Building an Archetype: AI and Roll-Call Data [†]

Rómulo A. Chumacero [‡]

Abstract

This paper studies whether ideological benchmarks can be constructed from roll-call proposal content rather than inferred from voting patterns alone. Standard scaling methods, such as W-NOMINATE and IDEAL, estimate legislators' ideal points from the vote matrix without using proposal content. The approach proposed here reverses that information structure: human coding and large language models (LLMs) infer how stylized ideological archetypes—communist, socialist, social democrat, liberal, conservative, libertarian, left-wing, and right-wing—would have voted on each proposal. Legislators are then assigned affinity scores with each archetype, defined as agreement rates over valid comparisons.

The application uses 4,609 roll calls from the Chilean Constitutional Convention of 2021–2022, benchmarked against a W-NOMINATE–IDEAL consensus, with Optimal Classification as an additional diagnostic. Archetype affinities reproduce the dominant left-right ordering, but benchmark agreement alone is not discriminating. The key condition is vote informativeness: proposals on which archetypes vote alike carry little information for distinguishing ideological profiles. Once restricted to archetype-informative votes, all LLMs recover the expected ideological signs, and affinities also separate observed political coalitions. Archetype affinities therefore complement conventional scaling by providing explicit, content-anchored benchmarks for interpreting latent ideological structure.

Keywords: Archetype, Roll call data, W-NOMINATE, IDEAL, Optimal Classification, Artificial Intelligence, Chile.

JEL Classification: C25, C80, D72.

This version: July 2026

[†] I would like to thank Jorge Abela, María José Abud, Rodrigo Cases, Alejandro D'Andrea, Rodrigo Fuentes, Sebastián Izquierdo, Carmen Le Foulon, Aldo Mascareño, Leónidas Montes, Ricardo Paredes, Klaus Schmidt-Hebbel, and Paola Soux for helpful comments and suggestions. Jaime Aránguiz and Arturo Ramirez provided able research assistance. The usual disclaimer applies.

[‡] Department of Economics, University of Chile. E-mail: rchumace@fen.uchile.cl.

1 Introduction

Roll-call votes are one of the most systematic records of political behavior. They register repeated choices made by legislators when facing alternative policy outcomes and, for that reason, have become a central empirical input for measuring ideology, polarization, party cohesion, and legislative conflict. The methodological tradition behind this use of roll-call data is grounded in spatial models of voting. In these models, legislators are represented by ideal points in a low-dimensional policy space, while each roll call presents a choice between alternatives located in that same space. Voting behavior is therefore interpreted as the observed implication of latent preferences, subject to error, strategic considerations, or institutional constraints (Davis et al., 1970; Enelow and Hinich, 1984; Poole and Rosenthal, 1985; Poole, 2005).

The dominant empirical implementation of this logic is the NOMINATE family of models. Poole and Rosenthal (1985, 1991, 1997, 2001) developed a sequence of spatial scaling procedures that recover legislator locations and roll-call parameters from voting matrices. W-NOMINATE and DW-NOMINATE became standard tools because they reduce high-dimensional roll-call records to low-dimensional coordinates that are interpretable, comparable across settings, and useful in subsequent empirical work. The same tradition includes Optimal Classification (Poole, 2000), the linear probability approach of Heckman and Snyder (1997), and software implementations that made these methods more accessible to applied researchers (Poole et al., 2011).

The state of the art is broader than NOMINATE. Bayesian ideal-point models place roll-call analysis within an item-response framework and provide a coherent way to quantify uncertainty, compare behavioral assumptions, and incorporate additional information about legislators, parties, or the agenda (Jackman, 2001; Martin and Quinn, 2002; Clinton et al., 2004). Subsequent work has compared NOMINATE and Bayesian approaches, studied uncertainty, relaxed functional-form assumptions, and explored multidimensional or dynamic extensions (Carroll et al., 2009; Clinton and Jackman, 2009; Imai et al., 2016; Binding and Stoetzer, 2023). Alternative

dimensionality-reduction approaches, including principal components, have also been proposed as simpler or more transparent ways to recover ideal points under some conditions (Potthoff, 2018).

A key advantage of these models is that they do not need to know what the votes are about. NOMINATE-type and ideal-point methods operate on the roll-call matrix itself. They infer ideological structure from patterns of agreement and disagreement across legislators. This is a powerful feature because it reduces the need for external interpretation, avoids hand-coding the substantive content of each proposal, and allows researchers to scale large voting matrices even when individual votes are not analyzed one by one.

That advantage also defines the main contrast with the approach considered in this paper. Standard scaling methods estimate latent positions from observed voting patterns, but the substantive interpretation of the recovered dimensions is largely assigned ex post. They indicate where legislators are located relative to one another, but they do not directly answer a different question: how close is a legislator's voting record to the behavior expected from a clearly defined ideological type? This question connects roll-call scaling with an older tradition in the measurement of ideology. Before and alongside NOMINATE-type methods, researchers and interest groups used explicit ideological ratings, such as ADA liberalism scores, to summarize legislative behavior. These indices were easy to interpret but depended on the selection of votes and on the judgment of the organization constructing the benchmark. Shaffer (1989), for example, evaluates whether ADA-selected roll calls form a reliable and valid measure of liberalism.

Large language models make this approach newly feasible at scale. Recent work shows that LLMs can perform political text annotation and classification tasks, sometimes approaching or exceeding crowdworker performance, while also stressing that accuracy depends on the task, prompt, language, text type, and model used (Gilardi et al., 2023; Heseltine and Clemm von Hohenberg, 2024; Ornstein et al., 2025). A related literature uses text-as-data methods and LLMs to estimate or

simulate political positions from manifestos, legislative speech, social media, campaign contributions, or political texts (Laver et al., 2003; Slapin and Proksch, 2008; Bonica, 2014; Barberá, 2015; Lauderdale and Herzog, 2016; Gentzkow et al., 2019; Le Mens and Gallego, 2025; Di Leo et al., 2025; Kreutner et al., 2026). Burnham (2024), for example, proposes Semantic Scaling, which uses LLMs to classify documents into survey-like responses and then estimates ideal points through item-response models.

This paper differs from those contributions in the object being validated. It does not ask whether an LLM can recover the position of an observed political actor or simulate the vote of a real legislator. It asks whether a profile constructed from the definition of an externally specified ideological type, and never fitted to any actor, is correctly oriented and coherent enough to recover real political structure when compared with actual votes. The information set is therefore different from that used by conventional roll-call scaling. W-NOMINATE, IDEAL, and OC use the observed vote matrix but not proposal content. Archetype-based benchmarking uses proposal content to construct ideological reference profiles before comparing them with actual votes. The archetypal profiles are generated without observing any constituent's votes, identities, or bloc membership.

The central empirical question is not whether archetype affinities can be highly correlated with an estimated ideal-point dimension. High correlations with the dominant dimension are not, by themselves, a discriminating validation test: as shown below, the observed voting records of the most extreme members on each side, being the most informative reference points for the ordering, also recover it very strongly when used as content-blind profiles. The relevant question is whether profiles constructed from proposal content alone are correctly oriented, internally coherent, and informative about ideological differences. This distinction is important because some votes may be politically informative but textually opaque. A vote that refers only to a numbered amendment may separate legislators sharply, but it cannot be classified by an archetype unless information external to the minutes is supplied. Conversely, a vote may be substantively classifiable from the text and still be

uninformative for ideological discrimination if all archetypes are predicted to vote the same way.

The paper makes three contributions. First, it proposes an archetype-based benchmarking procedure for roll-call analysis. The method constructs ideological reference profiles from proposal content and measures legislators' affinities to those profiles. Second, it shows that the relevant filtering condition is not simply whether a vote is substantive, but whether it distinguishes archetypes. Substantive votes are not necessarily informative votes. Third, it validates the resulting affinities against observable political structure: archetypal profiles constructed without observing constituent votes or bloc membership generate affinity measures that separate actual political coalitions in the Convention. The empirical evaluation also uses content-blind extreme-member profiles as a null benchmark for raw benchmark recovery and cross-model bootstrap comparisons as diagnostics.

The empirical application uses the Chilean Constitutional Convention of 2021–2022. The case is useful because it generated a dense roll-call record on a heterogeneous constitutional agenda. The analysis uses 4,609 roll calls and compares archetype affinities with a consensus benchmark constructed from W-NOMINATE and IDEAL, with Optimal Classification used as an additional diagnostic. The results show that, in the unfiltered baseline, left-leaning archetypes are recovered consistently across LLMs, whereas right-leaning archetypes are more model-dependent. This asymmetry is consistent with documented ideological priors in conversational models (Rozado, 2024). Once the comparison is restricted to votes on which archetypes are predicted to disagree, all LLMs recover the expected ideological signs. Remaining cross-model differences are small, although one model remains less sharp in recovering the right-leaning profile. The results also show that archetype affinities are consistent with observed political blocs in the Convention.

The rest of the paper evaluates these claims through four criteria: side placement, coherence, vote informativeness, and consistency with observable political structure. Section 2 presents the archetype-based benchmarking method and defines the affinity

and informativeness measures. Section 3 describes the Chilean Constitutional Convention and the roll-call data. Section 4 explains the empirical implementation, including the statistical benchmarks, human and LLM coding, vote sets, the content-blind extreme-member null benchmark, and the bootstrap procedure. Section 5 presents the results, including exploratory multidimensional and domain-specific evidence. Section 6 concludes.

2 Archetype-Based Benchmarking of Roll-Call Votes

The objective of this section is to define an archetype-based benchmark for roll-call analysis. The benchmark is constructed from the substantive content of proposals and from prior descriptions of ideological types, and is then compared with observed voting behavior through agreement rates. This reverses the information structure of standard ideal-point estimators: conventional scaling recovers latent positions from the vote matrix, whereas archetype-based benchmarking first constructs hypothetical voting profiles for stylized ideological types and only then compares them with legislators' votes.

2.1 Information sets and objects of comparison

Let V denote the $N \times J$ roll-call matrix, with element v_{ij} corresponding to the vote cast by legislator i=1,...,N on proposal j=1,...,J. The binary coding is $v_{ij} = 1$ for support and $v_{ij} = 0$ for opposition. Missing values correspond to abstentions, absences, or cases in which no valid binary vote is observed. The empirical treatment of these cases is described in Section 3.2.

Standard roll-call scaling methods use V to estimate latent positions. Let $\hat{\theta}_i$ denote the ideal point, score, or coordinate assigned to legislator i by a statistical benchmark such as W-NOMINATE, IDEAL, or Optimal Classification. These methods infer ideological structure from patterns of agreement and disagreement among legislators.

They do not require the researcher to classify the substantive content of each proposal.

Archetype-based benchmarking uses a different information set. For each proposal, the researcher observes a textual description of the matter being voted on and constructs a hypothetical vote for each ideological archetype. Let $a \in \mathcal{A}$ index archetypes and $m \in \mathcal{M}$ index the source that constructs the profile, such as human experts or an LLM. For each proposal j, source m assigns an archetypal vote b_{aj}^{m}. The sequence $B_{a}^{m} = (b_{a1}^{m},\ldots,b_{aJ}^{m})$ is the voting profile of archetype a according to source m.

The key restriction is that B_{a}^{m} is constructed without using legislators' votes, identities, bloc membership, or estimated benchmark positions. The archetypal profile is therefore not fitted to the vote matrix. It is an externally defined reference profile derived from proposal content and from the description of the ideological type.

The two objects of comparison are thus distinct. Statistical scaling produces coordinates $\hat{\theta}_{i}$ from the vote matrix. Archetype-based benchmarking produces profiles B_{a}^{m} from proposal content and then measures how closely each observed legislator resembles each profile. The purpose is not to re-estimate ideal points, but to provide an interpretable reference benchmark for observed voting behavior.

2.2 Archetypal voting profiles

An archetype is a stylized ideological type with a prior substantive definition. It is not an observed legislator, party, faction, or bloc. It is a reference voter whose expected position on each proposal is inferred from the proposal's content and from the ideological description supplied to the source constructing the profile.

Let x_{j} denote the textual information available for proposal j, and let z_{a} denote the description of archetype a. Source m maps $\left(x_{j}, z_{a}\right)$ into an archetypal vote,

$$b_{aj}^{m} = f_{m}(x_{j}, z_{a}),$$

where $b_{aj}^{m} \in \{0, 1, \varnothing\}$.

A value of 1 means that source m predicts archetype a would support the proposal. A value of 0 means that the archetype would oppose it. The value $\varnothing$ means that no defined archetypal vote is assigned. This may occur because the proposal is procedural, textually opaque, insufficiently substantive, or otherwise not classifiable from the available information.

The value $\varnothing$ is not an abstention by the archetype. It records a limitation of the information available to construct the profile. A proposal that refers only to a numbered amendment, for example, may separate legislators sharply in the observed vote matrix but still provide too little substantive information for a content-based classifier to assign an archetypal vote. Such a vote can be useful for statistical scaling but should not enter an archetype-based comparison unless the substantive content needed to define the archetypal position is available.

Conversely, a proposal may be textually clear and still add little to ideological discrimination. If all archetypes are predicted to vote the same way, the proposal helps describe a common position but does not distinguish ideological profiles. The distinction between textual classifiability and ideological informativeness is therefore central to the method and is formalized below.

2.3 Affinity measures

The basic measure of resemblance between a legislator and an archetype is voting affinity. For legislator i, archetype a, and source m, define the valid comparison set as:

$$J_{ia}^{m} = j : v_{ij} \in 0, 1, b_{aj}^{m} \in 0, 1.$$

This is the set of proposals on which both the legislator and the archetype have defined binary votes. For each $j \in J_{ia}^{m}$, define the agreement indicator

$$s_{iaj}^{m} = \mathbf{1}(v_{ij} = b_{aj}^{m}).$$

The affinity between legislator i and archetype a, according to source m, is

$$S_{ia}^{m} = \frac{1}{\mid J_{ia}^{m} \mid} \sum_{j \in J_{ia}^{m}} s_{iaj}^{m},$$

whenever $\left|J_{ia}^{m}\right| > 0$. Thus, S_{ia}^{m} is the fraction of valid comparisons on which legislator i and archetype a vote the same way. Higher values indicate greater resemblance between the legislator's observed voting record and the archetype's hypothetical voting profile.

The agreement logic is familiar. Interest-group ratings have long compared legislators' roll-call behavior with positions selected by organizations that define liberal, conservative, environmental, labor, or business-oriented standards (Shaffer, 1989). Party-loyalty measures similarly summarize the share of votes cast in agreement with party leadership (Lee et al., 2004). The contribution here is not the agreement rate itself, but the construction of the reference profile. Instead of using positions selected ex post by an organization or by party leaders, the benchmark profile is generated from proposal content for an explicitly defined ideological archetype and then compared with observed voting records.

Although simple, the agreement rate is useful because it is bounded, transparent, and directly interpretable: $S_{ia}^{m} = 0.70$, for example, means that legislator i agrees with archetype a in 70 percent of valid comparisons according to source m. This makes affinities comparable across legislators for a given archetype and source. More elaborate similarity measures could be used, including weighted agreement, cosine similarity, or entropy-weighted scores, but the baseline measure keeps the interpretation of the benchmark explicit. The empirical question is then whether these transparent agreement rates are correctly oriented, coherent, and informative relative to external benchmarks.

2.4 Substantive domains and domain-specific affinity

Proposal content can be used not only to assign archetypal votes, but also to classify proposals by substantive domain. Let q_j^m denote the domain assigned to proposal j by source m. The set of domains is denoted by $\mathcal{D}$, with $q_j^m = d$ for $d \in \mathcal{D}$. A residual category, denoted by $q_j^m = 0$, includes proposals that are procedural, organizational, textually opaque, or insufficiently substantive for domain classification.

For any substantive domain d, define the domain-specific comparison set as

$$J_{iad}^m = j \in J_{ia}^m : q_j^m = d.$$

The corresponding domain-specific affinity is

$$S_{iad}^m = \frac{1}{\mid J_{iad}^m \mid} \sum_{j \in J_{iad}^m} s_{iaj}^m,$$

whenever $\left|J_{iad}^m\right| > 0$. Thus, S_{iad}^m is the agreement rate between legislator i and archetype a, according to source m, restricted to proposals in domain d.

Domain-specific affinities are content-based decompositions of the baseline affinity score. They allow the researcher to ask whether resemblance to an archetype is concentrated in particular issue areas, such as economic, social-cultural, or legal proposals. They should not be interpreted, by themselves, as estimates of separate ideological dimensions. A domain may be substantively distinct while still aligned with the same dominant voting cleavage. For this reason, domain-specific affinities are useful descriptively and diagnostically, but evidence of multidimensionality requires additional analysis.

2.5 Vote informativeness

A proposal can be substantive without being informative for distinguishing archetypes. The method therefore separates three concepts. A vote is **textually classifiable** when the proposal contains enough information for source m to assign a defined archetypal vote. A vote is **politically informative** but textually opaque when it separates legislators in the observed vote matrix but cannot be classified from the

available proposal text. A vote is **archetype-informative** when the archetypes are predicted to disagree on it.

For source m, let

$$A_j^m = a \in \mathcal{A} : b_{aj}^m \in 0,1$$

be the set of archetypes with defined binary votes on proposal j, and let $n_j^m = \left|A_j^m\right|$. Define

$$y_j^m = \sum_{a \in A_j^m} b_{aj}^m$$

as the number of archetypes predicted to support the proposal. A simple measure of archetypal disagreement is

$$D_j^m = \min\left\{y_j^m, n_j^m - y_j^m\right\}.$$

This statistic counts the size of the smaller side in the archetypal division. If $D_j^m = 0$, all defined archetypes are predicted to vote the same way. The proposal may still be substantively meaningful, but it does not distinguish ideological profiles. If $D_j^m > 0$, at least one archetype is predicted to disagree with the rest.

The informativeness rule can be made stricter by requiring

$$D_j^m \geq \tau,$$

where τ is a pre-specified threshold. This avoids treating a proposal as informative when disagreement is driven by a single dissenting archetype. The empirical implementation specifies the threshold used in the baseline analysis and reports robustness checks when appropriate.

Let $I_j^m(\tau) = 1$ if $D_j^m \geq \tau$, and $I_j^m(\tau) = 0$ otherwise. The informative comparison set for legislator i, archetype a, and source m is

$$J_{ia}^m(\tau) = j \in J_{ia}^m : I_j^m(\tau) = 1.$$

The corresponding informative-vote affinity is

$$S_{ia}^{m}(\tau) = \frac{1}{\mid J_{ia}^{m}(\tau) \mid} \sum_{j \in J_{ia}^{m}(\tau)} s_{iaj}^{m},$$

whenever $\left|J_{ia}^{m}(\tau)\right| > 0$.

Votes on which all, or almost all, archetypes are predicted to vote alike can be important for the agenda, but they carry little information for distinguishing ideological profiles. Restricting the comparison to archetype-informative votes therefore focuses the benchmark on proposals for which ideological types imply different voting predictions.

2.6 Empirical evaluation logic

Archetype affinities are functions of two objects: the observed votes of legislators and the archetypal profiles constructed from proposal content. They are therefore not independent of the vote matrix. The relevant independence claim is narrower: the archetypal profiles are not constructed using legislators' votes, identities, bloc membership, or benchmark positions. The empirical question is whether these externally defined profiles generate meaningful and interpretable orderings once they are compared with observed votes.

Correlation with a latent benchmark is a useful orientation check, but it is not sufficient. A coherent voting profile located at an ideological extreme can also recover the dominant left-right ordering. The content-blind extreme-member profiles used below are designed to show how much benchmark recovery can be obtained from observed extreme voting records alone. They require the vote matrix to define the profiles and the benchmark to identify the extreme members. Archetypal profiles require neither. Their role is therefore not to validate the archetype method, but to evaluate whether raw affinity–benchmark correlation is a discriminating criterion.

The empirical evaluation relies on four criteria. First, side placement: affinities to left- and right-leaning archetypes should have the expected signs relative to the benchmark. Second, coherence: ideologically related archetypes and independently

constructed sources should produce similar orderings when they are intended to represent similar ideological profiles. Third, vote informativeness: recovery should improve when the comparison is restricted to proposals on which archetypes disagree. Fourth, external political structure: affinity scores should separate observable coalitions, parties, or blocs that were not used to construct the archetypal profiles.

These criteria do not carry equal evidentiary weight. Side placement and vote informativeness are necessary for the affinity scores to be meaningful. Coherence evaluates whether related archetypes and sources produce interpretable patterns rather than arbitrary rankings. External political structure provides the strongest validation check because it does not depend on the latent axis estimated from the vote matrix. Together, these criteria separate benchmark alignment from substantive validation.

3 Empirical Setting and Roll-Call Data

The empirical application is the Chilean Constitutional Convention of 2021–2022 (CC). The case is useful for two reasons. First, the CC produced a large and unusually rich set of recorded votes over a short period. Second, the issues voted on were substantively heterogeneous, covering economic, social, legal, institutional, environmental, indigenous, procedural, and symbolic matters. This makes the case appropriate for evaluating a method that relies not only on the vote matrix, but also on the semantic content of the proposal attached to each vote.

This section describes the institutional setting and the roll-call data. The construction of statistical benchmarks, archetypal profiles, substantive domains, and informativeness filters is described in Section 4.

3.1 Institutional setting

The Chilean Constitutional Convention was part of the institutional process that followed the social unrest of October 2019 and the Agreement for Social Peace and a

New Constitution of November 2019. In the October 2020 plebiscite, voters approved the drafting of a new constitution and selected a fully elected constitutional convention as the body in charge of preparing the draft (Alemán and Navia, 2023; Larraín et al., 2023).

The Convention was elected in 2021 and initially consisted of 155 members. Its electoral rules included gender parity, reserved seats for indigenous peoples, and the participation of independent lists. Constitutional norms required approval by a two-thirds threshold, and the final proposal was submitted to a national ratification referendum, where it was rejected in September 2022.

The CC operated under a broad constitutional mandate. Its work included proposals on political institutions, social rights, economic regulation, environmental protection, indigenous recognition, territorial organization, legal design, and procedural rules. This breadth is relevant for the roll-call data analyzed below because it produced votes that differ substantially in both substantive domain and textual transparency.

3.2 Roll-call data and coding

The empirical analysis uses the roll-call record of the CC's plenary sessions. The data contain J=4,609 recorded votes. The unit of observation is the vote cast by member i on proposal j, corresponding to the entries of the roll-call matrix V defined in Section 2.

Votes are coded as binary choices. A vote in favor of a proposal is coded as $v_{ij} = 1$, and a vote against the proposal is coded as $v_{ij} = 0$. Abstentions, absences, and other cases in which no valid binary vote is observed are treated as missing values. This coding keeps the roll-call matrix consistent with the statistical benchmarks used below, which are estimated from observed support–opposition choices. It also preserves the distinction between an explicit negative vote and the absence of a valid binary vote, although alternative treatments of abstention may be relevant in settings where abstention has strategic implications.

Although the Convention was elected with 155 members, one member ceased to participate during the process. The empirical analysis therefore uses N=154 members, corresponding to the final active composition of 77 women and 77 men. The resulting data matrix contains 154 members and 4,609 roll calls, with missing entries whenever a member did not cast a valid binary vote.

The roll-call descriptions attached to each proposal provide the textual information used later to construct archetypal voting profiles and domain classifications. These descriptions vary substantially in substantive transparency. Some identify a policy issue directly, while others refer to amendments, procedural motions, or technical formulations whose substantive content is not fully recoverable from the roll-call label alone. This variation is central for distinguishing the observed vote matrix from the subset of votes that can be used for archetype-based benchmarking.

The next section specifies how these data are used to construct the statistical benchmarks, archetypal profiles, domain classifications, and informativeness filters.

4 Empirical Implementation

This section describes how the methodological objects defined above are implemented in the Chilean Constitutional Convention data. The objective is to make explicit the information used at each stage: the statistical benchmarks are estimated from the roll-call matrix, while the archetypal profiles are constructed from proposal content without using members' votes, identities, bloc membership, or benchmark positions.

The implementation proceeds in five steps. First, conventional roll-call benchmarks are estimated using W-NOMINATE, IDEAL, and Optimal Classification, and a consensus benchmark is constructed from W-NOMINATE and IDEAL. Second, human experts and LLMs generate archetypal voting profiles for a common set of ideological types. Third, proposals are classified by substantive domain and by whether they produce disagreement among archetypes. Fourth, the analysis constructs three vote sets: all

valid votes, source-defined votes, and archetype-informative votes. Fifth, the empirical evaluation uses benchmark correlations, content-blind extreme-member profiles, cross-model bootstrap comparisons, and external validation using political blocs.

4.1 Statistical benchmarks

The statistical benchmarks are estimated from the binary roll-call matrix described in Section 3. Three procedures are used: W-NOMINATE, IDEAL, and Optimal Classification (OC). They differ in their assumptions, objective functions, and treatment of uncertainty, but share the same purpose: to recover latent ideological structure from patterns of agreement and disagreement in roll-call votes.

NOMINATE-type models are spatial voting models. Legislators are represented by ideal points, roll calls are represented by parameters describing the alternatives being voted on, and observed votes are modeled as choices generated by the distance between legislators and alternatives, with probabilistic error. The original NOMINATE framework was introduced by Poole and Rosenthal (1985) and later developed into several variants, including D-NOMINATE, DW-NOMINATE, and W-NOMINATE (Poole and Rosenthal, 1991, 1997, 2001; Poole, 2005). W-NOMINATE is used here as the main spatial-scaling benchmark. Poole et al. (2011) describe its R implementation, and Carroll et al. (2009) discuss its relationship with Bayesian ideal-point methods.

The second benchmark is IDEAL, the Bayesian ideal-point estimator associated with Clinton et al. (2004). IDEAL formulates roll-call analysis as an item-response model in which legislator ideal points and vote parameters are estimated jointly from observed votes. This framework allows uncertainty quantification, model checking, and extensions to richer behavioral specifications. Carroll et al. (2009) show that NOMINATE and IDEAL often produce similar one-dimensional orderings in large voting bodies, although differences may arise from functional form, identification, and estimation choices.

The third benchmark is OC. Unlike likelihood-based or Bayesian procedures, OC estimates legislator locations by maximizing the number of correctly classified voting decisions. It is therefore a nonparametric alternative to NOMINATE-type models (Poole, 2000). Poole and Rosenthal (2001) show that OC produces configurations similar to DW-NOMINATE in congressional data, with high rank correlations between the methods.

All three procedures summarize the same object: the binary roll-call matrix. The main benchmark used below is a consensus configuration constructed from W-NOMINATE and IDEAL and discussed in Section 5.1. Optimal Classification is held out as an additional diagnostic rather than included in the consensus. Multidimensional estimates are reported separately as exploratory diagnostics because axis-specific comparisons in higher dimensions depend on alignment and rotation choices.

4.2 Sources, archetypes, and prompts

Archetypal voting profiles are constructed from two sources: human coding and LLM coding. The human coding provides a transparent left–right benchmark. The LLM coding extends the exercise to a larger set of ideological archetypes and allows cross-model comparison.

The human expert retained only proposals on which broad Left and Right archetypes were expected to vote differently. On this set, the two profiles are complements, so recording the Right archetype's predicted vote fully determines the Left profile. Retained proposals were classified only as economic or social-cultural. Legal or institutional proposals could enter the human-coded set when their implications were primarily economic or social; otherwise, they were excluded. The resulting profile reflects expert judgment based on proposal content, not observed votes or Convention member identities.[1]

[1] Jorge Abela and Paola Soux provided valuable help in reviewing the CC proposals, classifying their substantive content, and defining the benchmark voting profiles used in the human-coded exercise.

The LLM analysis uses five models: ChatGPT, Claude, Copilot, Gemini, and Grok. Each model constructs voting profiles for eight archetypes: Communist, Socialist, Social Democrat, Liberal, Conservative, Libertarian, Left, and Right. The first six represent more specific ideological traditions, while Left and Right provide broader reference profiles. The archetypes are not observed members, parties, or blocs of the Convention; they are externally defined ideological voters.

For each proposal–archetype pair, the model receives the proposal text and the archetype definition. The output is support, opposition, or no defined vote. The last category is used when the proposal is procedural, textually opaque, insufficiently substantive, or otherwise not classifiable from the information provided. In addition, each LLM classifies proposals into economic, social-cultural, legal, or residual domains. The residual category includes procedural, organizational, textually opaque, or insufficiently substantive proposals.

All sources follow the same information restriction. They are not given members' observed votes, identities, party or bloc membership, or benchmark coordinates. The profiles are constructed from proposal content and archetype definitions, and are only later compared with the roll-call matrix through the affinity measures defined in Section 2. The full prompts and implementation details are reported in Appendix A.[2]

4.3 Vote sets and informativeness rule

The empirical analysis distinguishes three vote sets. The first is the full set of valid binary roll calls in V. This is the information set used by W-NOMINATE, IDEAL, and Optimal Classification. It includes all observed support–opposition choices, with abstentions, absences, and other non-binary outcomes treated as missing at the member level.

[2] Recent work emphasizes that LLM-based annotation can be sensitive to prompt design, model version, task structure, and implementation details (Halterman, 2025 and Lin and Zhang, 2025). Appendix A.4 reports the model information and reproducibility constraints for the present exercise.

The second set consists of **source-defined votes**. These are proposals for which a source assigns a defined archetypal vote from the available text. For the human coding, this set contains only proposals on which the Left and Right archetypes were expected to vote differently. Hence, the human-defined set is already an informative left–right set. For the LLMs, the source-defined set is broader: a proposal may receive defined predictions for some or all of the eight archetypes, depending on whether the model treats the text as sufficiently substantive and transparent.

The third set consists of **archetype-informative votes**. For the human coding, this set coincides with the human-defined set by construction. For the LLMs, informativeness is based on the disagreement statistic, D_j^m, defined in Section 2.5. With eight archetypes, the baseline rule retains proposals for which at least two archetypes are predicted to vote on each side. Equivalently, a proposal is LLM-informative when $D_j^m \geq 2$.

This rule excludes proposals on which all, or almost all, archetypes are predicted to vote alike. Such proposals may be substantively important, but they carry little information for distinguishing ideological profiles. The results therefore compare affinities computed from the full roll-call matrix, from source-defined votes, and from the subset of source-defined votes that are informative for distinguishing archetypes.

In the results, these sets are labeled as follows: Level 0 refers to the full binary roll-call matrix, Level 1 to source-defined votes, and Level 2 to archetype-informative votes. For the LLMs, the baseline Level 2 rule is $D_j^m \geq 2$.

4.4 Content-blind extreme-member null benchmark

Benchmark recovery is evaluated against a content-blind diagnostic based on the voting records of the most extreme members of the Convention. Let i_L and i_R denote the members with the lowest and highest values of the one-dimensional consensus

benchmark $\hat{\theta}_i^C$, respectively. Their observed voting records define two reference profiles: a left-extreme profile P_L and a right-extreme profile P_R.

Agreement with these profiles is computed using the same agreement-rate logic as the archetype affinities. For each member, the measure records the fraction of valid votes on which the member agrees with P_L or P_R. The anchor member is excluded from the corresponding correlation to avoid mechanical self-agreement. When the null benchmark is computed on source-defined or archetype-informative vote sets, it uses the same proposal set as the corresponding archetype comparison.

These profiles are not archetypes and do not use proposal content. They are constructed ex post from realized voting behavior and require the consensus benchmark to identify the extreme members. Their purpose is therefore not to validate the archetype method. Their role is to evaluate the informativeness of the benchmark-correlation criterion: if content-blind extreme-member profiles recover the dominant benchmark ordering very strongly, then high affinity–benchmark correlation alone is not a discriminating test. The substantive validation burden must instead fall on vote informativeness and external political structure.

4.5 Cross-model bootstrap comparisons

The analysis uses paired member-level bootstrap comparisons to assess whether differences across LLMs are meaningful. Each bootstrap replication resamples Convention members with replacement. The same resampled member indices are applied to the consensus benchmark and to all affinity vectors, so that each comparison preserves the pairing between a member's benchmark position and that member's archetype affinity. The consensus benchmark is treated as fixed and is not re-estimated in each replication.

The bootstrap is used to compare cross-model recovery of the benchmark. For a given archetype and vote set, the statistic of interest is the difference between the Spearman correlations obtained by two LLMs. Percentile intervals are computed from the

bootstrap distribution of these pairwise differences. This procedure summarizes the stability of cross-model differences to changes in the member composition of the sample. It conditions on the fixed consensus benchmark and on the generated LLM profiles; it does not represent uncertainty from re-estimating the roll-call benchmark or rerunning the prompts.

4.6 External validation using political blocs

The final validation check uses observed political structure in the Convention. Party and bloc membership are not used to construct the archetypal profiles, the statistical benchmarks, or the affinity scores. They therefore provide an external criterion for assessing whether content-derived archetype affinities separate politically meaningful groups.

If an archetype captures a meaningful ideological reference profile, members belonging to an observed coalition associated with that ideological position should display systematically higher affinity to that archetype than the rest of the Convention. Conversely, members located on the opposing side should display lower affinity. This check does not rely on the latent axis estimated from the roll-call matrix, although it can be compared with the benchmark results.

The empirical analysis focuses on blocs with clear ideological interpretation and sufficient membership to make the comparison informative. For each selected bloc, the results compare the distribution of the relevant archetype affinity inside and outside the bloc. The purpose is not to reclassify all members of the Convention, but to evaluate whether archetype affinities constructed without party or bloc information recover observable political structure in the expected direction.

5 Empirical Results

This section reports the empirical evidence. It begins by documenting the stability of the statistical benchmarks and the distribution of the one-dimensional consensus

score. It then introduces a content-blind extreme-member benchmark to show why high correlation with the latent dimension is not, by itself, evidence in favor of the archetype method: observed extreme voting records recover the same ordering without using proposal content. This null benchmark motivates the two validation exercises that follow. The first evaluates whether archetype affinities recover the expected ideological signs once the comparison is restricted to votes that distinguish archetypes. The second evaluates whether content-derived affinities separate observed political coalitions that were not used to construct the archetypes. Cross-model bootstrap comparisons assess the stability of differences across LLMs, and multidimensional and domain-specific results are treated as exploratory.

5.1 Statistical benchmark stability

The first empirical step is to establish the roll-call benchmark against which the archetype affinities are evaluated. W-NOMINATE, IDEAL, and Optimal Classification are estimated on the same binary roll-call matrix and the same roster of Convention members. Each method is estimated in one, two, and three dimensions, but the main analysis uses the first dimension because it is the most stable and directly comparable across procedures.

The consensus benchmark is constructed from the two metric procedures, W-NOMINATE and IDEAL, using Generalized Procrustes Analysis.[3] Before comparison, the recovered configurations are placed in a common consensus frame. The coordinates are centered, standardized, and, when required, aligned to that frame by orthogonal Procrustes rotation.[4] The remaining reflection is fixed by orienting the first dimension so that it is positively associated with the one-dimensional W-NOMINATE estimate.

[3] Generalized Procrustes Analysis iteratively aligns several configurations and computes the configuration that minimizes the total squared Procrustes distance to the aligned inputs; see Gower (1975).

[4] Orthogonal Procrustes alignment removes rotations and reflections while preserving distances within each configuration; see Schönemann (1966).

This normalization is necessary because ideal-point configurations are identified only up to arbitrary transformations of location, scale, rotation, and reflection.

Optimal Classification is excluded from the consensus and used as a held-out diagnostic. This choice keeps the consensus benchmark based on the two metric scaling procedures, while allowing the analysis to check whether the resulting ordering is also recovered by a classification-based method.

Table 1

One-dimensional agreement among benchmark methods

Method	W-NOMINATE	IDEAL	OC	Consensus
W-NOMINATE	1.000	0.997	0.971	0.998
IDEAL	0.988	1.000	0.963	0.999
OC	0.895	0.936	1.000	0.966
Consensus	0.997	0.997	0.918	1.000

Notes: The lower triangular entries report Pearson correlations. The upper triangular entries report Spearman rank correlations. All one-dimensional scores are standardized and sign-normalized before comparison. The consensus benchmark is constructed from W-NOMINATE and IDEAL only. OC is excluded from the consensus and used as a held-out comparison.

Table 1 reports the one-dimensional agreement among W-NOMINATE, IDEAL, Optimal Classification, and the consensus benchmark. The lower triangular entries report Pearson correlations, while the upper triangular entries report Spearman rank correlations. W-NOMINATE and IDEAL are almost identical in rank terms, with a Spearman correlation of 0.997, and very close in cardinal terms, with a Pearson correlation of 0.988. The consensus benchmark is correspondingly close to both metric estimates. Its Pearson correlation is 0.997 with W-NOMINATE and IDEAL, while the Spearman correlations are 0.998 and 0.999, respectively.

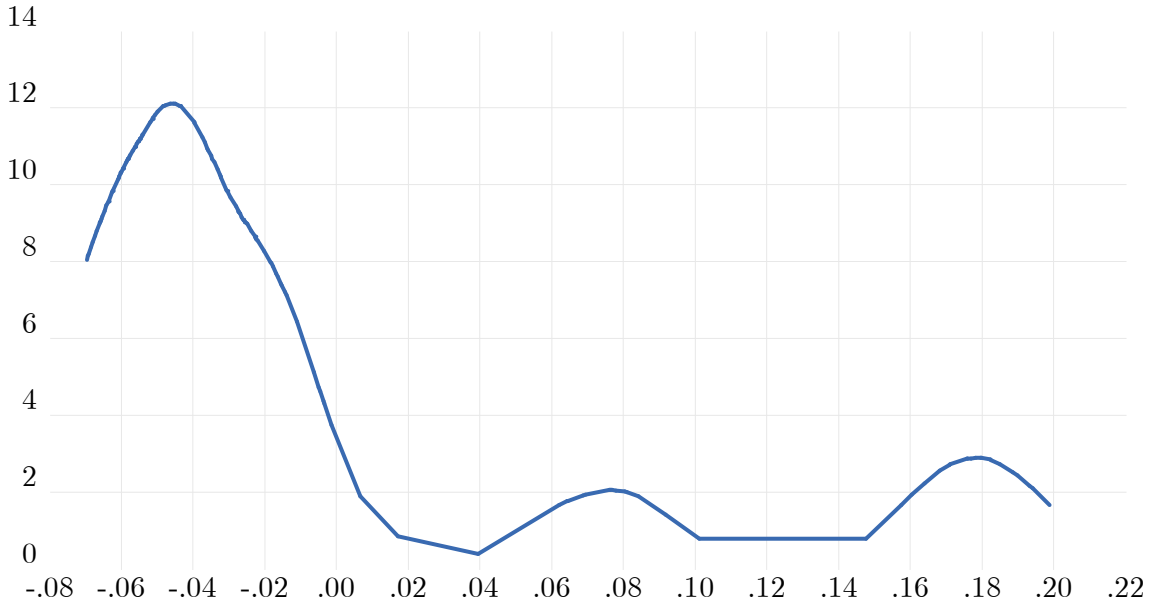
Optimal Classification is not used to construct the consensus, but it produces a closely related ordering. Its correlation with the consensus is 0.918 using Pearson correlation and 0.966 using Spearman rank correlation. Since Optimal Classification is primarily ordinal, the rank correlation is the more relevant comparison. The result shows that

the consensus benchmark is not only close to the two metric procedures used in its construction, but also strongly aligned with an alternative classification-based estimator.

Figure 1 reports the kernel density of the one-dimensional consensus score. The distribution summarizes how the benchmark orders the members of the CC along the dominant voting dimension. It is asymmetric, with a larger concentration of members on the left side of the scale, consistent with the political composition of the Convention. This provides a visual reference for the archetype-affinity results below, where left- and right-leaning profiles are expected to vary systematically across the same ordering.

Figure 1

Kernel density of the one-dimensional consensus benchmark



Notes: The figure reports the kernel density of the one-dimensional consensus benchmark constructed from W-NOMINATE and IDEAL. The horizontal axis is expressed in normalized ideal-point units and has no intrinsic substantive scale.

Appendix B.1 reports the multidimensional benchmark comparisons. Tables B.1 and B.2 show that the first two dimensions are the most stable. In the two-dimensional case, the canonical correlations between the consensus and W-NOMINATE are 0.997 and 0.936, while those between the consensus and IDEAL are 0.999 and 0.986. In the

three-dimensional case, the first two canonical correlations remain high for both metric methods, but the third dimension is less robust. This is especially clear in the direct W-NOMINATE–IDEAL comparison, where the third canonical correlation falls to 0.786, compared with 0.984 and 0.978 for the first two ranks.[5] The main analysis therefore focuses on the first benchmark dimension; the second dimension is informative, and the third is treated as exploratory.

5.2 Content-blind extreme-member null benchmark

Before evaluating the archetype affinities, it is useful to assess how much can be learned from benchmark recovery alone. The exercise in this subsection uses the observed voting records of the members located at the two extremes of the one-dimensional consensus benchmark. These profiles use no proposal content; they are constructed ex post, requiring the estimated benchmark to identify the extreme members and the vote matrix to define their profiles. They therefore provide a null benchmark for the affinity–benchmark correlation: if content-blind extreme profiles recover the dominant ordering very strongly, then high correlation with the benchmark is not, by itself, sufficient evidence in favor of the archetype method.

Agreement with the observed voting records of the most left-wing and most right-wing members is computed using the same agreement-rate logic as the archetype affinities. Table 2 reports the Spearman correlation between the consensus benchmark and agreement with each extreme-member profile. The anchor member is excluded from the corresponding correlation. The exercise is reported for the full vote matrix and for the vote sets used later in the archetype analysis.

[5] For two k-dimensional configurations, canonical correlation analysis yields k ordered correlations between optimally chosen linear combinations of the two coordinate matrices. These correlations are invariant to nonsingular linear transformations of either configuration, including rotations and reflections that leave ideal-point configurations unidentified, and therefore measure agreement between subspaces rather than agreement along particular recovered axes; see Hotelling (1936) and Mardia et al. (1979).

The table shows that the dominant benchmark ordering is recovered strongly by coherent voting records located at ideological extremes. Using all valid votes, agreement with the most left-wing member correlates with the consensus benchmark at -0.968, while agreement with the most right-wing member correlates at 0.962. The correlations remain large after restricting the comparison to Level 1 source-defined votes and Level 2 archetype-informative votes. The left-extreme profile remains close to -0.96 or stronger in all cases. The right-extreme profile is weaker on restricted vote sets, but still produces correlations around 0.90.

Table 2

Benchmark recovery by content-blind extreme-member profiles

	Human	ChatGPT	Claude	Copilot	Gemini	Grok
	Most left-wing member					
Level 0	-0.968	-0.968	-0.968	-0.968	-0.968	-0.968
Level 1	-0.961	-0.961	-0.960	-0.961	-0.964	-0.964
Level 2	-0.961	-0.964	-0.971	-0.957	-0.968	-0.967
	Most right-wing member					
Level 0	0.962	0.962	0.962	0.962	0.962	0.962
Level 1	0.904	0.910	0.906	0.906	0.909	0.911
Level 2	0.904	0.909	0.899	0.897	0.907	0.918

Notes: Entries report Spearman correlations between the first dimension of the consensus benchmark and agreement with a content-blind extreme-member profile. The profiles are the observed voting records of the members located at the two extremes of the consensus dimension. Level 0 uses all valid votes in the binary roll-call matrix and is therefore identical across coding sources. Level 1 restricts the comparison to source-defined votes. Level 2 further restricts the comparison to archetype-informative votes. The anchor member is excluded from the corresponding correlation. Correlations are oriented so that higher benchmark values denote more right-leaning positions.

This result does not validate the archetype method. It shows that high recovery of the first benchmark dimension is not a demanding criterion when the reference profile is coherent and located at an ideological extreme. The content-blind profiles recover the ordering without proposal text, ideological definitions, or substantive

classification. The relevant comparison is whether profiles constructed from proposal content alone can approach the recovery attained by vote-based extreme profiles, while also satisfying criteria that content-blind profiles cannot address, especially vote informativeness and consistency with external political structure.

5.3 Proposal classification and vote sets

The next step is to document how much of the roll-call agenda can be used for archetype-based benchmarking. The full roll-call matrix contains 4,609 proposals. Standard scaling methods can use all valid binary votes in this matrix, subject only to member-level missing values. Archetype-based benchmarking is more restrictive because a proposal must contain enough substantive information for a source to assign an archetypal vote.

Table 3

Proposal classifications by coding source

Source	Human	ChatGPT	Claude	Copilot	Gemini	Grok
Total	4609	4609	4609	4609	4609	4609
Classified	1452	2457	2469	2484	2252	2612
Economic	811	452	168	274	587	197
Social	641	782	746	1140	393	892
Legal	--	1223	1555	1070	1272	1523

Notes: "Total" refers to the total number of roll calls used to estimate the statistical benchmarks. "Classified" refers to proposals assigned by each coding source to one of the substantive categories. The human-coded benchmark classifies substantive votes as economic or social-cultural. The LLM-coded procedures also allow a legal category. Some proposals classified by LLMs as legal may correspond to economic votes in the human-coded benchmark when they involve fiscal resources, property rights, public spending, taxation, or economic authority. Votes classified as residual, procedural, ambiguous, or insufficiently informative are excluded from the three substantive categories.

Table 3 reports the number of proposals classified by each source. The human coding retained 1,452 proposals. This number should be interpreted as a left–right informative set, because the human expert retained only proposals on which the broad Left and Right archetypes were expected to vote differently. Of these proposals, 811 were classified as economic and 641 as social-cultural.

The LLMs classified larger sets of proposals, ranging from 2,252 proposals for Gemini to 2,612 for Grok. This reflects the broader LLM coding task, which includes economic, social-cultural, and legal domains. The legal category accounts for a substantial share of LLM-classified proposals, which is expected in a constitutional setting. At the same time, classification patterns differ across models. Claude and Grok assign a relatively large number of proposals to the legal category, while ChatGPT and Gemini classify more proposals as economic than the other LLMs. These differences matter because source-defined vote sets are model-specific.

These counts show why the distinction between source-defined and archetype-informative votes is necessary. A proposal may be classified by a source and still provide little information for distinguishing ideological profiles if all, or almost all, archetypes are predicted to vote alike. The next affinity comparisons therefore use two restricted vote sets. Level 1 uses source-defined votes. Level 2 further restricts the comparison to archetype-informative votes, defined in Section 4.3. For the human coding, these two sets coincide by construction. For the LLMs, Level 2 is a stricter subset of Level 1.

5.4 Baseline archetype affinities

This subsection reports the baseline association between archetype affinities and the one-dimensional consensus benchmark using source-defined votes. The comparison establishes whether content-derived profiles are oriented in the expected direction before applying the stricter informativeness rule. Left-leaning archetypes should correlate negatively with the right-oriented benchmark, while right-leaning archetypes should correlate positively.

For the human coding, the Right archetype produces a strong positive association with the consensus benchmark. The Spearman correlation is 0.884, and the Pearson correlation is 0.963. Since the consensus benchmark is oriented so that larger values correspond to more right-leaning positions, this is the expected sign. The corresponding Left profile is the complement of the Right profile on the human-coded set and therefore yields the opposite ordering.

Spearman correlations are emphasized in the remaining one-dimensional comparisons because the validation exercise focuses on whether affinity scores order members consistently with the benchmark. Pearson correlations are useful as cardinal diagnostics, but they are affected by the bounded scale of the affinity scores.

Table 4

Baseline association between LLM archetype affinities and the consensus benchmark

Archetype	ChatGPT	Claude	Copilot	Gemini	Grok
Communist	-0.932	-0.915	-0.917	-0.922	-0.922
Socialist	-0.919	-0.915	-0.917	-0.922	-0.922
Social Dem.	-0.918	-0.915	-0.917	-0.922	-0.922
Liberal	-0.917	-0.889	-0.917	-0.921	-0.895
Conservative	0.857	0.513	0.695	0.807	-0.588
Libertarian	0.808	0.513	0.693	0.805	-0.588
Left	-0.921	-0.915	-0.917	-0.922	-0.922
Right	0.822	0.513	0.694	0.807	-0.589

Notes: Entries report Spearman correlations between each LLM archetype-affinity score and the first dimension of the consensus benchmark. Correlations are computed using source-defined votes. The consensus benchmark is oriented so that larger values correspond to more right-leaning positions. Negative correlations are therefore expected for left-leaning archetypes and positive correlations for right-leaning archetypes.

Table 4 reports the baseline Spearman correlations between LLM archetype affinities and the consensus benchmark, using source-defined votes. Left-leaning archetypes are expected to correlate negatively with the right-oriented benchmark, while right-leaning archetypes are expected to correlate positively.

The baseline results show a sharp contrast between left- and right-leaning profiles. All models recover the expected negative association for Communist, Socialist, Social Democrat, Liberal, and Left archetypes. The correlations are also similar across models, generally clustered around -0.90. This indicates that the left side of the ideological scale is recovered consistently in the source-defined vote sets.

Right-leaning archetypes are less stable. ChatGPT, Copilot, and Gemini produce the expected positive correlations for Conservative, Libertarian, and Right profiles. Claude produces positive but much weaker correlations. Grok reverses the expected sign for the Conservative, Libertarian, and Right profiles in the baseline specification. This asymmetry is consistent with the possibility that model-specific priors or prompt interpretations matter more for right-leaning archetypes when the comparison includes all source-defined votes.

The baseline exercise therefore identifies the central empirical issue. Archetype affinities can align strongly with the dominant benchmark ordering, but this alignment is not equally stable across ideological sides or models. The next subsection evaluates whether this instability is reduced when the comparison is restricted to archetype-informative votes.

5.5 Vote informativeness and archetype recovery

The baseline results show that source-defined votes are not always sufficient for stable archetype recovery. The informativeness restriction provides a sharper comparison. It retains only proposals on which archetypes are predicted to disagree, using the rule defined in Section 4.3. For the human coding, the source-defined and archetype-

informative sets coincide by construction. For the LLMs, the restriction removes proposals on which all, or almost all, archetypes are predicted to vote alike.

Table 5 reports the Spearman correlations between LLM archetype affinities and the one-dimensional consensus benchmark using archetype-informative votes. The number of retained proposals differs by model, reflecting differences in both classification and archetypal disagreement. ChatGPT retains 1,874 proposals, Claude 1,260, Copilot 2,103, Gemini 1,525, and Grok 1,195.

Table 5

One-dimensional agreement using LLM archetype-informative votes, $D_j^m \geq 2$

Archetype	ChatGPT	Claude	Copilot	Gemini	Grok
Communist	-0.936	-0.936	-0.911	-0.931	-0.947
Socialist	-0.923	-0.936	-0.911	-0.931	-0.947
Social Dem.	-0.923	-0.936	-0.911	-0.931	-0.947
Liberal	-0.922	-0.917	-0.911	-0.930	-0.907
Conservative	0.871	0.878	0.820	0.873	0.868
Libertarian	0.843	0.878	0.820	0.873	0.868
Left	-0.926	-0.936	-0.911	-0.931	-0.947
Right	0.867	0.878	0.820	0.873	0.868
Votes	1874	1260	2103	1525	1195

Notes: Entries report Spearman correlations between each LLM archetype-affinity score and the first dimension of the consensus benchmark, using archetype-informative votes. The consensus benchmark is oriented so that larger values correspond to more right-leaning positions. Negative correlations are therefore expected for left-leaning archetypes and positive correlations for right-leaning archetypes. A proposal is archetype-informative when at least two archetypes are predicted to vote on each side.

The informativeness restriction changes the interpretation of the LLM results. All models now recover the expected signs for all archetypes. The negative correlations for Communist, Socialist, Social Democrat, Liberal, and Left profiles remain large in absolute value, generally between -0.90 and -0.95. More importantly, the instability

observed for right-leaning archetypes in the baseline results is removed. Conservative, Libertarian, and Right profiles now correlate positively with the consensus benchmark for every model.

The table also provides the main evidence on coherence. Within each model, ideologically related left-leaning archetypes generate very similar orderings, and the broad Left profile is closely aligned with the Communist, Socialist, and Social Democratic profiles. On the right, the informativeness restriction makes the Conservative, Libertarian, and Right profiles converge much more closely than in the source-defined baseline.

The strongest change concerns the models that were least stable in the source-defined comparison. Claude's right-leaning correlations increase from approximately 0.51 in the baseline table to about 0.88 after the informativeness restriction. Grok's right-leaning profiles, which had the wrong sign in the baseline specification, become positive and close to 0.87. Copilot remains the weakest right-leaning recovery among the five models, but its correlations remain positive and large, around 0.82.

These results support the main methodological claim. The relevant condition is not only whether a proposal is substantive enough to classify, but whether it distinguishes ideological profiles. Votes on which archetypes agree contribute little to ideological discrimination and can weaken the association between affinity scores and the benchmark. Once the comparison is restricted to archetype-informative votes, the affinity measures are correctly oriented for every model and every archetype, and cross-model differences are substantially reduced.

5.6 Cross-model bootstrap comparisons

The previous subsection shows that the informativeness restriction reduces the cross-model instability observed in the baseline results. This subsection evaluates whether the remaining differences across LLMs are substantively and statistically meaningful.

The comparison focuses on the broad Left and Right archetypes, which are common to all models and map directly onto the dominant benchmark dimension.

Table 6 reports pairwise differences in Spearman correlations between each model's archetype affinity and the one-dimensional consensus benchmark. Correlations are computed using each model's archetype-informative vote set. The table uses the signed correlations: stronger recovery is more negative for the Left archetype and more positive for the Right archetype. Differences are computed in the column order shown, as the earlier model minus the later model. Thus, for the Left archetype, a positive difference means weaker recovery by the earlier model; for the Right archetype, a positive difference means stronger recovery by the earlier model.

Table 6

Pairwise differences in LLM recovery of the consensus benchmark, archetype-informative votes

Archetype	ChatGPT	Claude	Copilot	Gemini	Grok
ChatGPT/Left	---	[0.002,0.021]	[-0.027,-0.005]	[-0.001,0.012]	[0.008,0.038]
Claude/Left	0.010	---	[-0.044,-0.011]	[-0.013,0.001]	[0.000,0.025]
Copilot/Left	-0.015	-0.025	---	[0.008,0.035]	[0.017,0.061]
Gemini/Left	0.005	-0.005	0.020	---	[0.003,0.033]
Grok/Left	0.021	0.011	0.036	0.016	---
ChatGPT/Right	---	[-0.024,0.000]	[0.024,0.077]	[-0.016,0.003]	[-0.025,0.022]
Claude/Right	-0.011	---	[0.030,0.092]	[-0.001,0.012]	[-0.015,0.034]
Copilot/Right	0.047	0.058	---	[-0.085,-0.027]	[-0.083,-0.020]
Gemini/Right	-0.006	0.005	-0.053	---	[-0.020,0.029]
Grok/Right	-0.001	0.010	-0.048	0.005	---

Notes: Entries below the diagonal report observed differences in Spearman correlations. Entries above the diagonal report 95 percent percentile bootstrap intervals from 10,000 paired member-resampling replications. In each replication, the same resampled member indices are applied to the consensus benchmark and to all affinity vectors. The consensus benchmark is treated as fixed and is not re-estimated. Differences are computed in the column order shown: ChatGPT, Claude, Copilot, Gemini, Grok.

The bootstrap comparisons reinforce the main result from the informativeness restriction. For the Left archetype, all models recover the expected negative association with the consensus benchmark, and the remaining differences are small. Some pairwise intervals exclude zero, but the magnitudes are limited: the largest observed difference is 0.036 in absolute value. Grok and Claude recover the Left profile somewhat more strongly, while Copilot is the weakest, but the differences do not alter the substantive conclusion that all models recover the left side of the benchmark ordering.

The Right archetype shows a sharper pattern. Among ChatGPT, Claude, Gemini, and Grok, all six pairwise bootstrap intervals include zero. Copilot is the exception. Its recovery of the Right archetype is weaker than that of the other four models, with observed differences of about 0.05 to 0.06 and bootstrap intervals that exclude zero in the relevant comparisons.

Thus, after restricting the comparison to archetype-informative votes, cross-model differences are substantially reduced. The main remaining heterogeneity is not a general instability of the method across LLMs, but a weaker recovery of the right-leaning profile by Copilot. This conclusion is conditional on the fixed consensus benchmark, the generated LLM profiles, and the vote sets used in the comparison. The bootstrap summarizes stability to member resampling; it does not represent uncertainty from re-estimating the roll-call benchmark or rerunning the prompts.

5.7 External validation using political blocs

The second validation exercise uses observed political blocs in the Convention. These blocs are not used to construct the archetypal profiles, estimate the statistical benchmarks, or compute the affinity scores. They therefore provide an external criterion for evaluating whether content-derived affinities recover politically meaningful structure.

Table 7 reports the separation between selected blocs and the rest of the Convention. The statistic is the area under the ROC curve (AUC), interpreted as the probability that a randomly selected bloc member has higher affinity to the relevant archetype than a randomly selected non-member. An AUC of 0.5 indicates no separation, while an AUC of 1 indicates perfect separation. The table considers two substantively clear comparisons: the Communist

archetype against Chile Digno, a coalition composed mainly of the Communist Party, and the Conservative archetype against UDI.

Table 7

External validation using observed political blocs

	Human	ChatGPT	Claude	Copilot	Gemini	Grok
	Communist archetype vs. Chile Digno $\left(n_1 = 11, n_0 = 143\right)$					
AUC	0.714	0.771	0.766	0.801	0.759	0.780
95% CI	[0.51,0.86]	[0.57,0.90]	[0.56,0.89]	[0.60,0.92]	[0.56,0.89]	[0.58,0.90]
P	0.018	0.003	0.003	0.001	0.004	0.002
	Conservative archetype vs. UDI $\left(n_1 = 21, n_0 = 133\right)$					
AUC	0.992	0.995	0.994	0.997	0.992	0.991
95% CI	[0.80,1.00]	[0.72,1.00]	[0.76,1.00]	[0.59,1.00]	[0.80,1.00]	[0.81,1.00]
P	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Notes: Entries report the separation between members of a political bloc and the remaining members of the Convention, using affinity to the indicated archetype computed on each source's archetype-informative vote set. The AUC equals the Mann–Whitney statistic divided by $(n_1 n_0)$, with ties counted as one half. Reported p-values are two-sided and come from the tie-corrected Wilcoxon rank-sum test of the null (AUC=0.5). Confidence intervals are 95 percent Hanley–McNeil intervals computed on the logit scale and back-transformed to (0,1). The wider intervals in the Chile Digno comparison reflect the smaller size of that bloc (n_1=11). Upper bounds reported as 1.00 are rounded from values below one. The human-coded benchmark does not define Communist or Conservative archetypes. The human column therefore uses (1-Right) affinity as a coarse non-right proxy in the Chile Digno comparison and Right affinity as a coarse right-wing proxy in the UDI comparison. The archetypal profiles are constructed from proposal content alone, without member identities or bloc labels.

The results show that archetype affinities recover observed political structure in the expected direction. For Chile Digno, all AUC estimates exceed 0.70 and are statistically different from 0.5. The LLM-specific Communist archetypes produce AUCs between 0.759 and 0.801, while the coarser human non-right proxy produces an AUC of 0.714. This difference is informative. It suggests that more specific archetype definitions contain within-left discriminatory information that is not captured by a broad non-

right profile. The separation is substantively meaningful, although not perfect, which is expected given the small size of the bloc and the fact that Chile Digno is compared with all other Convention members rather than only with ideologically distant members.

The UDI comparison is stronger. Affinity to the Conservative archetype almost perfectly separates UDI members from the rest of the Convention. The AUCs range from 0.991 to 0.997 across sources, and all p-values are below 0.001. Copilot, which is the weakest model in the right-side benchmark correlation comparison, produces the highest AUC in this bloc-validation exercise. This indicates that its weaker right-side recovery in Table 6 reflects calibration against the latent benchmark more than an inability to identify right-wing bloc membership.

This evidence carries different information from the benchmark correlations reported above. The bloc labels are external to both the scaling model and the archetype construction. The result therefore supports the substantive interpretation of the affinity scores: profiles constructed from proposal content alone generate affinities that recover observed political organization, not merely the ordering of a vote-derived latent dimension.

5.8 Exploratory multidimensional and domain-specific evidence

The preceding results focus on the first benchmark dimension because it is the most stable component of the roll-call scaling results and because it corresponds to the dominant left–right ordering recovered by the conventional estimators. The multidimensional estimates are nevertheless informative as diagnostics. Appendix B.1 reports the two- and three-dimensional benchmark comparisons. The first two dimensions are reasonably stable across W-NOMINATE and IDEAL, while the third dimension is less robust. For this reason, the multidimensional evidence is treated as exploratory rather than as a separate validation target.

The same caution applies to domain-specific affinities. Classifying proposals as economic, social-cultural, legal, or residual makes it possible to decompose archetype affinities by substantive area. This is useful for evaluating whether the same archetypal profile is driven mainly by one type of proposal or whether it is recovered across several issue areas. It should not be interpreted mechanically as evidence of separate ideological dimensions. In a constitutional setting, substantively different domains may still align with the same dominant cleavage.

The domain-specific results are therefore used to assess robustness and concentration, not to redefine the ideological space. If an archetype affinity is recovered only in one narrow domain, the interpretation is weaker than if it appears across substantively distinct vote sets. Conversely, differences across domains can reveal where a broad archetype is more or less informative. This is particularly relevant for legal and institutional proposals, which are central to constitutional drafting but may not map cleanly onto conventional economic or social categories.

Overall, the exploratory evidence reinforces the main interpretation. The first dimension remains the appropriate benchmark for the central validation exercises. Multidimensional and domain-specific analyses are useful for diagnosing residual structure and heterogeneity in the vote set, but they do not change the main conclusion: archetype affinities are most reliable when computed on votes that distinguish ideological profiles and when their interpretation is checked against external political structure.

6 Concluding remarks

This paper evaluates whether explicit ideological archetypes can serve as content-anchored benchmarks for roll-call analysis. Standard scaling methods recover latent positions from the vote matrix, without using the substance of the proposals. The archetype-based approach reverses this information structure. It constructs hypothetical voting profiles from proposal content and ideological definitions, and then compares observed legislators with those profiles through affinity scores.

The application to the Chilean Constitutional Convention shows that archetype affinities can reproduce the dominant left–right ordering recovered by W-NOMINATE and IDEAL. That result, however, is not sufficient by itself. Content-blind profiles based on the observed voting records of the most extreme members also recover the same ordering strongly. Benchmark recovery is therefore useful as an orientation check, but it does not carry the main validation burden.

The central condition is vote informativeness. Affinity scores are most reliable when computed on proposals where archetypes are predicted to disagree. Votes on which all, or almost all, archetypes vote alike may be substantively important, but they contribute little to distinguishing ideological profiles. In the baseline comparison, right-leaning LLM profiles are more model-dependent, and one model reverses the expected sign. Once the comparison is restricted to archetype-informative votes, all models recover the expected signs for both left- and right-leaning profiles, and cross-model differences become substantially smaller.

The external validation exercise provides the strongest evidence for the substantive interpretation of the affinities. Archetypal profiles constructed without observing member identities, party labels, bloc membership, or votes separate observed political coalitions in the expected direction. Affinity to the Communist archetype separates members of Chile Digno from the rest of the Convention, and affinity to the Conservative archetype almost perfectly separates UDI members. This result does not depend on the latent benchmark axis and therefore addresses a different question from standard roll-call scaling: whether content-derived ideological profiles recover observable political organization.

The method is not a substitute for W-NOMINATE, IDEAL, or Optimal Classification. It answers a different question. Scaling methods locate legislators relative to one another from their voting patterns. Archetype affinities evaluate how closely those voting records resemble explicit ideological reference profiles. Their value lies in making the substantive benchmark transparent. Their reliability depends on the quality of the

proposal text, the definition of the archetypes, the informativeness of the selected votes, and the stability of the coding source.

The results also qualify the use of LLMs in political measurement. LLM-generated archetypes can be informative, but model heterogeneity matters. In the unrestricted comparison, right-leaning archetypes are less stable across models than left-leaning archetypes. This asymmetry is consistent with concerns about model-specific ideological priors. The informativeness restriction reduces, but does not eliminate, cross-model differences. For that reason, applications of this approach should report prompts, model versions, vote-selection rules, and cross-model comparisons rather than treating a single model output as definitive.

Several limitations remain. The analysis depends on the textual information available for each proposal. Votes that are politically informative but textually opaque cannot be used to construct valid archetypal profiles without additional information. The archetypes considered here are stylized and deliberately simplified; other ideological definitions could generate different profiles. The empirical application is also a constitutional convention with an unusually broad agenda, so further work should evaluate whether the approach travels to ordinary legislatures, parliamentary systems, and settings with different party structures.

Archetype-based benchmarking is therefore best understood as a complement to conventional roll-call scaling. It adds explicit ideological reference profiles to the latent structure recovered from votes. When restricted to informative proposals and checked against external political structure, these profiles provide a transparent way to interpret legislative behavior relative to stylized ideological positions.

References

Alemán, E., and P. Navia (2023). "Chile's Failed Constitution: Democracy Wins." *Journal of Democracy*, 34(2), 90–104. DOI/link: https://doi.org/10.1353/jod.2023.0014

Barberá, P. (2015). "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis*, 23(1), 76–91. DOI/link: https://doi.org/10.1093/pan/mpu011

Binding, G., and L. F. Stoetzer (2023). "Non-Separable Preferences in the Statistical Analysis of Roll Call Votes." *Political Analysis*, 31(3), 352–365. DOI/link: https://doi.org/10.1017/pan.2022.11

Bonica, A. (2014). "Mapping the Ideological Marketplace." American Journal of *Political Science*, 58(2), 367–386. DOI/link: https://doi.org/10.1111/ajps.12062

Burnham, M. (2024). "Semantic Scaling: Bayesian Ideal Point Estimates with Large Language Models." *Working paper*. Link: https://arxiv.org/abs/2405.02472

Carroll, R., J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal (2009). "Comparing NOMINATE and IDEAL: Points of Difference and Monte Carlo Tests." *Legislative Studies Quarterly*, 34(4), 555–591. DOI/link: https://doi.org/10.3162/036298009789869727

Clinton, J. D., and S. Jackman (2009). "To Simulate or NOMINATE?" *Legislative Studies Quarterly*, 34(4), 593–621. DOI/link: https://doi.org/10.3162/036298009789869763

Clinton, J. D., S. Jackman, and D. Rivers (2004). "The Statistical Analysis of Roll Call Data." *American Political Science Review*, 98(2), 355–370. DOI/link: https://doi.org/10.1017/S0003055404001194

Convención Constitucional de Chile (2022). *Propuesta Constitución Política de la República de Chile*. Santiago: Convención Constitucional. Link: https://www.chileconvencion.cl/wp-content/uploads/2022/07/Texto-CPR-2022.pdf

Converse, P. E. (1964). "The Nature of Belief Systems in Mass Publics." In D. E. Apter, ed., *Ideology and Discontent*. New York: Free Press, 206–261.

Davis, O. A., M. J. Hinich, and P. C. Ordeshook (1970). "An Expository Development of a Mathematical Model of the Electoral Process." *American Political Science Review*, 64(2), 426–448. DOI/link: https://doi.org/10.2307/1953842

Di Leo, R., C. Zeng, E. Dinas, and R. Tamtam (2025). "Mapping (A)Ideology: A Taxonomy of European Parties Using Generative LLMs as Zero-Shot Learners." *Political Analysis*, 33(4), 456–463. DOI/link: https://doi.org/10.1017/pan.2025.7

Enelow, J. M., and M. J. Hinich (1984). *The Spatial Theory of Voting: An Introduction*. Cambridge: Cambridge University Press.

Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica*, 87(4), 1307–1340. DOI/link: https://doi.org/10.3982/ECTA16566

Gilardi, F., M. Alizadeh, and M. Kubli (2023). "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy*

of Sciences, 120(30), e2305016120. DOI/link: https://doi.org/10.1073/pnas.2305016120

Gower, J. (1975). "Generalized Procrustes Analysis." *Psychometrika*, 40(1), 33–51.

Halterman, A. (2025). "Synthetically Generated Text for Supervised Text Analysis." *Political Analysis*, 33(2), 181–194. DOI/link: https://doi.org/10.1017/pan.2024.31

Hanley, J. and B. McNeil (1982). "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) curve." *Radiology,* 143(1), 29–36.

Heckman, J. J., and J. M. Snyder Jr. (1997). "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators." *RAND Journal of Economics*, 28, S142–S189. Link: https://www.jstor.org/stable/2555817

Heseltine, M., and B. Clemm von Hohenberg (2024). "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics*, 11(1). DOI/link: https://doi.org/10.1177/20531680241236239

Hotelling, H. (1936). "Relations Between Two Sets of Variates." *Biometrika*, 28(3/4), 321–377.

Imai, K., J. Lo, and J. Olmsted (2016). "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review*, 110(4), 631–656. DOI/link: https://doi.org/10.1017/S000305541600037X

Jackman, S. (2001). "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis*, 9(3), 227–241. DOI/link: https://doi.org/10.1093/oxfordjournals.pan.a004873

Kreutner, M., M. Lutz, and M. Strohmaier (2026). "Persona-Driven Simulation of Voting Behavior in the European Parliament with Large Language Models." *Findings of the Association for Computational Linguistics: EACL 2026*, 490–511. DOI/link: https://doi.org/10.18653/v1/2026.findings-eacl.25

Larraín, G., G. Negretto, and S. Voigt (2023). "How Not to Write a Constitution: Lessons from Chile." *Public Choice*, 194, 233–247. DOI/link: https://doi.org/10.1007/s11127-023-01046-z

Lauderdale, B. E., and A. Herzog (2016). "Measuring Political Positions from Legislative Speech." *Political Analysis*, 24(3), 374–394. DOI/link: https://doi.org/10.1093/pan/mpw017

Laver, M., K. Benoit, and J. Garry (2003). "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review*, 97(2), 311–331. DOI/link: https://doi.org/10.1017/S0003055403000698

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

Lee, D. S., E. Moretti, and M. J. Butler (2004). "Do Voters Affect or Elect Policies? Evidence from the U.S. House." *Quarterly Journal of Economics*, 119(3), 807–859. DOI/link: https://doi.org/10.1162/0033553041502153

Lin, H., and Y. Zhang (2025). "Navigating the Risks of Using Large Language Models for Text Annotation in Social Science Research." *Social Science Computer Review*, online first. DOI/link: https://doi.org/10.1177/08944393251366243

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Academic Press.

Martin, A. D., and K. M. Quinn (2002). "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis*, 10(2), 134–153. DOI/link: https://doi.org/10.1093/pan/10.2.134

Morales-Bader, D., R. D. Castillo, R. F. A. Cox, and C. Ascencio-Garrido (2023). "Parliamentary Roll-Call Voting as a Complex Dynamical System: The Case of Chile." *PLOS ONE*, 18(4), e0281837. DOI/link: https://doi.org/10.1371/journal.pone.0281837

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott (2025). "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods*, 13(2), 264–281. DOI/link: https://doi.org/10.1017/psrm.2024.64

Poole, K. T. (2000). "Nonparametric Unfolding of Binary Choice Data." *Political Analysis*, 8(3), 211–237. DOI/link: https://doi.org/10.1093/oxfordjournals.pan.a029813

Poole, K. T. (2005). *Spatial Models of Parliamentary Voting*. Cambridge: Cambridge University Press. DOI/link: https://doi.org/10.1017/CBO9780511614644

Poole, K. T., and H. Rosenthal (1985). "A Spatial Model for Legislative Roll Call Analysis." American *Journal of Political Science*, 29(2), 357–384. DOI/link: https://doi.org/10.2307/2111172

Poole, K. T., and H. Rosenthal (1991). "Patterns of Congressional Voting." *American Journal of Political Science*, 35(1), 228–278. DOI/link: https://doi.org/10.2307/2111445

Poole, K. T., and H. Rosenthal (1997). *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Poole, K. T., and H. Rosenthal (2001). "D-NOMINATE after 10 Years: A Comparative Update to Congress: A Political-Economic History of Roll-Call Voting." *Legislative Studies Quarterly*, 26(1), 5–29. DOI/link: https://doi.org/10.2307/440401

Poole, K. T., J. B. Lewis, J. Lo, and R. Carroll (2011). "Scaling Roll Call Votes with wnominate in R." *Journal of Statistical Software*, 42(14), 1–21. DOI/link: https://doi.org/10.18637/jss.v042.i14

Potthoff, R. F. (2018). "Estimating Ideal Points from Roll-Call Data: Explore Principal Components Analysis, Especially for More Than One Dimension?" *Social Sciences*, 7(1), 12. DOI/link: https://doi.org/10.3390/socsci7010012

Rozado, D. (2024). "The Political Preferences of LLMs." *PLOS ONE*, 19(7), e0306621. DOI/link: https://doi.org/10.1371/journal.pone.0306621

Schönemann, P. (1966). "A Generalized Solution of the Orthogonal Procrustes Problem." *Psychometrika*, 31(1), 1–10.

Shaffer, W. R. (1989). "Rating the Performance of the ADA in the U.S. Congress." *Western Political Quarterly*, 42(1), 33–51. DOI/link: https://doi.org/10.1177/106591298904200103

Sheskin, D. (2000). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall.

Slapin, J. B., and S.-O. Proksch (2008). "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science*, 52(3), 705–722. DOI/link: https://doi.org/10.1111/j.1540-5907.2008.00338.x

Appendix A
Prompts and Implementation Details

This appendix documents the prompts and implementation details used to construct the LLM-based archetypal voting profiles. The prompts were designed to preserve the information restriction used in the paper: models received proposal content and archetype definitions, but not observed votes, Convention member identities, party or bloc membership, or benchmark coordinates. Section A.1 reports the prompt used to classify proposals by substantive domain. Section A.2 reports the archetype definitions. Section A.3 reports the requested output structure. Section A.4 describes the model information and reproducibility constraints associated with the LLM implementation.

A.1 Classification of proposal type

The following prompt was used to classify each proposal by substantive domain:

> *Act as an expert in economics, political science, and constitutional law. I will be giving you a proposal (in Spanish) of the voting minutes of the constituent process of Chile carried out between 2021 and 2022.*
>
> *Assign a numerical value indicating the type of proposal considered. Assign a value of 1 if the proposal refers to mainly economic issues; a value of 2 if it refers to mainly social, value or cultural issues; a value of 3 if it refers mainly to legal aspects; and a value of 0 otherwise. Always assign only a numerical value according to the instructions I provided.*

Thus, the proposal type is coded as:

$$q_j = \begin{cases} 1 & \text{if proposal } j \text{ refers mainly to economic issues} \\ 2 & \text{if proposal } j \text{ refers mainly to social, value, or cultural issues} \\ 3 & \text{if proposal } j \text{ refers mainly to legal issues} \\ 0 & \text{otherwise} \end{cases},$$

The residual category includes procedural or organizational matters, proposals for which the minutes do not provide enough information, and votes referring to documents or circumstances whose substantive content cannot be inferred from the available text.

A.2 Archetypal voting prompts

For each archetype, the model was given the following common instruction:

> *Act as an expert in economics, political science, and constitutional law. I will provide a proposal, written in Spanish, from the voting minutes of the Chilean constituent process carried out between 2021 and 2022.*

The model was then given a definition of the ideological archetype and asked to determine how that archetype should have voted. The output was coded as:

$$b_{aj} = \begin{cases} 0 & \text{if archetype } a \text{ should reject proposal } j \\ 1 & \text{if archetype } a \text{ should approve proposal } j, \\ 2 & \text{otherwise} \end{cases}$$

The residual value 2 includes cases in which there is not enough information, the archetype should be indifferent, or the vote is not relevant according to the archetype's position.

The common closing instruction was:

> *Determine how this archetypal constituent should have voted on the proposal. Assign a value of 1 if the archetype should have approved the proposal; a value*

of 0 if the archetype should have rejected it; and a value of 2 otherwise. Make sure that only one option can be chosen, and that only a numerical value is assigned.

The archetype definitions were the following:

Communist

A communist constituent has the following characteristics: in economics, this archetype supports a centrally planned economy with communal or state ownership of the means of production; in social aspects, it aims for equality of outcomes, supporting redistributive policies for wealth and opportunities; in legal aspects, it endorses a strong state role, which may limit individual freedoms for the perceived collective good.

Socialist

A socialist constituent has the following characteristics: in economics, this archetype favors a mixed economy with both state planning and market elements, significant public sector and social welfare programs; in social aspects, focuses on reducing inequalities through state-provided education, health, and social services; while in legal aspects supports stronger state regulation compared to capitalist systems but maintains democratic governance.

Social democrat

A social democrat constituent has the following characteristics: in economics, this archetype advocates for a market economy with substantial state intervention to correct market failures and provide social welfare; in social aspects, supports inclusive policies and social protection within a capitalist framework; and in legal matters, backs a strong rule of law to guarantee both economic and civil rights. It favors a mixed economy with both state planning and market elements, significant public sector and social welfare programs.

Liberal

A liberal constituent (in the American sense) has the following characteristics: in economics, generally supports market intervention with fairness-oriented regulations; in social issues, champions civil rights, gender equality, and minority rights; while on legal aspects defends individual freedoms and rule of law, focusing on social justice.

Conservative

A conservative constituent (in the American sense) has the following characteristics: In economics, prefers a free market with limited government intervention, low taxes, and minimal regulations; in social issues holds traditional values, possibly restrictive on immigration and reproductive rights; and in legal matters supports established order, law and order, and is cautious about rapid or extensive legislative changes.

Libertarian

A libertarian constituent has the following characteristics: in economics, favors a completely free market without state intervention, lower taxes and minimal regulations. In social aspects, it promotes individual liberties and civil rights, including the right to property, free association, and freedom of expression. In legal aspects, it advocates a minimum governance that is limited to protecting life, liberty, and property. An important emphasis is on promoting equality before the law and respect for the rule of law.

Left-wing

A left-wing constituent has the following characteristics: in economics, supports greater state intervention in the economy to promote social equality and regulate corporations; in social issues strongly defends civil rights, equality, and progressive policies; and in legal matters, often favors reforms that expand individual and collective rights and promote social justice.

Right-wing

A right-wing constituent has the following characteristics: In economics, tends towards minimal state intervention in the economy, favoring free markets and lower taxes; in social issues often prioritizes traditional values and a more nationalistic view; while on legal matters generally supports a conservative policy approach, emphasizing stability and order.

A.3 Combined prompt for structured output

In addition to the separate prompts, a combined prompt was used to generate a structured spreadsheet. The model was instructed to review Word files containing the CC voting minutes and produce a CSV-compatible table with the following columns:

"Minute", "Vote", "Type", "Communist", "Socialist", "SocialDem", "Liberal", "Conservative", "Libertarian", "Left", "Right".

The prompt defined the first two columns as the minute number and vote number. The variable Type followed the coding rule in Appendix A.1. The remaining columns followed the archetype-specific voting rule in Appendix A.2. The model was instructed to assign one and only one numerical value in each case.

A.4 Model information and reproducibility constraints

The LLM coding was implemented in June 2026 using five commercial conversational models accessed through their web interfaces: ChatGPT 5.5, Claude Opus 4.8, Microsoft Copilot, Gemini 3.5, and Grok 4. Microsoft Copilot did not display an exact version label to the user and is therefore reported by service name only.

All models received the same substantive-domain prompt, archetype definitions, and structured-output instructions reported in Sections A.1–A.3. No sampling

parameters, such as temperature, were exposed to the user. The prompts did not provide observed votes, Convention member names, party or bloc membership, benchmark coordinates, or information derived from the roll-call scaling procedures. The outputs were generated from proposal content and archetype definitions alone and were only later compared with observed voting behavior through the affinity measures defined in Section 2.

More refined prompts or newer model versions could plausibly improve baseline performance. The central improvement documented in the paper, however, does not come from changing prompts, rerunning the models, adding feedback, or changing the model version. The same generated profiles are evaluated under different vote-selection rules. Initially weaker cases, such as Grok's sign reversal and Claude's weak right-leaning correlations in the unrestricted comparison, improve when the comparison is restricted to archetype-informative votes.

Reproducibility therefore has two components. The first is procedural: prompts, output structure, coding rules, model information, and vote-selection rules are documented so that the exercise can be replicated as closely as possible. The second is robustness to model dependence: the empirical analysis compares five LLMs under the same coding structure and reports cross-model differences, rather than treating any single model output as definitive.

Appendix B
Benchmark Comparisons and Robustness

This appendix reports additional benchmark comparisons and robustness checks. The main text focuses on the first dimension of the consensus benchmark because it is the most stable component across roll-call scaling procedures and provides the clearest reference ordering for the archetype-affinity measures. The appendix documents the multidimensional comparisons underlying that choice, reports restricted-sample benchmark checks, and presents exploratory domain-specific results.

B.1 Multidimensional benchmark comparisons

Tables B.1 and B.2 compare W-NOMINATE, IDEAL, Optimal Classification (OC), and the consensus benchmark in two and three dimensions. The consensus benchmark is constructed from W-NOMINATE and IDEAL using Generalized Procrustes Analysis; OC is excluded from the consensus because it is primarily ordinal rather than cardinal.

All configurations are centered, standardized, aligned to the common consensus frame, and sign-normalized so that the first axis is positively associated with the one-dimensional W-NOMINATE estimate. In each block, lower triangular entries report Pearson correlations between matched axes after alignment, while upper triangular entries report ordered canonical correlations.

The two-dimensional results show close agreement between the consensus benchmark and the two metric procedures used to construct it. In Table B.1, the canonical correlations are 0.997 and 0.936 with W-NOMINATE, and 0.999 and 0.986 with IDEAL. The matched-axis Pearson correlations tell the same story. OC is also broadly consistent with this structure, although it is not used to define the consensus.

Table B.1

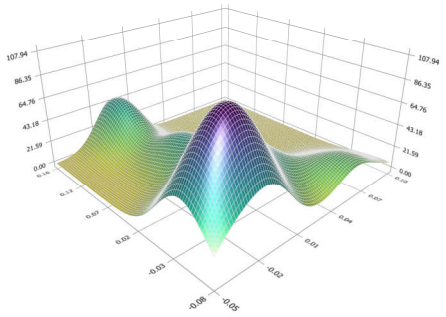
Two-dimensional agreement among benchmark methods

Method	W-NOMINATE	IDEAL	OC	Consensus
W-NOMINATE	1.000	0.983	0.917	0.997
IDEAL	0.862	1.000	0.982	0.999
OC	0.908	0.781	1.000	0.964
Consensus	0.903	0.996	0.818	1.000
W-NOMINATE	1.000	0.893	0.896	0.936
IDEAL	0.942	1.000	0.793	0.986
OC	0.902	0.885	1.000	0.809
Consensus	0.988	0.982	0.905	1.000

Notes: The table contains two 4×4 blocks. Lower triangular entries report Pearson correlations between matched axes after alignment to the common consensus frame; here the block index denotes the recovered axis. Upper triangular entries report canonical correlations; here the block index denotes the rank of the canonical correlation. Canonical correlations are ordered, basis-invariant measures of subspace agreement and do not correspond to individual recovered axes, so they are unaffected by the alignment. All configurations are centered, standardized, aligned to the common consensus frame, and sign-normalized so that the first axis is positively associated with the one-dimensional W-NOMINATE estimate. The consensus benchmark is constructed from W-NOMINATE and IDEAL only. OC is reported as an additional comparison but is not used to construct the consensus.

Figure B.1

Bivariate kernel density of the two-dimensional consensus benchmark



Notes: The figure reports the bivariate kernel density of the first two dimensions of the consensus benchmark constructed from W-NOMINATE and IDEAL. The axes are expressed in normalized ideal-point units and should be interpreted as relative coordinates, not as substantive ideological magnitudes.

Figure B.1 presents a surface plot of the bivariate kernel density for the first two dimensions of the consensus benchmark. The figure summarizes the distribution of Convention members in the two-dimensional consensus space. It shows the concentration of members along the dominant first dimension and the extent to which the second dimension adds dispersion around that ordering. The figure is used as a descriptive complement to the canonical-correlation results in Table B.1.

Table B.2

Three-dimensional agreement among benchmark methods

Method	W-NOMINATE	IDEAL	OC	Consensus
W-NOMINATE	1.000	0.984	0.961	0.996
IDEAL	0.843	1.000	0.971	0.997
OC	0.681	0.817	1.000	0.973
Consensus	0.958	0.962	0.782	1.000
W-NOMINATE	1.000	0.978	0.907	0.994
IDEAL	0.927	1.000	0.887	0.994
OC	0.895	0.877	1.000	0.901
Consensus	0.979	0.984	0.902	1.000
W-NOMINATE	1.000	0.786	0.653	0.946
IDEAL	0.946	1.000	0.839	0.943
OC	0.943	0.885	1.000	0.791
Consensus	0.986	0.986	0.927	1.000

Notes: The table contains three 4×4 comparison blocks. Lower triangular entries are Pearson correlations between matched axes after alignment to the consensus frame; upper triangular entries are ordered canonical correlations, which measure subspace agreement and do not correspond to individual axes. All configurations are centered, standardized, aligned, and sign-normalized so that the first axis is positively associated with the one-dimensional W-NOMINATE estimate. The consensus benchmark uses W-NOMINATE and IDEAL only; OC is reported as a held-out comparison.

The three-dimensional results are weaker, especially for the third canonical correlation. In Table B.2, the first two canonical correlations between the consensus and the two metric methods remain high: 0.996 and 0.994 for W-NOMINATE, and 0.997 and 0.994 for IDEAL. The third canonical correlations fall to 0.946 and 0.943, respectively. The direct W-NOMINATE–IDEAL comparison shows a sharper decline

in the third canonical correlation, from 0.984 and 0.978 in the first two ranks to 0.786 in the third.

The evidence supports the use of the first benchmark dimension as the main reference ordering in the empirical analysis. The second dimension is informative and broadly stable across the metric procedures, but the third dimension is less robust and is treated as exploratory.

B.2 Robustness to the human-coded vote set

The main benchmark is estimated using the full binary roll-call matrix. Archetype-based benchmarking, however, cannot use every roll call. Some proposals are procedural or organizational, others are textually opaque, and others do not contain enough information in the available minutes to assign a substantive archetypal vote. This raises a practical question: whether the dominant benchmark ordering depends on votes that are unavailable to the content-based procedure.

Table B.3

Robustness of one-dimensional benchmarks to the human-classified vote set

Method	W-NOMINATE	IDEAL	OC	Consensus
Pearson	0.996	0.995	0.989	0.997
Spearman	0.985	0.967	0.984	0.975

Notes: The table reports the correlation between each one-dimensional benchmark estimated using the full roll-call matrix and the same benchmark estimated using only votes classified by the human-coded procedure. All scores are standardized and sign-normalized so that they are positively associated with the one-dimensional W-NOMINATE estimate. The consensus benchmark is constructed from W-NOMINATE and IDEAL.

To assess this, the one-dimensional benchmarks are re-estimated using only the proposals retained by the human-coded procedure. This is a restrictive comparison because the human set includes only votes for which the broad Left and Right archetypes were expected to vote differently, and only proposals classified as economic

or social-cultural. Table B.3 reports correlations between each full-matrix benchmark and the corresponding benchmark estimated on the human-coded vote set.

The restricted vote set produces almost the same one-dimensional benchmark as the full matrix. For the consensus, the Pearson correlation between the full-matrix and restricted-matrix estimates is 0.997, while the Spearman correlation is 0.975. The corresponding correlations are also high for W-NOMINATE, IDEAL, and Optimal Classification. Thus, the dominant ideological ordering is not driven by votes that the human-coded procedure excludes as procedural, ambiguous, or insufficiently informative for a left–right archetype comparison.

B.3 Domain-specific archetype affinities

This subsection reports exploratory comparisons between domain-specific archetype affinities and the multidimensional consensus benchmark. The objective is to assess whether affinities computed separately by substantive domain contain information related to the two- and three-dimensional benchmark spaces. The exercise should not be interpreted as assigning particular substantive domains to particular benchmark axes. Canonical correlations measure association between spaces, not between individual domains and individual latent dimensions.

Table B.4 compares economic and social-cultural affinity vectors with the two-dimensional consensus benchmark. For the LLMs, the vectors contain the economic and social-cultural affinities for the Left and Right archetypes. The human comparison is reported only for the Right archetype, because the human-coded validation benchmark records the Right profile, with the Left profile implied as its complement.

The first canonical correlations in Table B.4 are high for both Left and Right affinities, generally above 0.92. This indicates that domain-specific affinities constructed from economic and social-cultural votes contain information closely aligned with the two-dimensional benchmark space. The second canonical

correlations are lower and more heterogeneous across models, especially for the Right archetype. This suggests that most of the association is concentrated in the dominant component of the comparison, while the second component varies more across sources.

Table B.4

Canonical correlations between domain-specific archetype affinities and the two-dimensional consensus benchmark

	Human	ChatGPT	Claude	Copilot	Gemini	Grok
Left/Rank 1	--	0.953	0.956	0.959	0.954	0.956
Left/Rank 2	--	0.451	0.623	0.299	0.022	0.348
Right/Rank 1	0.965	0.942	0.944	0.925	0.962	0.924
Right/Rank 2	0.430	0.295	0.613	0.275	0.142	0.125

Notes: Each cell reports an ordered canonical correlation between the domain-specific affinity vector for the indicated archetype and the two-dimensional consensus benchmark. For the LLMs, the affinity vector contains economic and social-cultural affinities. The human-coded comparison is reported only for the Right archetype because the human-coded benchmark used in the validation exercise records the Right profile and contains economic and social-cultural affinities. Canonical correlations measure association between spaces and should not be interpreted as correlations between specific substantive domains and specific latent benchmark axes.

Table B.5 extends the exercise to three dimensions for the LLMs. In this case, the domain-specific affinity vector includes economic, social-cultural, and legal affinities. The human-coded benchmark is not included because the human coding does not define a legal domain.

The three-dimensional results show the same pattern. The first canonical correlation is high for every model and both broad archetypes. The second canonical correlation remains sizable for several models, especially for the Left archetype and for Claude and Grok on the Right archetype. The third canonical correlation is generally weak, with the partial exception of Claude's Right profile. This is consistent with the benchmark evidence in Appendix B.1: the first dimension is the most stable

component, the second dimension contains additional structure, and the third dimension is less robust.

Table B.5

Canonical correlations between domain-specific archetype affinities and the three-dimensional consensus benchmark

	ChatGPT	Claude	Copilot	Gemini	Grok
Left/Rank 1	0.971	0.967	0.969	0.968	0.965
Left/Rank 2	0.784	0.801	0.613	0.501	0.800
Left/Rank 3	0.024	0.218	0.144	0.219	0.141
Right/Rank 1	0.968	0.969	0.964	0.976	0.970
Right/Rank 2	0.496	0.740	0.622	0.278	0.744
Right/Rank 3	0.036	0.498	0.149	0.205	0.030

Notes: Each cell reports an ordered canonical correlation between the domain-specific LLM affinity vector for the indicated archetype and the three-dimensional consensus benchmark. The affinity vector contains economic, social-cultural, and legal affinities. Canonical correlations measure association between two spaces and should not be interpreted as correlations between specific substantive domains and specific latent benchmark axes.

These results support using domain-specific affinities as exploratory diagnostics. They indicate that domain-level archetype affinities are related to the multidimensional benchmark space, but they do not establish a direct mapping between substantive domains and latent ideological axes. The main validation claims therefore remain those in the main text: affinities become reliable when votes are informative for distinguishing archetypes, and they recover observed political blocs that were not used in profile construction.

B.4 Robustness to alternative informativeness thresholds

The main analysis defines an LLM-coded proposal as archetype-informative when at least two archetypes are predicted to vote on each side, $D_j^m \geq 2$. This rule excludes proposals on which all archetypes agree and proposals on which disagreement is

driven by a single dissenting archetype. Two alternative thresholds are considered here.

Table B.6

One-dimensional agreement using LLM archetype-informative votes, $D_j^m \geq 3$

Archetype	ChatGPT	Claude	Copilot	Gemini	Grok
Communist	-0.938	-0.936	-0.911	-0.931	-0.947
Socialist	-0.928	-0.936	-0.911	-0.931	-0.947
Social Dem.	-0.925	-0.936	-0.911	-0.931	-0.947
Liberal	-0.925	-0.917	-0.911	-0.930	-0.907
Conservative	0.866	0.878	0.820	0.873	0.868
Libertarian	0.861	0.878	0.820	0.873	0.868
Left	-0.930	-0.936	-0.911	-0.931	-0.947
Right	0.870	0.878	0.820	0.873	0.868
Votes	1525	1260	2103	1517	1195

Notes: Entries report Spearman correlations between each LLM archetype-affinity score and the first dimension of the consensus benchmark, using archetype-informative votes. The consensus benchmark is oriented so that larger values correspond to more right-leaning positions. Negative correlations are therefore expected for left-leaning archetypes and positive correlations for right-leaning archetypes. A proposal is archetype-informative when at least three archetypes are predicted to vote on each side.

The weaker rule, $D_j^m \geq 1$, retains any proposal on which at least one archetype dissents. The results are almost identical to those reported in Table 5, indicating that the main findings do not depend on excluding one-dissenter cases. The stricter rule, $D_j^m \geq 3$, retains only proposals where at least three archetypes are on the smaller side of the predicted division. With eight archetypes, this rule selects votes closer to a bloc structure among the ideological profiles.

The stricter threshold produces results that are virtually unchanged from the baseline specification. All left-leaning archetypes retain the expected negative sign, and all right-leaning archetypes retain the expected positive sign. The magnitudes remain

high across models. The number of retained proposals falls for ChatGPT, Gemini, and Grok, but the substantive conclusions are unaffected. For Claude and Copilot, the retained vote counts are unchanged, indicating that the proposals satisfying $D_j^m \geq 2$ also satisfy the stricter bloc-like division.

The threshold checks therefore support the interpretation that the central result is not driven by a narrow cutoff. What matters is whether the comparison excludes votes on which archetypes do not meaningfully disagree. Once the vote set is restricted to proposals with archetypal disagreement, the expected ideological ordering is recovered across all LLMs under both weaker and stricter definitions of informativeness.