverse_category_order	nato	0.804	<0.001	—	53

Table 12: Mean party range and standard deviation (unweighted polarization measures) across all topics and years. Ratio columns show model values as proportion of RDU.

Model	Mean Range	Mean SD	Range/RDU	SD/RDU
RDU	2.879	1.101	1.000	1.000
GPT-4.1	2.762	0.983	0.959	0.893
GPT-5-mini	2.485	0.877	0.863	0.797
Mistral	1.698	0.632	0.590	0.574
Qwen3	1.267	0.456	0.440	0.414

10.11 Appendix K: Year-to-Year Polarization Changes

Year-to-year changes in the Dalton index identify periods of increasing or decreasing polarization. Table 13 presents the largest Dalton index shifts observed in RDU and whether models correctly identify the direction of change.

GPT-4.1 and GPT-5-mini correctly identify the direction of major polarization increases in migration policy, including the dramatic increase from 2006 to 2010 (RDU ΔDalton = +1.725) corresponding to the intensification of immigration debate, and from 2010 to 2014 (ΔDalton = +1.612) during the Sweden Democrats' rise. However, models show inconsistent directional tracking for NATO polarization shifts: both models fail to capture the increase from 2002 to 2006 (RDU ΔDalton = +1.249), though GPT-4.1 successfully tracks the decline from 2018 to 2022 (ΔDalton = -2.579). This mixed performance indicates that speech-based measures capture broad polarization trends but may miss specific temporal inflection points, particularly for security and defense issues where parliamentary speech may lag or lead public opinion shifts.

Mistral and Qwen3 show poor change detection: Mistral correctly identifies only 33% of major directional changes (2 of 6), while Qwen3 achieves 50% accuracy (3 of 6). Both models fail to capture the NATO polarization increase from 2002 to 2006, and show inconsistent performance on other security and energy issues. Full year-to-year change statistics available in `images/polarization_output/data/04_polarization_cl`

10.12 Prompting

The following is the prompt template used for retrieval-augmented generation of party position estimates from parliamentary speech:

```
Kontexten utgörs av utdrag från en politisk debattör i Sveriges Riksdag. Din uppgift är att

**Kategorier:**
1. "Mycket bra förslag" - Debattören uttrycker ett mycket starkt stöd för förslaget.
2. "Ganska bra förslag" - Debattören uttrycker sitt stöd för förslaget men medger samtidigt
3. "Varken bra eller dåligt förslag" - Debattören uttrycker sig neutralt till förslagett och
4. "Ganska dåligt förslag" - Debattören uttrycker sig avvisande till förslaget men belyser s
5. "Mycket dåligt förslag" - Debattören uttrycker sig starkt avvisande till förslaget.
```

Table 13: Major polarization changes (|ΔDalton| >1.0 in RDU) and model directional agreement. GPT models correctly identify some but not all major polarization shifts.

Topic	Period	RDU ΔDalton	GPT-4.1 ΔDalton	GPT-5-mini ΔDalton	Direction Agreement
flyktingar	2006→2010	+1.725	+2.561	+3.427	Both
flyktingar	2010→2014	+1.612	+1.542	+1.675	Both
nato	2002→2006	+1.249	-0.997	-0.155	GPT-4.1, GPT-5-mini
nato	2018→2022	-2.579	-2.383	-1.416	Both
kärnkraft	2014→2018	+1.197	-0.630	+0.741	GPT-4.1, GPT-5-mini
nato	1998→2002	-1.057	+1.377	+0.358	Both

```
6. "Ingen uppfattning" - Ingen åsikt eller otillräcklig information.

**Regler:**
 - Om debattören uttalar sig kring förslaget men det är oklart hurvida man anser att det är
 - Om kontexten saknar relevant information för att besvara frågan, välj "Ingen uppfattning'

**Kontext:**
{CONTEXT}

**Förslag:**
{QUERY}
```

10.12.1 Prompt variations:

Variant 1: Reverse Category Order (Test for Anchoring Bias) Your current prompt lists categories 1→ 5($very good \rightarrow very bad$).$LLM scan exhibit primacy effects, favoring earlier-listed options. Test whether reversing the order changes distributions$:

```
**Kategorier:**
1. "Mycket dåligt förslag" - Debattören uttrycker sig starkt avvisande till förslaget.
2. "Ganska dåligt förslag" - Debattören uttrycker sig avvisande till förslaget men belyser s
3. "Varken bra eller dåligt förslag" - Debattören uttrycker sig neutralt till förslaget och
4. "Ganska bra förslag" - Debattören uttrycker sitt stöd för förslaget men medger samtidigt
```

```
5. "Mycket bra förslag" - Debattören uttrycker ett mycket starkt stöd för förslaget.
6. "Ingen uppfattning" - Ingen åsikt eller otillräcklig information.
```

Variant 2: Minimal Instruction (Test for Over-Specification) Your current prompt provides detailed guidance ("Om debattören uttalar sig kring förslaget men det är oklart..."). Test whether simpler instructions produce equivalent results:

```
Analysera följande utdrag från en riksdagsdebatt. Bedöm debattörens ställning till det givna

Skala:
1 = Mycket bra förslag
2 = Ganska bra förslag
3 = Varken bra eller dåligt förslag
4 = Ganska dåligt förslag
5 = Mycket dåligt förslag
0 = Ingen uppfattning / ej relevant

Kontext:
{CONTEXT}

Förslag:
{QUERY}

Ange kategori (0 -5) och kort motivering.
```

References

L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, 2 2023. ISSN 1476-4989. doi: 10.1017/pan.2023.2. URL http://dx.doi.org/10.1017/pan.2023.2.

M. Arslan, S. Munawar, and C. Cruz. Political-RAG: using generative AI to extract political information from media content. *Journal of Information Technology & Politics*, 22(4):479–494, 2025. doi: 10.1080/19331681.2024.2417263. URL https://doi.org/10.1080/19331681.2024.2417263.

S. Axelsson, S. Källman, and S. Matti. *Hur polariserad är den svenska klimat- och energipolitiken? En jämförelse mellan väljare och valda.* Number 192 in Göteborg Studies in Politics. Statsvetenskapliga institutionen, Göteborgs universitet, Göteborg, första upplagan edition, 2025. ISBN 978-91-984403-3-1. URL https://gu.se/sites/default/files/2025-04/Riksdagens_landskap_kap4.pdf. Valforskningsprogrammet.

J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models.

Political Analysis, 32(4):401–416, 5 2024. ISSN 1476-4989. doi: 10.1017/pan.2024.5. URL http://dx.doi.org/10.1017/pan.2024.5.

R. J. Dalton. The Quantity and the Quality of Party Systems: Party System Polarization, Its Measurement, and Its Consequences. *Comparative Political Studies*, 41(7):899–920, 2 2008. ISSN 1552-3829. doi: 10.1177/0010414008315860. URL http://dx.doi.org/10.1177/0010414008315860.

S. Ekin. Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices. 5 2023. doi: 10.36227/techrxiv.22683919.v2. URL http://dx.doi.org/10.36227/techrxiv.22683919.v2.

J. Grimmer and B. M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013. ISSN 1047-1987. doi: 10.1093/pan/mps028. URL http://dx.doi.org/10.1093/pan/mps028.

M. LAVER, K. BENOIT, and J. GARRY. Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(02), 5 2003. ISSN 1537-5943. doi: 10.1017/s0003055403000698. URL http://dx.doi.org/10.1017/S0003055403000698.

A. Spirling. Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957):413–413, 4 2023. ISSN 1476-4687. doi: 10.1038/d41586-023-01295-4. URL http://dx.doi.org/10.1038/d41586-023-01295-4.

E. Strubell, A. Ganesh, and A. McCallum. Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, 4 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i09.7123. URL http://dx.doi.org/10.1609/aaai.v34i09.7123.

Y. Wang. Topic Classification for Political Texts with Pretrained Language Models. *Political Analysis*, 31(4):662–668, 3 2023. ISSN 1476-4989. doi: 10.1017/pan.2023.3. URL http://dx.doi.org/10.1017/pan.2023.3.

X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, and W. Han. Chatie: Zero-Shot Information Extraction via Chatting with ChatGPT, 2023. URL https://arxiv.org/abs/2302.10205.

X. Wu, R. Duan, and J. Ni. Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence*, 2(2):102–115, 3 2024. ISSN 2949-7159. doi: 10.1016/j.jiixd.2023.10.007. URL http://dx.doi.org/10.1016/j.jiixd.2023.10.007.

Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The

Ugly. *High-Confidence Computing*, 4(2):100211, 6 2024. ISSN 2667-2952. doi: 10.1016/j.hcc.2024.100211. URL http://dx.doi.org/10.1016/j.hcc.2024.100211.

C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291, 2024. ISSN 1530-9312. doi: 10.1162/coli_a_00502. URL http://dx.doi.org/10.1162/coli_a_00502.