Social Computing Methodology

Hongkun Zhang, Rong Bai, Yuanze Liu
Joshua Conrad Jackson

University of Chicago, University of Southern California,University of Illinois
Urbana-Champaign

August 2025

1 Introduction

Section lead: Yuanze

Values play a central role in guiding human decision-making at both individual and collective levels (Rokeach, 1973). Political speeches, in turn, serve as a unique arena where these guiding principles are articulated, contested, and transmitted to broader publics (Van Dijk, 1997). Yet systematic, multilingual analysis of values in political speech has remained elusive. We present a large language model–based pipeline that enables scalable cross-national annotation of values and implement it in parliament speeches across Europe. Our pipeline opens new possibilities for comparative research on how values shape political life.

1.1 The Significance of Values in Social Science Research

Values lie at the heart of many social science disciplines, including sociology (Weber, 2019), history (Berlin, 2013), economics (Ben-Ner and Putterman, 1998), political science (Feldman, 1988), and psychology (Schwartz, 1992; Sagiv and Schwartz, 2022). Although each tradition offers its own definition, most agree that values are abstract ideals that serve as guiding principles in people's lives (Schwartz and Bilsky, 1987; Rokeach, 1973; Maio, 2010).

By this definition, values can be distinguished from other related concepts that also shape behavior. First, values are inherently evaluative beliefs, reflecting subjective preferences and ideals (Rohan, 2000; Maio, 2010; Leung and Bond, 2004). Other concepts, such as knowledge or lay beliefs, may also guide behavior, but are not necessarily evaluative. For instance, the statement "hard work should be rewarded" is a value while the statement "hard work will be rewarded" is not. Second, values are more abstract than attitudes (Rohan, 2000; Schwartz, 1992). Although attitudes are also evaluative beliefs, they target specific objects or situations, whereas values represent fundamental principles that apply broadly across contexts. For example, a person who values equality as a guiding principle may, in a specific context, hold a positive attitude toward policies such as affirmative action. These features make values particularly important: they provide overarching guidelines that direct behavior across domains of social life, shaping how individuals perceive their environment, make decisions, and interpret outcomes in systematic, directional, and predictable ways (Rokeach, 1973; Sagiv and Roccas, 2021).

1.2 Political Values: Why They Matter

Political values are reflections of general human values in the political domain, and they occupy a central position in people's political life (Caprara and Zimbardo, 2004; Piurko et al., 2011). For ordinary people, political values provide the standards by which they evaluate politicians, political parties, policies, and political events (Feldman, 2003; Rathbun et al., 2016; Kam, 2005). Political values also structure a wide range of political behaviors, including voting decisions and participation in collective action (Caprara et al., 2006; Schwartz et al., 2010; van Stekelenburg and Klandermans, 2017). For political elites, values are equally consequential: they inform the priorities of policymaking (Young, 1977), shape governance styles (Kooiman and Jentoft, 2009), and guide responses to crises by providing enduring principal guidelines (Boin and Lodge, 2021).

Beyond individual-level attitudes and behaviors, political values can also function as powerful group signals in intergroup relations, shaping trust, perceived legitimacy, and dynamics of political coalition and conflicts (Nelson and Garst, 2005). For example, research shows that when people perceive congruence between their own political values and those embodied by institutions, they are more likely to view those institutions as legitimate and trustworthy (Grosfeld et al., 2022; Dowling and Pfeffer, 1975). In addition, research demonstrates that value conflicts themselves can produce intergroup hostility, fostering negative attitudes and distrust between groups who perceive one another as violating their fundamental values, regardless of political party or ideology (Brandt and Crawford, 2020). For insrance, experimental studies manipulating perceived value similarity find that participants show greater support for political candidates or groups framed as sharing their values (Crawford and Pilanski, 2014; Wetherell et al., 2013). Moreover, nationally representative surveys reveal that partisans vastly perceive out-partisans do not share core democratic values with them, with Democrats perceiving their peers to value democracy up to 77% more than

Republicans and Republicans perceiving their peers to value democracy up to 88% more than Democrats (Pasek et al., 2022).

1.3 Political Values are Distinct from Political Ideology

Political values are often compared with political ideologies, commonly defined as sets of beliefs about the proper order of society and how it should be achieved (Jost et al., 2009). Like values, ideologies are often evaluative in nature, specifying preferred societal arrangements and means of realizing them. Empirically, however, while models of values typically identify four to ten core dimensions (Schwartz et al., 2012; Inglehart and Baker, 2000), political ideologies are usually portrayed along a single left–right axis (Bobbio, 1996). This makes ideology a more parsimonious, yet possibly oversimplistic, more context-dependent depiction of people's potential variation space in political values.

A large body of research has criticized the left–right dichotomy for oversimplifying political differences, which are in fact multidimensional, heterarchical, intrapersonally eclectic, and contextually activated (Costello et al., 2023). Studies show that individuals' political values often form "mosaic-like" constellations rather than coherent left–right systems (Feldman and Johnston, 2014; Boutyline and Vaisey, 2017). Moreover, cross-cultural research further demonstrates that the left–right template, rooted in Western history (e.g., the French Revolution), is mainly applicable in western context, but often defied elsewhere (Malka et al., 2019). For example, studies find that the left-right dichotomy of political ideology corresponds to the sorting pattern of political values in Western European countries pretty well, but can barely explain the distribution of values in post-communist countries (Piurko et al., 2011; Beattie et al., 2022).

Beyond oversimplicity and context dependence, values are less tied to political sophistication than political ideologies. Value endorsement is often affect-driven, meaning that people rely more on intuitive emotional reactions, such as pride, guilt, or anger, than on elaborate cognitive reasoning when affirming their importance (Rohan, 2000; Maio, 2010). Due to this affective basis, the influence of values on political attitudes and behaviors is less restricted by levels of political knowledge. In contrast, classic survey research found that ideologically coherent belief systems were rare in the general public (Converse, 1964). Although more recent work suggests that this ideological incoherence has abated to some degree, there remains a broad consensus that party elites typically express more distinctive and coherent ideological beliefs than ordinary citizens (Goren et al., 2022; Kalmoe, 2020). Both elites and the public, however, hold values that are rooted in affect, and thus values provide a broader foundation for explaining political attitudes and behaviors across the population.

Empirical studies confirm the distinction of values relative to ideology. Feldman (1988) showed that core values such as equality and economic individualism predicted policy preferences and candidate evaluations even after controlling for ideology and partisanship. Similarly, Schwartz et al. (2010) found that values predicted voting behavior above and beyond ideological self-placement. These findings highlight the distinctive role of values in shaping political attitudes and behavior.

1.4 Values Taxonomies and Theoretical Frameworks

Instead of the dominance of a single-dimension model in political ideology studies, scholars of human values have developed several influential taxonomies and theoretical frameworks. In this paper, we primarily focus on one of the most widely used approaches—Schwartz's theory of basic personal values (Schwartz, 1992)—while also comparing it with other prominent models, including Inglehart's two-dimensional value framework (Inglehart and Baker, 2000), Hofstede's cultural dimensions (Hofstede, 1983), and the moral foundations model (Graham et al., 2009a).

1.4.1 Schwartz's Universal Values Theory

Schwartz's theory of basic human values aims to provide a universal, hierarchical structure of motivational goals that guide behavior across cultures (Schwartz and Bilsky, 1987; Schwartz, 1992). The theory begins from three basic human requirements—biological needs, social coordination, and group survival—that give rise to ten broad value types: self-direction emphasizes independent thought and action, such as choosing, creating, and exploring. Stimulation involves the pursuit of excitement, novelty, and challenge in life. Hedonism reflects pleasure and sensuous gratification for oneself. Achievement centers on personal success through demonstrating competence according to social standards. Power concerns social status and prestige, and control or dominance over people and resources. Security emphasizes safety, harmony, and stability of society, relationships, and self. Conformity entails the restraint of actions or impulses that might upset others or violate social expectations. Tradition involves respect, commitment, and acceptance of cultural or religious customs. Benevolence is about preserving and enhancing the welfare of people with whom one has frequent personal contact. Finally, Universalism involves understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.

These values are arranged into two higher-order dimensions—openness to change versus conservation, and self-enhancement versus self-transcendence—forming a circular motivational continuum in which adjacent values are compatible and opposing values are in tension. Later refinements further distinguished these ten value types into 19 more fine-grained dimensions, such as separating self-direction into "thought" and "action," or universalism into "tolerance," "concern," and "nature protection" (Schwartz et al., 2012). This hierarchical and circular model has been validated across diverse societies and provides a comprehensive framework for explaining cultural differences as well as political attitudes and behaviors (Sagiv and Schwartz, 2022).

1.4.2 Other Value Frameworks

There are some other influential approaches proposed to conceptualize basic human values. For example, Inglehart's model emphasizes macro-historical processes of modernization through the dimensions of traditional versus secular-rational values and survival versus self-expression values, focusing on shifts in religiosity, authority, and post-materialist orientations (Inglehart and Baker, 2000). Hofstede's cultural dimensions framework, originally developed in organizational research, highlights contrasts such as individualism–collectivism, power distance, masculinity–femininity, and uncertainty avoidance (Hofstede, 1983). Moral Foundations Theory (MFT) focuses more on moral intuitions, distinguishing between individualizing foundations (care, fairness) and binding foundations (loyalty, authority, sanctity) (Graham et al., 2009a). Each of these models illuminates important facets of cultural variation, and overlaps with Schwartz's model (Kaasa, 2021). Yet, they are often tailored to particular domains—modernization processes, workplace culture, or moral psychology—rather than providing a fully generalizable framework.

In this study, our aim is to develop an LLM-based annotation of the value of texts which is in principle compatible with any of these value models. We focus on Schwartz here because it has been most widely used and validated in cross-cultural surveys. In particular, the ten basic values can be organized into four higher-order clusters, and this structure has been consistently replicated across societies. This widespread adoption - especially in the European Social Survey (ESS) - makes it a conceptual backbone that is uniquely practical for comparative political research (Davidov et al., 2008a). By contrast, MFT has struggled to show a consistent dimensional structure across cultures (Atari et al., 2023), while Hofstede's dimensions are derived from a macro-level survey whose specific items often lack face validity (Blodgett et al., 2008).

1.5 The Value of LLMs for Advancing Our Understanding of Values

Section lead: Hongkun

1.5.1 Traditional Methods in Value Research Methods

Traditional approaches to measuring and analyzing values in political and social science research have relied primarily on three methodological paradigms: survey-based instruments, manual content analysis, and dictionary-based text analysis. While these methods have generated substantial insights into value structures and their political implications, their inherent limitations force researchers to choose between depth and breath.

Survey-Based Instruments

For decades, the dominant approach to measuring value has replied on standardized survey instruments, a paradigm that has generated substantial insights into value structure and their social implications. Instruments, like Schwartz Value Survey (SVS) and the Portrait Values Questionnaire (PVQ) (Schwartz et al., 2001; Schwartz, 2003), can enable researchers to collect comparable, standardized data from large and diverse samples. Major cross-national efforts built on this foundation have produced a rich and profound understanding of human value patterns. While this method is powerful for achieving breadth, its strengths are offset by inherent limitations in depth and flexibility.

These instruments face several fundamental and inevitable challenges. First, collecting survey data at scale is resource-intensive. Large, cross-national efforts like the European Social Survey or World Values Survey demand extensive funding, coordination, and time, limiting the frequency and geographic coverage of data collection(Heath et al., 2005). This is especially problematic when studying rapid value shifts triggered by political or economic shocks, where annual or multi-year wave survey cannot capture changes in real time. Second, survey-based methods are constrained by their a *priori* design. Researchers can only collect data on the limited set of value items designed in advance, which may fail to capture emergent or context-specific values that fall outside the pre-existing theoretical framework (Schuman and Presser, 1977; Schwartz, 2003). Third, meaning loss arising from linguistic and cultural heterogeneity is a persistent problem. Even carefully translated items may fail to convey equivalent meanings across contexts, undermining the validity of cross-national comparisons (Davidov et al., 2008b). The abstract nature of value concepts further exacerbates this challenge: terms such as ”tradition” or ”self-direction” can evoke fundamentally different associations in people’s mind across societies. Indeed, this shortcoming can be alleviated by surveys that define terms clearly or translate well, though such remedies demand considerable effort and unwavering rigor from the researcher.

Manual Content Analysis Limitations

As an alternative to surveys, content analysis of political texts allows researchers to study values in their naturally occurring context, offering deep, qualitative insights into how values are expressed and used in authentic discourse (Tetlock, 1983; Tetlock et al., 1984). This approach provides a powerful lens for fine-grained analysis of specific documents or speeches. While this method generates valuable depth, its inherent limitations force a trade-off against analytical breadth, reliability, and theoretical flexibility.

The most significant constraint of manual coding is its lack of scalability. The process is extraordinarily labor-intensive: training human coders to reliably identify value expressions requires substantial time, and the coding itself is slow and costly. Consequently, studies are typically confined to small corpora or narrow temporal windows, making it difficult to analyze large-scale patterns or track value dynamics over time.

Second, Ensuring inter-rater reliability is another challenge. Value are often expressed implicitly–in metaphors, justificatory language, or rhetorical strategies–rather than as explicit value statement, which forces coders to make subjective judgments (Entman, 1993). Consequently, extensive personnel training for calibration is required, further increasing resource demands.

Finally, the predetermined coding schemes required for manual analysis also impose theoretical constraints. Researchers must specify value categories in advance, potentially missing emergent or context-specific values that do not fit existing frameworks(Holloway and Todres, 2003; Braun and Clarke, 2006), especially for Western-derived taxonomies.

Dictionary-Based Approach Shortcomings

Computational text analysis using value dictionaries has emerged as a more scalable alternative, exemplified by tools like the Moral Foundations Dictionary (Graham et al., 2009b) and the Linguistic Inquiry and Word Count (LIWC) program (Pennebaker et al., 2015). These methods operate by calculating the usage intensity of keywords, following specific rules, from pre-compiled lists where each term is associated with a specific value or psychological construct, allowing researchers to quantify these concepts across vast text corpora (Grimmer and Stewart, 2013). Yet dictionary methods are fundamentally insensitive to context: the same word can signal different values depending on how it is used—“security” might mean national defense, economic stability, or personal safety—yet lexicons typically treat every instance the same. This context-insensitive property significantly dampens the depth of value excavation, forcing a trade-off where breadth is achieved at the expense of nuanced understanding.

Dictionary-based methods also struggle with temporal and cultural validity. Value expressions evolve over time as political discourse adapts to new issues and contexts, rendering static dictionaries increasingly obsolete (Garten et al., 2018). Cultural differences in value expression further complicate dictionary development, as direct translations often fail to capture culturally specific modes of expressing values (Van de Vijver and Leung, 2021).

The fixed nature of dictionary categories creates additional rigidity. Researchers must either accept predefined value dimensions that may not align with their theoretical framework or invest substantial effort in developing custom dictionaries, which then face validation challenges (Iliev et al., 2016). This inflexibility particularly limits exploratory research aimed at discovering novel value dimensions or configurations.

Cross-Cutting Methodological Challenges

Beyond method-specific limitations, traditional approaches share several overarching constraints. Most critically, they struggle to capture the dynamic, contextual nature of value expression in political discourse. Values are not simply present or absent but are emphasized, combined, and framed in complex ways that simple frequency counts or survey ratings cannot adequately represent (Sagi and Dehghani, 2014).

Furthermore, the inability to efficiently process large-scale, real-time data has prevented researchers from capturing value dynamics as they unfold. Traditional methods cannot feasibly analyze the millions of political texts generated daily across social media, news outlets, and government communications, leaving vast amounts of value-relevant data unexplored (Lazer et al., 2020).

Ultimately, existing methods force researchers into a fundamental trade-off between depth and breadth. Qualitative approaches can capture nuanced value expressions but lack scalability, while quantitative methods achieve scale at the cost of contextual sensitivity (Loughran and McDonald, 2011; Grimmer and Stewart, 2013). This trade-off has prevented the development of comprehensive value maps that combine fine-grained analysis with broad temporal and substantive scope.

These accumulated limitations have shaped a research landscape heavily reliant on a few large-scale, resource-intensive survey projects. While invaluable, this dependency has inadvertently narrowed the scope of empirical inquiry, often limiting analysis to general populations and pre-defined questions. This makes it difficult to study the values of specific groups, such as political elites, or to capture value expression as it unfolds in dynamic, real-world discourse. As a result, both theoretical development and the practical application of value research to political dynamics have been constrained. The emergence of large language models offers potential solutions to these challenges, as explored in the following section.

1.5.2 Large Language Models as a Methodological Solution

The emergence of large language models (LLMs) represents a paradigm shift in computational text analysis, offering solutions to many fundamental limitations that have constrained traditional value research methods (Ziems et al., 2024; Gilardi et al., 2023). Recent advances in transformer-based architectures and their training on massive text corpora have produced models capable of sophisticated semantic understanding, contextual interpretation, and flexible adaptation to diverse analytical tasks (Vaswani et al., 2017; Brown et al., 2020). These capabilities directly address the core challenges of value identification and quantification in political discourse.

Contextual Understanding

Unlike dictionary-based approaches that treat words as isolated units, LLMs process text through attention mechanisms that capture complex contextual relationships across entire documents (Devlin et al., 2019). This contextual processing enables LLMs to disambiguate value expressions based on surrounding discourse. For instance, when analyzing political speeches, LLMs can distinguish whether "freedom" refers to economic liberalization, civil liberties, or national sovereignty based on the broader argumentative context (Liu et al., 2019). Studies have demonstrated that LLMs can identify implicit value expressions that traditional methods miss, such as values embedded in policy justifications, metaphorical language, or historical references (Mendelsohn et al., 2023; Islam and Goldwasser, 2025; Tong et al., 2024). For example, Kiesel et al. (2022) note that many arguments convey values without naming any value words—for instance, “anyone who commits a crime should be prosecuted” implicitly appeals to values like law-abidingness and public safety—highlighting why context-sensitive models (e.g., LLMs) are needed beyond keyword dictionaries (Kiesel et al., 2022).

This contextual sensitivity extends to understanding value trade-offs and tensions within texts. Rather than simply counting value mentions, LLMs can detect when speakers acknowledge competing values, prioritize certain values over others, or reframe value conflicts (Park et al., 2024). This nuanced understanding is crucial for political discourse analysis, where value complexity and ambiguity are common features rather than exceptions.

Theoretical Flexibility

A critical advantage of LLMs is their capacity to adapt to multiple theoretical frameworks without extensive retraining or manual reconfiguration. Through prompt engineering, researchers can instruct LLMs to identify values according to Schwartz's basic values, Inglehart's materialist-postmaterialist dimensions, Moral Foundations Theory, or custom value taxonomies (Abdulhai et al., 2023; Hadar-Shoval et al., 2024; Zhu et al., 2025; Tao et al., 2024). This flexibility eliminates the need to develop separate measurement instruments for each theoretical approach. However, these methods remain framework-dependent at the point of analysis. They require researchers to commit to a single theoretical lens for each analytical run, creating inefficiencies for comparative research and potentially overlooking values that do not map cleanly onto the chosen framework. This limitation highlights the need for an approach that separates open-ended value identification from subsequent theoretical mapping, a core innovation of our proposed pipeline.

Recent work has further demonstrated the power of LLMs to work within and between pre-defined frameworks. For example, some approaches can reliably map between different value schemas, identifying conceptual overlaps that facilitate theoretical integration (Yao et al., 2024). Similarly, the zero-shot and few-shot learning capabilities of modern LLMs enable rapid adaptation to novel or culturally specific value concepts once they have been defined by the researcher (Kojima et al., 2022). While powerful, these applications still rely on a top-down approach where the analytical constructs are specified in advance, reinforcing the need for a more bottom-up, data-driven method for value discovery.

Scalable Analysis

LLMs have fundamentally transformed the scalability constraints of value research. Modern APIs can process millions of documents at speeds incomparable to human coding, while maintaining consistent analytical criteria across the entire corpus (Korinek, 2023). This scalability enables unprecedented research designs, such as analyzing value dynamics across decades of parliamentary debates or tracking real-time value shifts during political crises.

The cost-effectiveness of LLM-based analysis has democratized large-scale value research. (Gilardi et al., 2023) calculated that using GPT-3.5 for political text classification costs approximately 0.003% of equivalent human annotation, while achieving comparable accuracy. This dramatic cost reduction makes comprehensive value analysis feasible for researchers without access to large coding teams or substantial funding.

Scalability also enables iterative refinement of value measurement. Researchers can rapidly test different prompting strategies, adjust value definitions, or explore alternative theoretical frameworks across entire datasets (Ziems et al., 2024). This iterative capability facilitates methodological innovation and validation that would be prohibitively expensive with traditional approaches.

1.6 Preview of the Present Research

In brief, Schwartz values are central to parsing political rhetoric, and large language models provide a scalable, context-sensitive means of measuring them. Section 2 ("The Present Research") translates this motivation into a concrete design: a GPT-4o–based pipeline applied to a temporally balanced corpus of congressional speeches, mapping text data onto target theoretical dimensions. We also justify key technical choices (model selection, maximum-token threshold, and error handling) and preview reliability and validity assessments that reveal coherent internal structure and known-groups differences. In doing so, the present research operationalizes the paper's core claim—integrating contextual nuance with theoretical flexibility and scale—and establishes a reproducible approach for measuring political values in text.

2 The Present Research

2.1 Research Overview

This study introduces a novel computational pipeline for extracting and quantifying political values from textual data, addressing key limitations of existing approaches while leveraging recent advances in large language models. Our methodology represents a significant departure from traditional dictionary-based and manual coding approaches by combining the contextual sensitivity of human interpretation with the scalability and consistency of automated analysis.

The core innovation of our approach lies in its three-stage architecture that moves from open-ended value identification to systematic quantification. Rather than imposing predetermined value categories or relying on static keyword dictionaries, our pipeline first allows for the spontaneous identification of value concepts as they naturally emerge in political discourse. This initial flexibility is then channeled through a structured mapping process that connects these organically identified values to established theoretical frameworks, specifically Schwartz's theory of basic human values.

Our methodology employs OpenAI's GPT-4o model (OpenAI, 2024) for value identification and scoring. By using GPT-4o, we enable a more nuanced and contextually sensitive quantifica as the primary analytical engine, chosen for its demonstrated capabilities in complex text understanding and its ability to process nuanced semantic relationships. The selection of GPT-4o over alternative large language models reflects several key considerations: its superior performance on language understanding benchmarks (OpenAI, 2024), its ability to handle extended context windows effectively (Ryan, 2024), and its demonstrated reliability in structured output generation tasks (Croft, 2024). These capabilities directly align with our pipeline's core requirements: processing political speeches of varying lengths, distinguishing contextual value expressions from surface-level keyword mentions, generating structured JSON outputs with consistent formatting, and providing coherent explanations for analytical decisions across large-scale datasets.

The three-stage pipeline operates as follows: First, we identify the most salient political values expressed in each speech through guided prompting that encourages the model to recognize value-laden language combined with importance weights without imposing theoretical constraints. Second, we systematically map these spontaneously identified value terms to target theoretical dimensions through calibrated relevance scoring. Third, we aggregate these mappings using weighted averages that preserve the relative importance of different values as expressed in the original text.

Throughout all stages of the pipeline, we require GPT-4o to provide explicit justifications for its analytical decisions alongside its primary outputs. This methodological choice serves multiple critical functions. Research demonstrates that requiring explanations significantly improves the quality and reliability of large language model responses by encouraging more deliberate and systematic processing (Zhao et al., 2024). The explanations also provide crucial transparency, allowing researchers to understand the reasoning behind each analytical decision and identify potential systematic biases or errors in the extraction process. Furthermore, these justifications enable post-hoc validation and quality control, as researchers can evaluate whether the model's reasoning aligns with theoretical expectations and domain expertise. This explanatory requirement transforms what might otherwise be a "black box" analytical process into a transparent, interpretable system that maintains the benefits of automated analysis while preserving the accountability expected in rigorous social science research.

Our pipeline employs a carefully calibrated temperature setting of 0.2 for all API operations, reflecting the specific requirements of systematic social science research. Temperature controls the randomness in language model outputs, with lower values producing more deterministic and consistent responses while higher values increase creativity and variability (Noble, 2024). The selection of 0.2 represents an optimal balance between the competing demands of our analytical task: we require sufficient consistency to ensure that identical or similar speeches produce comparable value extractions across multiple runs, supporting the reliability and replicability essential to scientific inquiry. Simultaneously, we need enough variability to prevent overly rigid responses that might miss nuanced value expressions or fail to adapt to diverse rhetorical styles across our corpus. This moderate temperature setting ensures that our value extraction maintains the systematic consistency necessary for large-scale quantitative analysis while preserving the semantic flexibility required to capture the rich diversity of value expression in political discourse.

We validate this methodology through application to a carefully constructed dataset of 8,270 Congressional speeches spanning 2015-2024, selected through stratified temporal sampling to ensure balanced representation across this politically consequential decade. Our reliability and validity assessments demonstrate that the extracted value profiles exhibit theoretically expected patterns, including the characteristic circumplex structure of Schwartz's model and meaningful differences between Democratic and Republican speakers. (This paragraph should be modified according to our later results)

2.2 A Two-Step Approach to Extracting Values from Text

The development of our value quantification pipeline responds to fundamental limitations that have constrained existing approaches to measuring political values in textual data. A systematic review of contemporary methodologies reveals three critical gaps that our approach is designed to address.

First, existing LLM pipelines exhibit rigid theoretical framework dependencies that limit analytical flexibility. Contemporary approaches are typically designed around a single value taxonomy, requiring complete reconstruction to accommodate alternative frameworks. For example, recent work has developed LLM-based systems specifically for moral foundations (Abdulhai et al., 2023), separate pipelines exclusively for Schwartz's framework (Hadar-Shoval et al., 2024), and argument value identification systems around predetermined human values categories(Kiesel et al., 2022) . Researchers seeking to apply multiple theoretical lenses to the same corpus must entirely reconstruct analytical pipelines, reprocess datasets, and often retrain model components, creating significant inefficiencies in comparative research.

Our pipeline fundamentally addresses this inflexibility through its novel two-stage architecture that separates value identification from theoretical mapping. Unlike existing LLM approaches constrained to single theoretical frameworks, our method can be seamlessly adapted to measure any value system—whether Schwartz's ten basic human values, materialist/post-materialist orientations, moral foundations, or entirely novel value constructs—without reprocessing the original texts. The initial value identification step captures spontaneously expressed values without imposing theoretical constraints, creating a framework-agnostic inventory that can then be systematically mapped to various theoretical frameworks through modified prompting in the second stage, enabling efficient multi-framework analysis of the same corpus.

Second, direct measurement approaches suffer from bias amplification problems inherent in single-step prompting. When large language models are directly prompted to identify and score political values within a single analytical step, they manifest systematic tendencies to artificially agree with all value statements, likely stemming from social desirability bias. Research demonstrates that LLMs exhibit pronounced social desirability bias when directly prompted with assessments, systematically skewing responses toward socially desirable ends even when presented with individual questions (Salecha et al., 2024).

Third, existing methods struggle to capture dynamic contextual value expression. Dictionary-based approaches impose fixed value categories (Garten et al., 2018), while direct measurement methods often fail to recognize how values are embedded within argumentative frameworks, rhetorical strategies, and implicit appeals in policy justifications that predetermined lexicons cannot detect.

Our Two-Stage Solution

Our pipeline addresses these limitations through a novel architecture that separates value identification from theoretical mapping. The first stage identifies the most salient political values expressed in each speech through open-ended generation without imposing theoretical constraints, creating a framework-agnostic inventory of contextually-identified values with importance weights. This design directly enables theoretical flexibility by allowing the same value inventory to be seamlessly mapped to any theoretical framework—whether Schwartz's basic values, moral foundations, or materialist-postmaterialist dimensions—without reprocessing original texts or reconstructing analytical pipelines.

The second stage systematically maps these spontaneously identified values to target theoretical frameworks through calibrated relevance scoring. Crucially, this indirect approach achieves bias mitigation by preventing the model from knowing which specific values are being assessed during the initial identification phase, eliminating the systematic response bias that afflicts direct measurement approaches where models artificially agree with all value statements.

Throughout both stages, our method achieves contextual sensitivity by abandoning predetermined dictionaries entirely and allowing values to emerge organically from the data itself. The LLM processes entire speech contexts to understand how values are embedded within argumentative frameworks and rhetorical strategies, while importance weighting quantifies the relative prominence each value receives, ensuring measurements reflect actual political priorities rather than superficial word counts.

Unlike existing approaches that force researchers into suboptimal trade-offs between theoretical frameworks, measurement validity, and contextual nuance, our pipeline demonstrates that these limitations reflect design choices rather than inherent constraints of computational value measurement.

3 Methodology

We present a proof-of-concept implementation of our two-stage value-extraction pipeline using a sample of political speeches from the United States Congress (2015–2024). The purpose of this application is demonstrative—to establish the method's feasibility, scalability, and measurement quality (reliability and validity)—rather than to advance definitive substantive claims about U.S. political discourse. We use the U.S. Congress because it provides rich metadata, broad temporal coverage, and theory-informed expectations that aid validation. The approach itself is general and can be applied to political discourse in other linguistic and cultural contexts.

3.1 Data Processing

3.1.1 Data Sources

Our analysis draws on two primary data sources to construct a comprehensive dataset of political speeches with speaker biographical information. The first source is the Congressional Record dataset maintained by the United States project on GitHub (Judd et al., 2017), which provides structured JSON files containing the complete proceedings of both chambers of Congress. This dataset captures all official congressional speeches, debates, and proceedings in a standardized digital format, offering unprecedented access to the full scope of legislative discourse. The second source is the Congress Legislators dataset (unitedstates, 2025), also maintained by the United States project, which provides comprehensive biographical and political information for all members of Congress, including demographic attributes, political affiliations, and geographical representation.

From the Congressional Record repository, we extracted all speeches delivered between 2015 and 2024, focusing specifically on records classified as ”speech” with valid speaker bioguide identifiers, yielding a corpus of 363,849 speeches. Each speech record includes essential metadata such as volume number, session date, chamber location, page references, and document structure, alongside the complete speech text, speaker information, and sequential positioning within the proceedings.

To transform these speech records into analytically meaningful data with comprehensive legislator profiles, we integrated the Congressional Record data with the biographical information from the Congress Legislators dataset using bioguide identifiers as the linking key. This integration process enriches each speech with detailed demographic attributes such as gender and age, political affiliations, and geographical representation including state and district information, enabling sophisticated analysis of how speaker characteristics relate to value expression in political discourse.

3.1.2 Data Processing Strategy

Our data processing followed a systematic three-stage approach designed to balance analytical comprehensiveness with computational efficiency. In the initial stage, we converted the raw JSON files into a structured CSV format, creating individual rows for each speech while preserving all relevant metadata. This transformation resulted in a denormalized structure where document-level information such as session details and chamber location was replicated across all speeches from the same proceeding, ensuring that contextual information remained accessible for each individual speech while facilitating subsequent analysis.

The second stage involved four critical processing steps. First, we consolidated the annual congressional record files chronologically into a unified dataset spanning the entire temporal range while maintaining referential integrity between sessions and years. Second, we enriched our speech records through biographical integration, implementing an inner join strategy that matched congressional speeches to legislator profiles using bioguide identifiers. This approach prioritized data completeness by excluding speeches from legislators not present in the biographical database, ensuring that all retained records included comprehensive demographic and political information necessary for meaningful analysis.

Third, we quantified text length using tokenization algorithms consistent with our target language model (GPT-4o), calculating token counts for each speech to enable downstream processing decisions and computational cost estimation. This tokenization step was essential because our main value extraction pipeline relies on the OpenAI GPT-4o API, which processes and bills text based on token units rather than word counts, making accurate token quantification crucial for both technical implementation and resource planning.

Fourth, we implemented an optional speech-consolidation step to maximize information density while reducing compute time and cost on large corpora. The core pipeline does not require consolidation; it is purely a scalability optimization and can be disabled without affecting methodological validity. Concretely, the procedure identifies speeches with identical contextual metadata (same chamber/session/day and conversational/legislative context) and applies a bin-packing rule to group them chronologically under a 10,000-token cap per consolidated segment. Using this method, each consolidated entry contains as much information as possible while preserving the logical flow across utterances, because we respect both contextual alignment and strict temporal order during implementation. For example, if Senator Smith delivered four speeches in the same Senate session on the same day with token counts of 2,500, 3,200, 2,800, and 2,100, the first three (8,500 tokens) would form one segment and the fourth would start a new segment. The 10,000-token cap balances content richness with processing efficiency; we justify this choice in the next section. Speeches that individually exceed the cap are retained as singleton segments.

Stratified Sample

The final stage implemented a two-step sampling strategy to produce a temporally balanced dataset while trimming clearly non-substantive items. First, we applied a conservative filter that removes speeches with fewer than 20 tokens, which are typically procedural and too short to contain analyzable content (see examples in Table 1). Applying this rule reduced the dataset from 208,321 to 205,957 records (a 1.13% decrease). Second, we then employed temporal stratified sampling, extracting exactly 1,000 records per year through simple random sampling within each annual stratum, using a fixed random seed to ensure reproducibility.

Table 1: Illustrative examples: very short speeches vs. consolidated segments (two utterances)

#	**Very short speeches (<20 tokens)**	**Consolidated samples (two utterances)**
1	Mr. SCHATZ. I ask for the yeas and nays.	Mr. McCAIN. Mr. President, I ask unanimous consent that the Senate be in a period of morning business, with Senators permitted to speak therein for up to 10 minutes each. Mr. McCAIN. Mr. President, I ask unanimous consent that the Senator from South Carolina and I be permitted to engage in a colloquy.
2	Mr. DUNN. Mr. Speaker, I demand a recorded vote.	Mr. WHITEHOUSE. Mr. President, I ask unanimous consent that the Committee on the Judiciary be discharged from further consideration of H.R. 1002 and the Senate proceed to its immediate consideration. Mr. WHITEHOUSE. I ask unanimous consent that the bill be considered read a third time and passed and that the motion to reconsider be considered made and laid upon the table.
3	Mr. McCONNELL. What is the pending business?	Mr. McCONNELL. Mr. President, I ask unanimous consent that the Senate proceed to the immediate consideration of Calendar No. 364, H.R. 1660. Mr. McCONNELL. I ask unanimous consent that the bill be read a third time and passed and that the motion to reconsider be considered made and laid upon the table with no intervening action or debate.

Notes. Left: examples of sub-20-token utterances that are typically procedural and non-substantive. Right: consolidated segments constructed by grouping utterances with matched contextual metadata in strict chronological order under a 10,000-token cap. We display pairs here for space; longer consolidated segments are analogous.

Our stratified sample comprises 10,000 congressional speeches evenly distributed across the decade from 2015 to 2024, with exactly 1,000 speeches per year. This represents approximately 4.9% of the filtered dataset and ensures balanced temporal representation regardless of variations in yearly congressional activity levels. The stratified sampling design enables robust longitudinal analysis while preventing any single year from dominating the dataset.

API-Based Content Filtering

Despite our comprehensive metadata-based consolidation strategy designed to maximize the informational content of each speech entry, we observed that some texts classified as ”speech” in the Congressional Record still lacked substantive political content suitable for value analysis. To ensure the highest data quality for our value extraction pipeline, we implemented an additional API-based content filtering system applied to our final stratified sample of 10,000 speeches.

This filtering process employed GPT-4o-mini to distinguish between substantive political discourse and procedural statements through structured semantic analysis. The system was designed to remove purely procedural content (voting procedures, scheduling announcements, administrative remarks), formal ceremonies (oath-taking, ceremonial greetings), and technical corrections while preserving any speech containing policy discussions, political arguments, social commentary, or legislative debates, regardless of their length.

Our choice of GPT-4o-mini for this content filtering task was guided by both methodological and practical considerations. Research demonstrates that smaller language models can achieve comparable performance to larger models on focused classification tasks such as content filtering and text categorization, particularly when the task does not require complex reasoning or extensive domain knowledge (Chae and Davidson, 2025; OpenAI, 2024). GPT-4o-mini, scoring 82% on the MMLU benchmark while being significantly more cost-efficient than larger models, represents an optimal balance between classification accuracy and computational efficiency for binary content filtering tasks. The model's pricing at 15 cents per million input tokens compared to substantially higher costs for full-scale models makes it particularly well-suited for large-scale content processing applications where cost efficiency and processing speed are prioritized over complex reasoning capabilities.

The filtering system employed structured prompting to generate binary keep or remove decisions accompanied by explanatory reasoning for each judgment.

This API-based filtering process resulted in the removal of 1,730 entries from our 10,000-speech sample, yielding a final analytical dataset of 8,270 speeches suitable for comprehensive political value analysis. The filtered content consisted primarily of brief procedural acknowledgments, ceremonial statements, and administrative remarks that, while officially classified as speeches, contained insufficient substantive political content for meaningful value extraction.

Importantly, the final sample maintains temporal balance across the study period, with the year-wise distribution as follows: 2015 (862 speeches), 2016 (868), 2017 (852), 2018 (821), 2019 (826), 2020 (817), 2021 (806), 2022 (808), 2023 (816), and 2024 (794). This balanced temporal representation ensures robust longitudinal analysis capabilities while preserving the integrity of our stratified sampling design, creating an ideal foundation for examining temporal patterns in political discourse and values expression over this consequential decade in American politics.

3.1.3 Maximum Token Determination

The selection of an optimal maximum token threshold represents a critical methodological decision that directly impacts both the quality and efficiency of our value extraction pipeline. Large language models exhibit varying performance characteristics across different input lengths, with potential trade-offs between contextual comprehension and processing consistency (Levy et al., 2024). To empirically determine the optimal threshold for our GPT-4o-based value extraction system, we conducted a systematic controlled experiment that tested model performance across a range of input lengths.

Experimental Design

Our token determination experiment followed a four-phase approach designed to isolate the effects of input length on value extraction consistency while controlling for content variation and methodological factors.

Phase 0: Sample Construction

We began by applying our speech consolidation algorithm (described in Section 3.1.2) to the complete Congressional Record dataset (2015-2024), but without imposing any maximum token threshold. This consolidation process, which groups speeches sharing identical contextual metadata using a bin-packing approach, produced a dataset where some entries represented individual long speeches while others consisted of multiple related utterances merged into coherent discourse segments.

From this consolidated dataset, we identified all entries exceeding 10,000 tokens, yielding 152 long-form texts after initial filtering. Critically, this sample included both naturally lengthy individual speeches and consolidated speech segments created by combining multiple shorter utterances from the same legislative context. These texts underwent our API-based content quality assessment, resulting in 149 substantively political speech segments suitable for experimentation. From this pool, we randomly sampled 50 texts to ensure manageable computational costs while maintaining statistical power for consistency analysis. This sample comprised 7 individual speeches that naturally exceeded 10,000 tokens without any consolidation processing, and 43 consolidated segments created through our metadata-based speech consolidation algorithm.

This sampling strategy enabled us to test model performance on identical content across varying input lengths, eliminating confounding factors that might arise from comparing different speeches or speakers. Importantly, our test sample reflected the real-world diversity of our processing pipeline, including both individual lengthy speeches and consolidated multi-utterance segments, ensuring that our token threshold determination would be valid for both types of input data our system would encounter in practice.

Phase 1: Systematic Content Slicing

We implemented a sophisticated slicing strategy that created twenty distinct token-length conditions ranging from 500 to 10,000 tokens in 500-token increments. Using the GPT-4o tokenizer for precise boundary detection, we divided each speech into consecutive 500-token segments, then systematically merged these segments to create cumulative datasets:

- merged_500: First segment only (500 tokens)
- merged_1000: First two segments (1,000 tokens)
- merged_1500: First three segments (1,500 tokens)
- *continuing through*
- merged_10000: First twenty segments (10,000 tokens)

This cumulative approach ensured that each higher token threshold contained all content from lower thresholds plus additional context, allowing us to isolate the pure effects of increased input length on model performance. The consistent ordering across thresholds eliminated potential confounds from content selection or arrangement.

Phase 2: Consistency Testing Protocol

For each of the 1,000 text segments (50 speeches × 20 token thresholds), we conducted value extraction using our standard GPT-4o pipeline with one crucial modification: each segment was processed three times with identical prompts to assess consistency across multiple runs. This triple-run strategy enabled robust measurement of extraction stability, a critical factor for methodological reliability.

The structured prompt requested identification of 1-5 political values in JSON format—identical to the prompt used in our main value extraction pipeline. The expected output consisted of three parallel arrays: the first containing extracted value terms (such as ”value1”, ”value2”, etc.), the second containing corresponding numerical weights that sum to 1.0, and the third containing explanatory reasons for each identified value.

Quality controls ensured weights summed to 1.0 and maintained equal-length arrays across all components, with automatic retry mechanisms handling failed extractions.

Phase 3: Semantic Consistency Measurement Rather than relying on simple string matching, we implemented a sophisticated consistency measurement approach combining semantic embedding with optimal matching algorithms. This methodology addresses the fundamental challenge that identical political values may be expressed using different terminology across API runs (e.g., ”environmental protection” vs. ”ecological safety”).

BERT Embedding and Similarity Calculation

We employed the SentenceTransformer model ('all-MiniLM-L6-v2') to encode all extracted value terms as high-dimensional semantic vectors, enabling measurement of conceptual similarity rather than mere lexical overlap (Wang et al., 2020). This model was selected for several methodological reasons: first, it was specifically optimized for semantic textual similarity tasks and demonstrates strong performance on standardized similarity benchmarks while maintaining computational efficiency through knowledge distillation. Second, unlike static word embeddings that treat each word independently, sentence transformers capture contextual meaning and can effectively handle multi-word value expressions such as ”environmental protection” or ”economic prosperity” (Devlin et al., 2019). Third, the teacher model's (Liu et al., 2019) training on diverse text corpora enables robust semantic understanding across different domains, including political discourse. This diverse pre-training enables the distilled model to develop robust semantic understanding across varied contexts, thereby supporting our capability to analyze value-laden language in this setting. For each triplet of value sets extracted from the same text segment, we computed pairwise cosine similarity matrices between all combinations of values.

Hungarian Algorithm for Optimal Matching

The core innovation in our consistency measurement involved applying the Hungarian algorithm to solve the optimal bipartite matching problem between value sets. This approach addresses a key methodological challenge: when comparing two sets of extracted values, how do we determine which values correspond across runs when they may be expressed differently?

Example of Optimal Matching: Consider a speech segment where three API runs extract the following value sets:

- Run 1: [”economic growth”, ”equality”, ”innovation”, ”national security”]
- Run 2: [”prosperity”, ”social justice”, ”technological advancement”, ”public safety”]
- Run 3: [”financial development”, ”fairness”, ”modernization”, ”defense”]

Using BERT embeddings, we calculate semantic similarity between all cross-run pairs:

- ”economic growth” ↔ ”prosperity”: 0.544
- ”economic growth” ↔ ”financial development”: 0.482
- ”equality” ↔ ”social justice”: 0.404
- ”equality” ↔ ”fairness”: 0.472

The Hungarian algorithm then finds the optimal one-to-one matching that maximizes overall similarity across the value set, preventing issues where multiple values from one run might be matched to a single highly similar value in another run.

Multi-dimensional Consistency Scoring

For each speech segment, we calculated pairwise consistency scores between all three API runs using the optimal matching similarities. The overall consistency score represents the mean of these three pairwise similarities, providing a comprehensive measure of extraction stability across multiple API calls.

Empirical Findings

Our systematic analysis revealed that GPT-4o maintains remarkably stable performance across input lengths up to 10,000 tokens, with overall consistency scores consistently hovering around 0.9 across the full range from 500 to 10,000 tokens fig. 1. This stability demonstrates that the model effectively processes extended political discourse without substantial information loss or performance degradation, unlike some language models that show declining accuracy on longer inputs.

The empirical results directly support our selection of 10,000 tokens as the optimal threshold. While GPT-4o performance remains stable beyond this point, processing costs and time increase linearly with input length, making 10,000 tokens an economically efficient boundary that balances comprehensive contextual analysis with computational practicality. Additionally, analysis of our Congressional dataset revealed that this threshold captures the vast majority of individual speeches while effectively accommodating our consolidation strategy for grouping related legislative discourse.

This empirically-grounded threshold selection provides a principled foundation for large-scale political value analysis, ensuring that our methodology reflects genuine model performance characteristics rather than arbitrary cutoffs or convenient approximations.

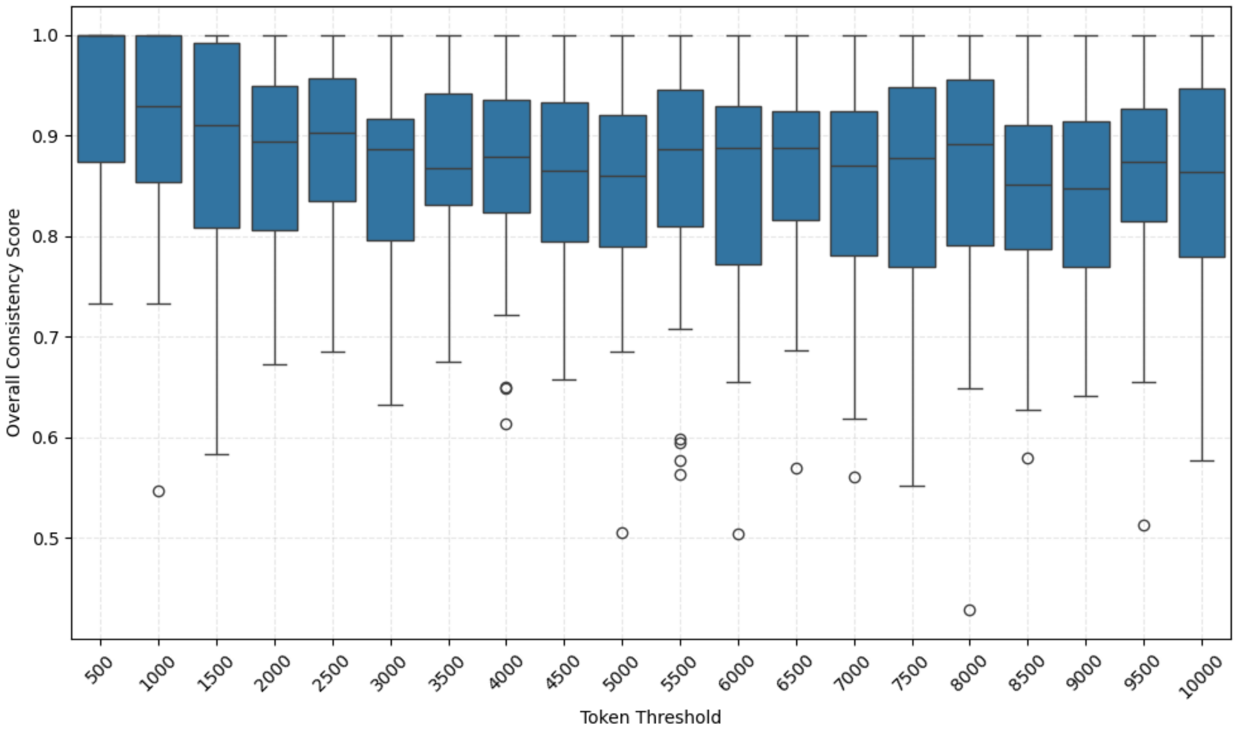

Figure 1: Overall Consistency by Token Threshold

3.2 Value Quantification Pipeline for Political Speeches

As noted earlier, our pipeline consists of three main steps: Value Word Identification and Weighting, Value Quantification, and Weighted Score Aggregation. The following provides an illustration of each step:

Step 1: Value Word Identification and Weighting

For each political speech, we first identify the most relevant value words and their relative importance through a computational text analysis pipeline. Rather than imposing predetermined value categories, we prompt a large language model (OpenAI's API) to identify the most salient value concepts—at least one and up to five—expressed in each speech. We did not allow GPT the option to refuse generating related value words, since our API filtering procedure already ensures that each text segment contains substantive political content.

Crucially, this stage also assigns importance weights to each identified value, reflecting the relative emphasis placed on different values within the speech context. Importance weights are numerical values (ranging from 0 to 1 and summing to 1.0 across all identified values) that quantify the relative prominence or emphasis each value receives within the specific speech context, enabling proportional representation of differential value emphasis in the final quantification.

This weighting mechanism addresses a fundamental limitation of binary presence-absence approaches: political speeches do not treat all mentioned values with equal prominence. A speaker might briefly acknowledge ”security concerns” while devoting substantial rhetorical attention to ”economic prosperity,” and our quantification should reflect this differential emphasis. The importance weights ensure that values receiving greater attention in the original discourse have proportionally greater influence on the final value profile, leading to more accurate and contextually sensitive measurements of political priorities.

This approach provides the model with sufficient flexibility, while maintaining appropriate constraints to ensure the extraction of meaningful values without generating an excessive number of irrelevant concepts. The addition of importance weighting enables more precise quantification by acknowledging that different values may be expressed with varying degrees of emphasis within a single speech.

Here is our definition of value:

Values are beliefs about desirable goals in life and modes of conducting oneself that guide how we evaluate behaviors, people, and events (Schwartz, 1994).

Specifically, we process each speech through the following procedure:

Value identification and weighting: We prompt the language model to identify the most prominent political values expressed in the speech using a carefully constructed prompt that avoids leading the model toward particular value frameworks. The model is required to output three components for each speech: (1) a list of value words or phrases (e.g., ”justice,” ”environmental safety,” ”national security,” ”economic prosperity”), (2) corresponding importance weights that sum to 1.0, reflecting the relative emphasis of each value within the speech, and (3) clear explanations for each identified value. Example outputs are shown in table 2.

Table 2: Output Examples from Step 1 – Value Word Identification and Weighting

Text	Values	Weights	Reasons
Mr. SCHUMER. Madam President, pursuant to S. Res. 27, the Committee on Health, Education, Labor, and Pensions being tied on the question of reporting, I move to discharge the Committee on Health, Education, Labor, and Pensions from further consideration of the nomination of Jennifer Ann Abruzzo, of New York, to be General Counsel of the National Labor Relations Board. Mr. SCHUMER. I yield the floor.	[‘democratic process’, ‘responsibility’, ‘fairness’]	[0.5, 0.3, 0.2]	[‘The speech reflects the value of democratic process as it involves a formal motion to discharge a committee, indicating adherence to procedural rules and legislative processes.’, ‘The value of responsibility is reflected in the act of moving to discharge the committee, which suggests a sense of duty to ensure that the nomination is considered and not stalled.’, ‘Fairness is a value reflected in the context of the speech, as the motion to discharge the committee from further consideration implies a need to ensure that the nomination process is conducted equitably, without unnecessary delays.’]
Ms. CASTOR of Florida. Mr. Speaker, I ask unanimous consent to bring up H.R. 1217, the bipartisan expanded background checks legislation, to honor the memory of Martavious Carn, age 3, a Florida victim of gun violence who never received a moment of action on the House floor.	[’justice’, ’safety’, ’remembrance’]	[0.4, 0.4, 0.2]	[’The call for expanded background checks legislation reflects a desire for justice, aiming to address and rectify the issue of gun violence, particularly in memory of a young victim.’, ’The mention of expanded background checks highlights the value of safety, as it seeks to prevent future incidents of gun violence and protect the community.’, ’The speech honors the memory of Martavious Carn, emphasizing the value of remembrance by acknowledging the loss and ensuring it is not forgotten.’]

Step 2: Value Quantification

After identifying the most relevant value words for each speech, we construct a political value dictionary that maps each value word onto Schwartz’s basic value dimensions using the Portrait Values Questionnaire (PVQ-21) framework. Rather than directly mapping to the ten basic value dimensions, we employ the 21 specific value items from the PVQ-21 instrument (Schwartz et al., 2001), which provide more granular measurement precision before aggregation into the ten higher-order dimensions.

To quantify these associations, we employ an API-based rating approach, assigning a relevance score between 1 and 6 to each value word for each of the 21 PVQ items (where 1 indicates ”not like [value word] at all” and 6 indicates ”very much like [value word]”). This 6-point scale mirrors the standard PVQ response format, ensuring methodological consistency with established value measurement protocols.

To generate these ratings, we use a precisely calibrated prompt that asks the language model to evaluate how closely each identified value word aligns with each of the 21 PVQ items, such as ”Important to think new ideas and being creative” (ipcrtiv) or ”Important to live in secure and safe surroundings” (impsafe). Crucially, the model must provide both a numerical score and explicit reasoning for each rating, enhancing transparency and enabling quality control. These 21 item-level scores are subsequently aggregated into the ten basic Schwartz value dimensions following standard PVQ-21 scoring procedures, creating comprehensive value profiles that maintain both granular precision and theoretical coherence.

Table 3: Value Words Relevance by Dimension

Dimension Name	**Dimension Content**	**5 Most Relevant Words**	**5 Least Relevant Words**
ipcrtiv	Important to think new ideas and being creative	academic and intellectual freedom, professional relevance, progress and reform	legal simplicity, constitutional loyalty, historical pride
imprich	Important to be rich, have money and expensive things	economic success, private interest, critique of elitism	legal sovereignty, legal simplification, opposition to populism
ipeqopt	Important that people are treated equally and have equal opportunities	global thinking, social equality, social consensus	skepticism towards new technology, skepticism towards opposition, militarism
ipshabt	Important to show abilities and be admired	attention, admiration, career achievement	regulatory intervention, commitment to legal framework, constitutionality
impsafe	Important to live in secure and safe surroundings	national priorities, order and stability, public health and well-being	freedom_of_expression, freedom_of_speech, pioneering
impdiff	Important to try new and different things in life	innovation and exploration, evolution and change, innovation	traditional customs, respect for tradition, respect for customs
ipfrule	Important to do what is told and follow rules	legal authority, legal obligation, constitutionality	non-conformity, ethical non-conformity, opposition
ipudrst	Important to understand different people	cultural diversity, global thinking, international cooperation	nationalism, legal sovereignty, national interest
ipmodst	Important to be humble and modest, not draw attention	modesty, humility, moderation	attention, economic success, admiration
ipgdtim	Important to have a good time	leisure and entertainment, enjoyment, pleasure	asceticism, sacrifice, commitment to hard work
impfree	Important to make own decisions and be free	freedom of choice, individual freedom, freedom_of_expression	authority and control, authoritarian control, legal obligation
iphlppl	Important to help people and care for others’ well-being	humanitarian aid, social support, public health and well-being	self-interest, competition, individualism
ipsuces	Important to be successful and that people recognise achievements	career achievement, achievement and success, economic success	failure acceptance, acceptance of limitations, modesty

Continued on next page

Table 3 – continued from previous page

Dimension Name	Dimension Content	5 Most Relevant Sample Words	5 Least Relevant Sample Words
ipstrgv	Important that government is strong and ensures safety	governmental authority, legal authority, order and stability	anarchism, anti-government sentiment, freedom from government
ipadvnt	Important to seek adventures and have an exciting life	adventure, innovation and exploration, excitement	routine, stability preference, predictability
ipbhprp	Important to behave properly	proper behavior, social norms, appropriate conduct	rebelliousness, non-conformity, unconventional behavior
iprspot	Important to get respect from others	respect and recognition, social status, admiration	humility, anonymity, modesty
iplylfr	Important to be loyal to friends and devote to people close	loyalty, friendship, personal relationships	betrayal, disloyalty, self-interest over relationships
impenv	Important to care for nature and environment	environmental protection, nature conservation, sustainability	environmental destruction, resource exploitation, disregard for nature
imptrad	Important to follow traditions and customs	traditional values, respect for tradition, cultural heritage	innovation over tradition, modernization, change over continuity
impfun	Important to seek fun and things that give pleasure	fun and enjoyment, pleasure seeking, recreational activities	seriousness over fun, duty over pleasure, restraint

Step 3: Weighted Score Aggregation Finally, we calculate each speech's scores on the 10 Schwartz value dimensions through a two-stage weighted aggregation process. First, we compute weighted averages of the PVQ-21 item scores for each speech's most representative political value words, using the importance weights determined in Step 1. Second, we aggregate these weighted PVQ-21 item scores into the ten basic Schwartz value dimensions following standard PVQ-21 scoring procedures (Schwartz et al., 2001). This represents a significant methodological improvement over simple arithmetic averaging, as it accounts for the varying emphasis placed on different values within each speech while maintaining measurement precision through the granular PVQ-21 framework.

For example, if a speech's three most salient value words are "justice" (weight: 0.5), "environmental safety" (weight: 0.3), and "law's power" (weight: 0.2), we first compute the weighted average of their respective scores across each of the 21 PVQ items, then aggregate these weighted item scores into the 10 Schwartz dimensions using the standard PVQ-21 mapping shown in Table 4. This weighted aggregation approach ensures that values with greater prominence in the original speech text have proportionally greater influence on the final value profile, leading to more accurate and contextually sensitive quantification of political values.

The weighted aggregation is computed as:

$$\text{PVQ\_Item\_Score}_{\text{item}} = \frac{\sum_{i=1}^{n} w_i \times s_{i,\text{item}}}{\sum_{i=1}^{n} w_i}$$

where w_i represents the importance weight of value word i, $s_{i,\text{ item}}$ represents the relevance score of value word i for PVQ item item, and n is the number of value words identified for the speech. The resulting PVQ item scores are then aggregated into Schwartz dimensions following the mapping in table 4, where each dimension score is calculated as the arithmetic mean of its constituent items (e.g., Security = mean(impsafe, ipstrgv), Universalism = mean(ipeqopt, ipudrst, impenv)).

Table 4: PVQ-21 items used for each value

10 Schwartz Values	21 PVQ Items
Security	impsafe_Judge, ipstrgv_Judge
Conformity	ipfrule_Judge, ipbhprp_Judge
Tradition	ipmodst_Judge, imptrad_Judge
Benevolence	iphlppl_Judge, iplylfr_Judge
Universalism	ipeqopt_Judge, ipudrst_Judge, impenv_Judge
Self-Direction	ipcrtiv_Judge, impfree_Judge
Stimulation	impdiff_Judge, ipadvnt_Judge
Hedonism	ipgdtim_Judge, impfun_Judge
Achievement	ipshabt_Judge, ipsuces_Judge
Power	imprich_Judge, iprspot_Judge

4 Reliability

Lead: Hongkun and Rong

4.1 Aggregation to Higher-Order Domains (10 → 4)

To summarize the ten values, we form four higher-order composites along two canonical bipolar axes: Openness-to-Change vs. Conservation and Self-Transcendence vs. Self-Enhancement. Because Hedonism lies on the border between Openness and Self-Enhancement, we report two conventional variants: Variant A places Hedonism with Openness; Variant B places Hedonism with Self-Enhancement. All composites are available-case, unweighted means of their constituent dimensions.

Table 5: Higher-order composites (two variants)

Composite	Variant A (Hedonism → Openness)	Variant B (Hedonism → Self-Enhancement)
Conservation	Security, Conformity, Tradition	Security, Conformity, Tradition
Self-Transcendence	Benevolence, Universalism	Benevolence, Universalism
Openness-to-Change	Self-Direction, Stimulation, Hedonism	Self-Direction, Stimulation
Self-Enhancement	Achievement, Power	Achievement, Power, Hedonism

4.2 Estimands and Estimation

We estimate internal consistency with three complementary indices and report them across composites:

- **Cronbach's α** — descriptive lower bound under (essential) tau–equivalence.
- **McDonald's ω_t (total)** — congeneric one–factor reliability allowing unequal loadings.
- **Split–half (Spearman–Brown)** — average over 200 random splits with prophecy correction.

For each composite we report point estimates and row–bootstrap 95% CIs ($B = 200$).[1] Because items are Likert–type (1–6), we add an *ordinal* robustness check (rank–based/ polychoric variants) alongside Pearson to verify conclusions are unchanged.[2]

Two–indicator composites (k=2) receive special handling: standardized α equals the inter–item correlation and does *not* directly reflect the reliability of the *average* of two indicators, so we emphasize the

[1] Bootstrap: Efron and Tibshirani (1993).

[2] Ordinal reliability guidance: Gadermann et al. (2012); Zumbo et al. (2007).

Spearman–Brown reliability of the two–item mean and treat ω_t as descriptive in these cases (Eisinga et al., 2013; Revelle and Zinbarg, 2009). For interpretability we also report the standard error of measurement (SEM) using ω,

$$\text{SEM} \; = \; \text{SD} \times \sqrt{1 - \omega},$$

where SD is the observed standard deviation of the composite.

Table 6: Internal-consistency indices at a glance

Index	**What it estimates**	**Key assumptions**	**When to prefer**	**How we use it here**
Cronbach's α	Lower-bound internal consistency of a summed score	Essential tau–equivalence; uncorrelated errors	Quick descriptive bound; widespread comparability	Report point estimate + bootstrap 95% CI; Pearson main, ordinal check; caution for k=2[a]
McDonald's ω_t	Reliability under congeneric one–factor model (unequal loadings allowed)	Unidimensional latent structure; freely estimated loadings	Items with differing loadings; SEM/latent-style reliability	Primary coefficient for $k \geq 3$ composites; compute SEM via ω; Pearson main, ordinal check[b]
Split–half (SB)	Correlation of two parallel halves, corrected to full length	Depends on split; SB prophecy corrects length	Two–item composites; sanity check for longer scales	Average over 200 random splits; report SB mean/median and CI; emphasize for k=2[c]

[a] Cronbach (1951); Sijtsma (2009); McNeish (2018).

[b] McDonald (1999); Revelle and Zinbarg (2009); Raykov (1997); Green and Yang (2009).

[c] Spearman (1910); Brown (1910); Eisinga et al. (2013). Bootstrap CIs follow Efron and Tibshirani (1993). Ordinal robustness per Olsson (1979); Zumbo et al. (2007); Gadermann et al. (2012).

4.3 Results

Under Variant A (Hedonism → Openness), internal consistency is acceptable for Conservation and strong to excellent for the remaining domains. Collapsing all ten dimensions into a single omnibus score produces poor reliability, indicating that a one-trait summary is not defensible in this corpus. Under Variant B, Openness remains high, but Self-Enhancement drops to a moderate level, reinforcing Variant A for substantive analyses. Party-stratified estimates are broadly comparable for Democrats and Republicans, and the omnibus 10D total remains unreliable in both groups.

Table 7: Reliability of four composites — Variant A (Hedonism → Openness)

Composite	**k**	α	**95% CI (α)**	ω	**SB**
Conservation	3	0.705	[0.695, 0.715]	0.595	0.677
Self-Transcendence	2	0.841	[0.834, 0.848]	0.667	0.842
Openness-to-Change (A)	3	0.885	[0.880, 0.888]	0.599	0.883
Self-Enhancement (A)	2	0.878	[0.873, 0.883]	0.667	0.913
Overall 10D total	10	0.295	[0.272, 0.319]	0.079	0.268

Table 8: Reliability of four composites — Variant B (Hedonism → Self-Enhancement)

Composite	**k**	α	**95% CI (α)**	ω	**SB**
Openness-to-Change (B)	2	0.927	[0.924, 0.930]	0.667	0.929
Self-Enhancement (B)	3	0.748	[0.740, 0.757]	0.591	0.733
Overall 10D total	10	0.295	[0.272, 0.319]	0.079	0.268

Table 9: Party-stratified α (Variant A)

Composite	**Democrats α**	**Republicans α**
Conservation	0.705	0.701
Self-Transcendence	0.790	0.863
Openness-to-Change (A)	0.879	0.891
Self-Enhancement (A)	0.882	0.872
Overall 10D total	0.317	0.302

4.4 Robustness of Reliability Estimates

Treating scores as ordinal produced Spearman-based α and ω that replicate the Pearson ordering, indicating that the interval treatment does not drive conclusions. Across 200 split-half draws per composite, Spearman–Brown medians and 95% split intervals remain within the same qualitative bands as the point estimates. Average inter-indicator correlations are moderate for 3-indicator composites and high for 2-indicator composites, consistent with the α pattern. Using ω as the reliability coefficient, standard errors of measurement yield narrow 95% score bands for Openness and Self-Enhancement. Leave-one-out α for 3-indicator composites shows no single indicator disproportionately inflates or depresses reliability. Bootstrap confidence intervals for Democrat–Republican differences in α span zero for all composites, corroborating subgroup stability.

Table 10: Robustness summary (Variant A) - Part 1: Split-half and AIC

Composite	**SB mean**	**SB median**	**SB 95% CI**	**AIC (Pearson)**	**AIC (Spearman)**
Conservation	0.678	0.647	[0.534, 0.875]	0.435	0.436
Self-Transcendence	0.842	0.842	[0.842, 0.842]	0.726	0.684
Openness-to-Change (A)	0.883	0.894	[0.814, 0.952]	0.730	0.726
Self-Enhancement (A)	0.913	0.913	[0.913, 0.913]	0.839	0.803
Overall 10D total	0.330	0.465	[-0.622, 0.851]	—	—

Table 11: Robustness summary (Variant A) - Part 2: Standard deviation and SEM

Composite	**Observed SD**	**SEM (ω)**	**$\pm$1.96·SEM**
Conservation	0.444	0.283	0.554
Self-Transcendence	0.570	0.329	0.645
Openness-to-Change (A)	0.487	0.308	0.604
Self-Enhancement (A)	0.445	0.257	0.503
Overall 10D total	0.203	0.194	0.381

4.5 Appendix — Full Robustness Tables

Table 12: Pearson vs Spearman α/ω (ordinal robustness)

Composite	**k**	α **Pearson**	ω **Pearson**	α **Spearman**	*omega* **Spearman**
Conservation	3	0.697	0.595	0.699	0.596
Self-Transcendence	2	0.842	0.667	0.813	0.667
Openness-to-Change (A)	3	0.89	0.599	0.888	0.599
Self-Enhancement (A)	2	0.913	0.667	0.891	0.667
Overall 10D total	10	0.335	0.079	0.416	0.095

Table 13: Split-half distribution (mean, median, 95% CI)

Composite	**k**	**SB mean**	**SB median**	**SB CI low**	**SB CI high**
Conservation	3	0.678	0.647	0.534	0.875
Self-Transcendence	2	0.842	0.842	0.842	0.842
Openness-to-Change (A)	3	0.883	0.894	0.814	0.952
Self-Enhancement (A)	2	0.913	0.913	0.913	0.913
Overall 10D total	10	0.33	0.465	-0.622	0.851

Table 14: Average inter-indicator correlations (Pearson & Spearman)

Composite	**k**	**AIC Pearson**	**AIC Spearman**
Conservation	3	0.435	0.436
Self-Transcendence	2	0.726	0.684
Openness-to-Change (A)	3	0.73	0.726
Self-Enhancement (A)	2	0.839	0.803
Overall 10D total	10	0.048	0.066

Table 15: Standard Error of Measurement (SEM) and 95% score bands (using ω)

Composite	**k**	**Observed SD**	**Reliability for SEM(ω)**	**SEM**	**$\pm$ 1.96 $\times$ SEM**
Conservation	3	0.444	0.595	0.283	0.554
Self-Transcendence	2	0.57	0.667	0.329	0.645
Openness-to-Change (A)	3	0.487	0.599	0.308	0.604
Self-Enhancement (A)	2	0.445	0.667	0.257	0.503
Overall 10D total	10	0.203	0.079	0.194	0.381

Table 16: Leave-one-out α for k=3 composites

Composite	**k**	**Base α**	**Dropped indicator**	**α after drop**	**$\Delta\alpha$**
Conservation	3	0.697	Security_Judge	0.685	-0.012
Conservation	3	0.697	Conformity_Judge	0.243	-0.454
Conservation	3	0.697	Tradition_Judge	0.784	0.086
Openness-to-Change (A)	3	0.89	Self-Direction_Judge	0.847	-0.043
Openness-to-Change (A)	3	0.89	Stimulation_Judge	0.74	-0.15
Openness-to-Change (A)	3	0.89	Hedonism_Judge	0.929	0.039
Overall 10D total	10	0.335	Security_Judge	0.53	0.195
Overall 10D total	10	0.335	Conformity_Judge	0.487	0.152
Overall 10D total	10	0.335	Tradition_Judge	0.326	-0.009
Overall 10D total	10	0.335	Benevolence_Judge	0.196	-0.139
Overall 10D total	10	0.335	Universalism_Judge	0.237	-0.098
Overall 10D total	10	0.335	Self-Direction_Judge	0.268	-0.067
Overall 10D total	10	0.335	Stimulation_Judge	0.193	-0.142
Overall 10D total	10	0.335	Hedonism_Judge	0.169	-0.166
Overall 10D total	10	0.335	Achievement_Judge	0.208	-0.127
Overall 10D total	10	0.335	Power_Judge	0.305	-0.03

Table 17: Democrat–Republican $\Delta\alpha$ (bootstrap CI)

Composite	**k**	**$\Delta\alpha$ (Dem-Rep)**	**95% CI**
Conservation	3	0.004	[-0.017, 0.025]
Self-Transcendence	2	-0.076	[-0.092, -0.060]
Openness-to-Change (A)	3	-0.013	[-0.022, -0.003]
Self-Enhancement (A)	2	0.007	[-0.001, 0.014]
Overall 10D total	10	0.045	[0.001, 0.085]

5 Validity

We assess validity on three fronts: (i) *internal structure* of the ten value dimensions (convergent/discriminant and circumplex expectations), (ii) *known–groups* differences (Democrats vs. Republicans), and (iii) a light *criterion* check (party classification). Throughout we benchmark against the Schwartz value model and its expected motivational oppositions and compatibilities (Schwartz, 1992; Schwartz and Boehnke, 2004; Schwartz et al., 2012; Sagiv and Schwartz, 2022).

5.1 Internal Structure

The 10×10 Pearson and Spearman correlation matrices display the expected circumplex pattern—positive correlations among adjacent values and negative correlations for opposing regions. We then derived two-dimensional layouts and quantified circular agreement with the canonical Schwartz order after optimal rotation using the mean absolute angular deviation (MAD°), the resultant length R, and a permutation test on MAD°.

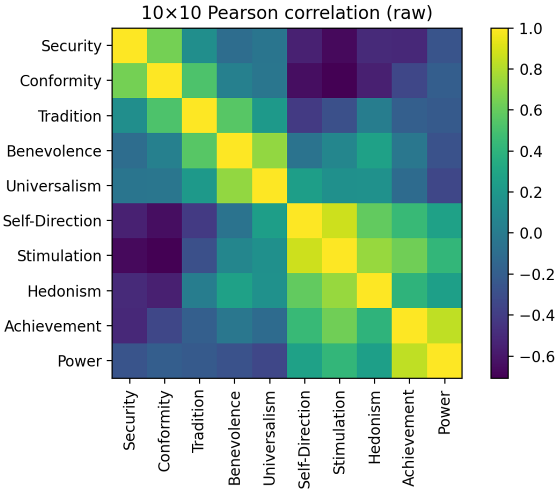

Figure 2: Pearson correlation heatmap for 10 values

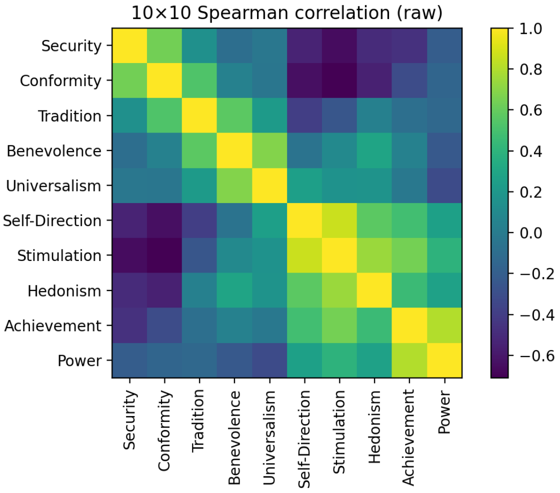

Figure 3: Spearman correlation heatmap for 10 values

Two alternative 2D layouts with matched axis limits for visual comparability.

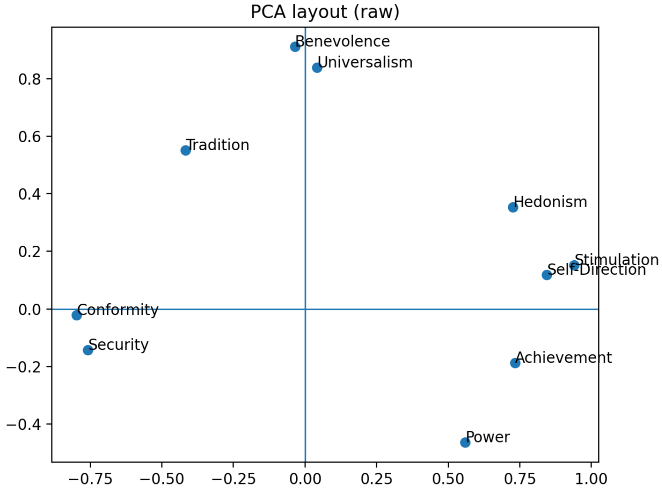

Figure 4: PCA-based 2D layout

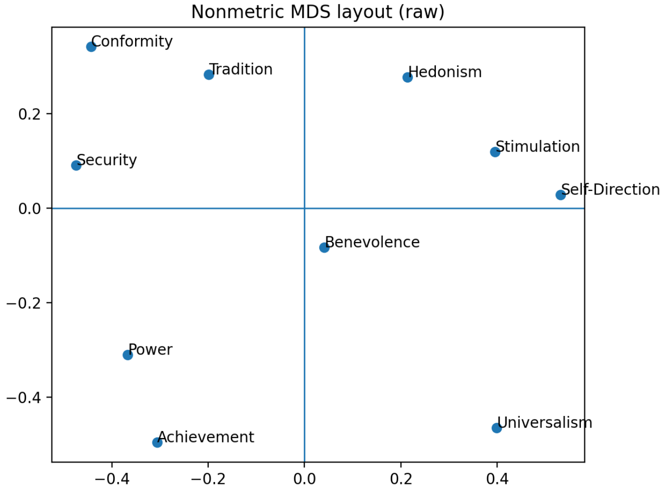

Figure 5: Nonmetric MDS-based 2D layout

As shown in Table 18, nonmetric MDS yields substantially stronger circular recovery (MAD° $= 46.2$, $R = 0.516$, $p = .005$) than PCA (MAD° $= 77.5$, $R = 0.205$, $p = .231$). Accordingly, we adopt the nonmetric MDS layout in the main text and report the PCA solution as a robustness check. The MDS configuration also aligns with the theorized quadrants (Openness-to-Change vs. Conservation; Self-Transcendence vs. Self-Enhancement), with the familiar proximity of *Self-Direction/Stimulation/Hedonism* and *Conformity/Tradition/Security*, and the expected opposition between *Universalism/Benevolence* and *Power/Achievement*.

5.2 Known-groups validity

We compare party means on the four composites (unadjusted). Directionality matches theory: Democrats score higher on Openness-to-Change and Self-Transcendence, Republicans on Conservation and Self-Enhancement. We report mean differences (Dem$-$Rep), Hedges' g with 95% CIs, and both Welch and Brunner–Munzel tests; positive g indicates Democrats $>$ Republicans.

Table 18: Circular agreement summary

Method	**MAD°**	R	**Permutation p (MAD°)**
PCA (on R)	77.50	0.205	0.231
Nonmetric MDS (on $1-R$)	46.24	0.516	0.005

Note: Configurations are circularly aligned to the canonical Schwartz order via circular Procrustes; lower MAD° and higher R indicate better circular recovery; permutation test uses $B = 10{,}000$ labelings.

Table 19: Axis-level evidence (Variant A; Fisher-z 95% CIs)

Pair	r	**95% CI**	N
Openness vs. Conservation	-0.701	[-0.712, -0.690]	8212
Self-Transcendence vs. Self-Enhancement	-0.186	[-0.207, -0.165]	8212
Openness vs. Self-Transcendence (cross-axis)	0.154	[0.133, 0.175]	8212
Conservation vs. Self-Enhancement (cross-axis)	-0.400	[-0.418, -0.381]	8212

Note: CIs computed via the Fisher z-transform (two-sided).

Table 20: Known-Groups Validity: Democrats vs. Republicans

Composite	**Group Means**			**Effect Size**		**p-values**		**N (D, R)**
	Dem	**Rep**		**Hedges' g**	**95% CI**	**Welch**	**BM**	
Conservation	3.139	3.234	-0.096	-0.216	$[-0.260, -0.173]$	$< .001$	$< .001$	4184, 4028
Self-Transcendence	4.166	3.878	$+0.288$	$+0.522$	$[+0.478, +0.566]$	$< .001$	$< .001$	4184, 4028
Openness-to-Change[a]	2.334	2.259	$+0.075$	$+0.154$	$[+0.111, +0.197]$	$< .001$	$< .001$	4184, 4028
Self-Enhancement[a]	2.254	2.363	-0.109	-0.248	$[-0.291, -0.204]$	$< .001$	$< .001$	4184, 4028

[a] Adjusted composite scores.

Note: Unadjusted group comparisons. Positive effect sizes indicate Democrats > Republicans. BM = Brunner–Munzel test for nonparametric comparison. All statistical tests are two-tailed with $\alpha = .05$.

5.3 Criterion indication

As a light predictive check, we fit 10-fold cross-validated logistic models using (a) the four composites and (b) the two axial contrasts. Mean AUCs are reported with across-fold SD; folds were stratified by party.

Table 21: Criterion indication: cross-validated party prediction

Model	**CV10 mean AUC**	**SD across folds**	**Folds**
Composites (4D)	0.683	0.025	10
Axis contrasts (2D)	0.672	0.028	10

6 Discussion

6.1 Methodological Contributions

- Advances over existing approaches
- Novel aspects of the pipeline
- Potential applications

6.2 Limitations and Future Directions

- Current constraints of the method
- Areas for improvement
- Future research opportunities

6.3 Implications for Social Computing

- Broader applications in computational social science
- Integration with other methodologies
- Scalability considerations

7 Conclusion

- Summary of key contributions
- Validation of the methodology
- Call for broader adoption and testing

8 Appendices

A. Complete prompt texts

B. Technical implementation details

C. Additional validation results

D. Comparison with alternative approaches

References

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., and Jaques, N. (2023). Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., and Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Beattie, P., Chen, R., and Bettache, K. (2022). When left is right and right is left: The psychological correlates of political ideology in china. *Political Psychology*, 43(3):457–488.

Ben-Ner, A. and Putterman, L., editors (1998). *Economics, values, and organization*. Cambridge University Press, Cambridge.

Berlin, I. (2013). *Against the Current: Essays in the History of Ideas*. Princeton University Press.

Blodgett, J. G., Bakir, A., and Rose, G. M. (2008). A test of the validity of Hofstede's cultural framework. *Journal of Consumer Marketing*, 25(6):339–349.

Bobbio, N. (1996). *Left and right: The significance of a political distinction*. University of Chicago Press, Chicago.

Boin, A. and Lodge, M. (2021). Responding to the COVID-19 crisis: A principled or pragmatist approach? *Journal of European Public Policy*, 28(8):1131–1152. Publisher: Taylor & Francis.

Boutyline, A. and Vaisey, S. (2017). Belief network analysis: A relational approach to understanding the structure of attitudes. *American Journal of Sociology*, 122(5):1371–1447.

Brandt, M. J. and Crawford, J. T. (2020). Worldview conflict and prejudice. In *Advances in Experimental Social Psychology*, volume 61, pages 1–66. Elsevier.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3:296–322.

Caprara, G. V., Schwartz, S., Capanna, C., Vecchione, M., and Barbaranelli, C. (2006). Personality and politics: Values, traits, and political choice. *Political Psychology*, 27(1):1–28. Place: United Kingdom Publisher: Blackwell Publishing.

Caprara, G. V. and Zimbardo, P. G. (2004). Personalizing politics: A congruency model of political preference. *American Psychologist*, 59(7):581–594.

Chae, Y. and Davidson, T. (2025). Large language models for text classification: from zero-shot learning to instruction-tuning. *Sociological Methods & Research*, page 00491241251325243.

Converse, P. (1964). The nature of belief systems in mass publics. *Ideology and discontent*, pages 75–169. Publisher: Free Press of Glencoe.

Costello, T. H., Zmigrod, L., and Tasimi, A. (2023). Thinking outside the ballot box. *Trends in Cognitive Sciences*, 27(7):605–615.

Crawford, J. T. and Pilanski, J. M. (2014). Political intolerance, right *and* left. *Political Psychology*, 35(6):841–851.

Croft, J. (2024). Using Structured Outputs in Azure OpenAI's GPT-4o for consistent document data processing. https://techcommunity.microsoft.com/blog/azureforisvandstartupstechnicalblog/using-structured-outputs-in-azure-openai Microsoft Azure for ISV and Startups Technical Blog.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Davidov, E., Schmidt, P., and Schwartz, S. H. (2008a). Bringing values back in: The adequacy of the european social survey to measure values in 20 countries. *Public Opinion Quarterly*, 72(3):420–445.

Davidov, E., Schmidt, P., and Schwartz, S. H. (2008b). Bringing values back in: The adequacy of the european social survey to measure values in 20 countries. *Public opinion quarterly*, 72(3):420–445.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dowling, J. and Pfeffer, J. (1975). Organizational Legitimacy: Social Values and Organizational Behavior. *The Pacific Sociological Review*, 18(1):122–136.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, New York.

Eisinga, R., te Grotenhuis, M., and Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Spearman–Brown, and Cronbach's alpha. *International Journal of Public Health*, 58(4):637–642.

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Feldman, S. (1988). Structure and consistency in public opinion: The role of core beliefs and values. *American Journal of Political Science*, 32(2):416.

Feldman, S. (2003). Values, ideology, and the structure of political attitudes. In *Oxford handbook of political psychology*, pages 477–508. Oxford University Press, New York, NY, US.

Feldman, S. and Johnston, C. (2014). Understanding the determinants of political ideology: Implications of structural complexity. *Political Psychology*, 35(3):337–358.

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for likert-type and ordinal item response data: A practical guide and tutorial. *Practical Assessment, Research & Evaluation*, 17(3):1–13.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., and Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior research methods*, 50(1):344–361.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Goren, P., Smith, B., and Motta, M. (2022). Human values and sophistication interaction theory. *Political Behavior*, 44(1):49–73. ISBN: 0123456789 Publisher: Springer US.

Graham, J., Haidt, J., and Nosek, B. A. (2009a). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Graham, J., Haidt, J., and Nosek, B. A. (2009b). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Green, S. B. and Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74(1):155–167.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Grosfeld, E., Scheepers, D., and Cuyvers, A. (2022). Value alignment and public perceived legitimacy of the European Union and the court of justice. *Frontiers in Psychology*, 12:785892.

Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y., and Elyoseph, Z. (2024). Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using schwartz's theory of basic values. *JMIR Mental Health*, 11:e55988.

Heath, A., Fisher, S., and Smith, S. (2005). The globalization of public opinion research. *Annu. Rev. Polit. Sci.*, 8(1):297–333.

Hofstede, G. (1983). National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization*, 13(1-2):46–74.

Holloway, I. and Todres, L. (2003). The status of method: flexibility, consistency and coherence. *Qualitative research*, 3(3):345–357.

Iliev, R., Hoover, J., Dehghani, M., and Axelrod, R. (2016). Linguistic positivity in historical texts reflects dynamic environmental and psychological factors. *Proceedings of the National Academy of Sciences*, 113(49):E7871–E7879.

Inglehart, R. and Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, 65(1):19.

Islam, T. and Goldwasser, D. (2025). Can llms assist annotators in identifying morality frames?-case study on vaccination debate on social media. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 169–178.

Jost, J. T., Federico, C. M., and Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology*, 60(1):307–337.

Judd, N., Drinkard, D., Carbaugh, J., and Young, L. (2017). congressional-record: A parser for the Congressional Record. https://github.com/unitedstates/congressional-record. Software, BSD 3-Clause License. Accessed 2025-08-31.

Kaasa, A. (2021). Merging hofstede, schwartz, and inglehart into a single system. *Journal of Cross-Cultural Psychology*, (1):002202212110112–002202212110112.

Kalmoe, N. P. (2020). Uses and abuses of ideology in political psychology. *Political Psychology*, 41(4):771–793.

Kam, C. D. (2005). Who toes the party line? Cues, values, and individual differences. *Political Behavior*, 27(2):163–182.

Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., and Stein, B. (2022). Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Kooiman, J. and Jentoft, S. (2009). Meta-governance: Values, norms and principles, and the making of hard choices. *Public Administration*, 87(4):818–836.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., et al. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Leung, K. and Bond, M. H. (2004). Social axioms: A model for social beliefs in multicultural perspective. In *Advances in experimental social psychology, Vol. 36*, pages 119–197. Elsevier Academic Press, San Diego, CA, US.

Levy, M., Jacoby, A., and Goldberg, Y. (2024). Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Maio, G. R. (2010). Mental Representations of Social Values. In *Advances in Experimental Social Psychology*, volume 42, pages 1–43. Elsevier.

Malka, A., Lelkes, Y., and Soto, C. J. (2019). Are cultural and economic conservatism positively correlated? A large-scale cross-national test. *British Journal of Political Science*, 49(3):1045–1069.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment.* Lawrence Erlbaum Associates, Mahwah, NJ.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3):412–433.

Mendelsohn, J., Bras, R. L., Choi, Y., and Sap, M. (2023). From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. *arXiv preprint arXiv:2305.17174*.

Nelson, T. E. and Garst, J. (2005). Values-based political messages and persuasion: Relationships among speaker, recipient, and evoked values. *Political Psychology*, 26(4):489–516.

Noble, J. (2024). What is LLM Temperature? https://www.ibm.com/think/topics/llm-temperature. IBM Think—Artificial Intelligence, Compute and servers, IT automation.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460.

OpenAI (2024). GPT-4o mini: advancing cost-efficient intelligence. Published July 18, 2024. Accessed 2025-08-31.

OpenAI (2024). Gpt-4o via openai api. Accessed: 2025-04-26.

OpenAI (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Milestone announcement dated May 13, 2024.

Park, J., Liscio, E., and Murukannaiah, P. K. (2024). Morality is non-binary: Building a pluralist moral sentence embedding space using contrastive learning. *arXiv preprint arXiv:2401.17228*.

Pasek, M. H., Ankori-Karlinsky, L.-O., Levy-Vene, A., and Moore-Berg, S. L. (2022). Misperceptions about out-partisans' democratic values may erode democracy. *Scientific Reports*, 12(1):16284.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015.

Piurko, Y., Schwartz, S. H., and Davidov, E. (2011). Basic personal values and the meaning of left-right political orientations in 20 countries. *Political Psychology*, 32(4):537–561.

Rathbun, B. C., Kertzer, J. D., Reifler, J., Goren, P., and Scotto, T. J. (2016). Taking Foreign Policy Personally: Personal Values and Foreign Policy Attitudes. *International Studies Quarterly*, 60(1):124–137.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21(2):173–184.

Revelle, W. and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma (2009). *Psychometrika*, 74(1):145–154.

Rohan, M. J. (2000). A rose by any name? The values construct. *Personality and Social Psychology Review*, 4(3):255–277.

Rokeach, M. (1973). *The nature of human values.* The nature of human values. Free Press, New York, NY, US. Pages: x, 438.

Ryan, A. (2024). GPT-4o: The Next Leap by OpenAI. https://quantilus.com/article/gpt-4o-the-next-leap-by-openai/. Blog post; accessed on October 28, 2025.

Sagi, E. and Dehghani, M. (2014). Measuring moral rhetoric in text. *Social science computer review*, 32(2):132–144.

Sagiv, L. and Roccas, S. (2021). How Do Values Affect Behavior? Let Me Count the Ways. *Personality and Social Psychology Review*, 25(4):295–316.

Sagiv, L. and Schwartz, S. H. (2022). Personal values across cultures. *Annual Review of Psychology*, 73(1):517–546.

Schuman, H. and Presser, S. (1977). Question wording as an independent variable in survey analysis. *Sociological Methods & Research*, 6(2):151–170.

Schwartz, S. H. (1992). Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Elsevier.

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire package of the european social survey*, 259(290):261.

Schwartz, S. H. and Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53(3):550–562.

Schwartz, S. H. and Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality*, 38(3):230–255.

Schwartz, S. H., Caprara, G. V., and Vecchione, M. (2010). Basic personal values, core political values, and voting: a longitudinal analysis. *Political Psychology*, 31(3):421–452.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J. E., Demirutku, K., Dirilen-Gumus, O., and Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., and Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, 32(5):519–542.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1):107–120.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3:271–295.

Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Tetlock, P. E. (1983). Cognitive style and political ideology. *Journal of Personality and social Psychology*, 45(1):118.

Tetlock, P. E., Hannum, K. A., and Micheletti, P. M. (1984). Stability and change in the complexity of senatorial debate: Testing the cognitive versus rhetorical style hypotheses. *Journal of Personality and Social Psychology*, 46(5):979.

Tong, X., Choenni, R., Lewis, M., and Shutova, E. (2024). Metaphor understanding challenge dataset for llms. *arXiv preprint arXiv:2403.11810*.

unitedstates (2025). Congress-legislators: Members of the united states congress, 1789–present. CC0 1.0 Universal (public domain).

Van de Vijver, F. J. and Leung, K. (2021). *Methods and data analysis for cross-cultural research*, volume 116. Cambridge University Press.

Van Dijk, T. A. (1997). What is political discourse analysis? *Belgian Journal of Linguistics*, 11:11–52.

van Stekelenburg, J. and Klandermans, B. (2017). Individuals in movements: A social psychology of contention. In Roggeband, C. and Klandermans, B., editors, *Handbook of Social Movements Across Disciplines*, pages 103–139. Springer International Publishing, Cham.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Weber, M. (2019). *Economy and society: A new translation*. Harvard University Press, Cambridge, Massachusetts.

Wetherell, G. A., Brandt, M. J., and Reyna, C. (2013). Discrimination across the ideological divide: The role of value violations and abstract values in discrimination by liberals and conservatives. *Social Psychological and Personality Science*, 4(6):658–667.

Yao, J., Yi, X., and Xie, X. (2024). Clave: An adaptive framework for evaluating values of llm generated responses. *Advances in Neural Information Processing Systems*, 37:58868–58900.

Young, K. (1977). 'values' in the policy process. *Policy & Politics*, 5(3):1–22.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Zhu, W., Xie, Y., Song, G., and Zhang, X. (2025). Eavit: Efficient and accurate human value identification from text data via llms. *arXiv preprint arXiv:2505.12792*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods*, 6(1):21–29.

A Maximum Token Threshold Determination