Elite Cleavages: Concept, Measurement, and a Verifiable AI Search Agent Methodology *

John Gerring † Connor T. Jerzak‡ Erzen Öncel §

Abstract

Scholars agree that the nature of ethnic cleavages hinges not on diversity per se but on the degree to which political elites reproduce social boundaries. Yet, we still lack a replicable, continuous, and cross-national metric of elite ethnic segregation of the party system. This paper proposes a new measure that treats the distribution of ethnic groups across legislative party delegations as an optimal-transport problem. By comparing the observed joint distribution of *groups* × *parties* with a counterfactual of perfect integration, we derive a segregation index bounded in $[0, 1]$ that is comparable across time, chambers, and countries with different party and ethnic configurations. We implement the index on an original dataset covering 368 parliamentary delegations (1990–2022), assembled from the Global Leadership Project and augmented by an AI search agent that, under country-specific closed codebooks, harvests and cites public-source evidence to infer missing party and ethnicity labels, abstains when evidence is weak or conflicting, and routes disagreements to human coders—thereby expanding coverage by 20.8% (12,898 leader cases) while preserving auditability and cross-national comparability. Validation against existing binary and survey-based measures shows that our cleavage metrics capture known cases while revealing substantial within-country and within-party variation that is missed by existing approaches. Substantively, we find patterns consistent with a simple demand-and-supply model of ethnic representation. The search-agent augmented index invites new tests of classic theories of ethnic politics, affords fine-grained diagnostics for institutional engineering, and is readily extensible to other ascriptive cleavages. We make the 770,000 distinct records on political elites, obtained by agent, open source.

Keywords: Descriptive representation; Political parties; Social groups; AI search agents.

*Authors are listed in alphabetical order. Preliminary and incomplete. For more information about dataset access, see `GlobalLeadershipProject.net/cleavage`.

†Professor, Department of Government, University of Texas at Austin. Email: `jgerring@austin.utexas.edu` URL: `liberalarts.utexas.edu/government/faculty/jg29775`

‡Assistant Professor, Department of Government, University of Texas at Austin. Email: `connor.jerzak@austin.utexas.edu` URL: `ConnorJerzak.com`

§Assistant Professor, Department of International Relations, Özyeğin University. Email: `erzen.oncel@ozyegin.edu.tr` URL: `ozyegin.edu.tr/en/faculty/erzenoncel`

Introduction

An enormous literature wrestles with challenges to governance and development posed by ethnic diversity (Alesina and La Ferrara, 2005; Easterly and Levine, 1997). Although the concern is well-founded, one must also appreciate that ethnic differences by themselves are neither unusual nor problematic. Every society is diverse in some respects, and in most instances this diversity does not prevent cooperation; it may even promote better solutions (Page, 2008). Diversity becomes consequential when latent differences are politicized (Fraenkel and Crawley, 2025; Posner, 2004). This is why it is important to distinguish ethnic differences (as measured, e.g., by ethnic fractionalization indices) from ethnic *cleavages*, our subject in this study.

Of particular concern to political scientists is the prospect that cleavages of an ethnic nature might undermine stability and democracy. The first generation of studies offers a pessimistic view, according to which ethnic politics is an invitation to civil conflict (Geertz, 1963; Horowitz, 1985; Rabushka and Shepsle, 1972; Rae and Taylor, 1970; Rustow, 1970). Some more recent studies confirm this pessimistic view (Houle, 2018).

Others are more optimistic, pointing out that the existence of ethnic cleavages does not necessarily convert elections into an ethnic census (Elischer, 2013; Fraenkel and Crawley, 2025). Moreover, the potentially damaging effects of ethnic cleavages may be mitigated by political institutions such as federalism, electoral rules, or power-sharing agreements (Chandra, 2005; Lijphart, 1977; Reilly, 2001) or by alliances with the private sector (Arriola, 2012). Indeed, the theory of consociationalism suggests that partisan cleavages based on ethnicity can be successfully mediated by appropriate institutions (Lijphart, 1969). Evidently, the existence of ethnic parties does not doom prospects for democracy (Birnir, 2006; Chandra, 2007; Ishiyama, 2009).

Even so, where voters choose leaders on the basis of who they are rather than what they stand for, this may undermine mechanisms of accountability essential for good governance. Accordingly, ethnic politics, associated with clientelistic policymaking and particularistic goods, is commonly counterposed to class politics, associated with programmatic policymaking and the provision of public goods (Chandra, 2007; Dixit and Londregan, 1996; Kitschelt, 2007; Lemarchand, 1972).

Thus, the role of ethnic identities in governance remains a point of contention. Issues are difficult to resolve because there is no generally recognized metric of ethnic cleavages. What makes one polity more ethnically divided than another, one party more ethnically defined than another, or one ethnic group more politicized than another?

In this study, we offer a new approach to the measurement of ethnic cleavages. This approach has several distinctive characteristics that, collectively, set it apart from extant work. It treats ethnicity as a political feature rather than a demographic fact. It is centered on elites rather than voters. It approaches cleavages as matters of degree rather than of kind. And it is applicable to any setting where the ethnic identity of parliamentarians can be ascertained.

Our key empirical indicator is the ethnic composition of party delegations. We describe delegations as *integrated* if there are few ethnic differences across parties and *segregated* if each party represents a different ethnic group and these groups are similar in size. To provide a precise measure of this fuzzy concept, we adopt an algorithm derived from optimal transport theory (Lott and Villani, 2009). Applied to party delegations, this generates an *ethnic cleavage index*. A low score on this index indicates a party system where ethnicity plays little role: either there are no sizeable ethnic groups (MPs are ethnically homogeneous) or parties have balanced delegations. A high score indicates a party system defined by ethnicity. Adaptations of optimal transport theory apply the same principles to political parties and ethnic groups. An *ethnic party index* measures the degree

to which individual parties are defined by ethnicity. An *ethnic cohesion index* measures the degree to which ethnic groups are concentrated in the same party.

After introducing these measures, we turn to the task of data collection. Evidently, mathematically elegant indices are useful only if they can be successfully applied to topics of theoretical interest. To meet this challenge, we propose an approach to coding the ethnicity of legislators around the world that leverages hand-coding along with a new set of LLM tools. This allows... [describe dataset]

In the final section of the paper we take up the task of explanation. Why are ethnic cleavages more marked in some polities than in others? We argue that this is largely a product of the composition of ethnic groups and of political parties. Greater fractionalization generally translates into stronger cleavages.

1 Measurement

Cleavage structures have been a preoccupation of political science and sociology for the better part of a century. Originally centered on social class (Alford, 1962; Converse, 1958), researchers soon broadened their purview to include other aspects of status and identity such as language, religion, and region (Lipset and Rokkan, 1967).[1] Contemporary studies of ethnic politics follow in this venerable tradition, with ethnicity as the umbrella term encompassing related factors such as religion, language, and region.

Two principal approaches may be identified from the literature, centered respectively on ethnic groups and political parties (Huber, 2012). One focuses on the degree to which each ethnic group consolidates behind a single party; the other focuses on the degree to which each party represents a unique ethnic group. Consider a setting with four ethnic groups (*A*, *B*, *C*, *D*) and two parties (*P1*, *P2*). If all voters from groups *A* and *B* support *P1*, and all voters from groups *C* and *D* support *P2*, *ethnic voting* is maximal and *ethnic parties* is minimal.

Beyond this conceptual disagreement lie empirical differences. When attempting to ascertain ethnic voting, some researchers focus on vote shares across constituencies. This sort of data is usually plentiful but raises problems of ecological inference since the characteristics of individual voters must be inferred from the characteristics of constituencies (obtained from census data). Other researchers rely on public opinion surveys. This provides individual-level data but is limited to contexts where national surveying is common and reliable, and where questionnaires include vote-choice (or party membership) and ethnic identity. Both data sources must contend with the specificity of census and survey data and the highly contextual dynamics of ethnicity, which may impair comparability across settings and even across surveys or censuses in the same setting (if the coding of ethnicity varies).[2]

In light of these obstacles, it is not surprising that most work on ethnic politics is limited to individual countries or regions. Recent studies center on Latin America (Madrid, 2012), Southeast Asia (Liu and Ricks, 2022; Reilly, 2021), the OECD (Hamza and Maeda, 2025), and Africa, where the topic is ubiquitous (Huber, 2012; Ishiyama, 2012). A few studies are more extensive, notably Houle, Park, and Kenny (2019), which covers sixty-five countries with data from the World Values Survey, and Fraenkel and Crawley (2025), which covers 132 countries with a wide variety of surveys.

A rather different approach categorizes individual parties as "ethnic" or "nonethnic" based on a

[1] For a recent effort, see Marks et al. (2023).

[2] For discussion of these and other difficulties encountered by survey-based research see Fraenkel and Crawley (2025, pp. 15–18).

variety of characteristics such as party name, rhetoric, policies, leadership, constituency, and expert judgments (Chandra, 2011). This corresponds to the party-centered approach to ethnic politics introduced above. Although most studies in this vein are limited to a single country or region, a few are more wide-ranging. Ishiyama (2009) codes ethnic parties in the developing world during the 1990s. Strijbis and Kotnarowski (2015, p. 463) code ethnic parties at various points over the past two decades in a handful of European countries, along with Canada and Australia. Van Cott (2007) codes indigenous parties in South America. Lublin (2014, App II) codes ethnoregional parties (considered jointly) in 80+ countries, observed at some point between 1990 and 2012.

While these studies offer broader coverage than the typical study of ethnic voting, it is important to appreciate the considerable loss of information that arises when parties are reduced to a single binary code — ethnic or nonethnic. This simplification complicates inferences one might draw about other parties (relegated to a large residual category) and the party system as a whole. Moreover, the complexity of the coding criteria means that raters must juggle a variety of dimensions that do not always point in the same direction. Since different studies of ethnic parties invoke different coding criteria, results differ somewhat across studies and are not easy to replicate.

Against this backdrop, our approach has five distinguishing features. First, it is centered on elites (representatives) rather than masses (voters).[3] Second, it treats cleavages as matters of degree rather than of kind. Third, it assigns scores to individual parties (mirroring the party-centered approach) and ethnic groups (mirroring the group-centered approach) as well as polities (the overall cleavage). Fourth, it is applicable to any setting—local, regional, or national—where the party and ethnic identity of MPs can be ascertained, raising the possibility of a truly comprehensive analysis on a global scale.

1.1 Ethnic Cleavages

In modern contexts, ethnic cleavages of any political significance are usually manifested in political parties. Where a cleavage is politically salient, parties will presumably be differentiated by ethnicity. Since the legislature is the preeminent representative body it is natural to look to legislatures if we wish to understand the character of a party at elite levels. Helpfully, legislative parties are sizeable enough to offer a basis for judgment.[4]

Our guiding assumption is that a party's ethnic orientation is reflected in the descriptive characteristics of its leadership. If the party's mission is to represent the interests of a particular ethnic group, that group is likely to dominate its parliamentary delegation. If its mission is to represent a variety of different ethnic groups, this will be reflected in a multi-ethnic delegation.[5]

This does not mean that parties with an ethnic base must trumpet their ethnic character. In Africa, most parties are formally non-ethnic; yet, it is an open secret that many are vehicles for particular ethnic groups (Berman, Eyoh, and Kymlicka, 2004). The same was true of the U.S. Republican Party through most of its history. Temperance, education, Sabbatarianism, anti-bossism, and other "reform" issues were calculated to please Protestant constituencies despite the

[3] As such, we sidestep complicated questions about voter motivation that are central to the literature on ethnic voting (Adida et al., 2017).

[4] By contrast, top leadership positions—presidents, prime ministers, party leaders—are encapsulated in a single office, which by construction can be occupied only by a single individual and thus allows no basis for distinguishing an ethnic party from a multi-ethnic one.

[5] It is of course possible for parties to dissemble. Leaders may nominate members of social groups they have no intention of representing (substantively), and for a while, they may be successful in this ruse. However, it is unlikely they will be successful for very long.

displeasure they caused for Catholics. Although the party did not proclaim itself Protestant, it was clearly responding to a constituency with roots in Protestant communities outside the South, a feature reflected in its electoral base and in its staunchly Protestant leadership (Gould, 2007; Layman, 2001; Silbey and Bogue, 1978). Accordingly, we consider the ethnic identity of a party's legislative delegation to be strong evidence of its ethnic (or non-ethnic) orientation.

To operationalize the concept of an ethnic cleavage, we focus on the degree of alignment that exists between ethnicity and party delegations. Where cleavages are extreme, each party represents one and only one ethnic group, a setting we describe as ethnic *segregation*. In a fully *integrated* system, ethnic groups are distributed across parties in the same proportion as across the legislature at-large, or there are no discernible ethnic distinctions among MPs at all. Here, ethnicity appears to play no role in party cleavages. Our approach thus melds group- and party-centered approaches to the measurement of cleavages.[6]

To make these intuitions more concrete, several stylized scenarios are illustrated in Table 1. In each scenario, a legislature is divided between three parties of varying sizes and composition. In Legislature I, only one ethnic group (A) gains entrance into the legislature.[7] In Legislature II, three ethnic groups (A-C) are equally represented across three parties ($P1$-$P3$). In both of these scenarios, there is perfect integration. In Legislature VI, each ethnic group is represented by a different political party. This exemplifies the extreme case of segregation, where all parties are ethnic parties and all ethnic groups are affiliated with a single party.

Note that Legislatures II and VI provide exactly the same parliamentary representation for the three ethnic groups. With respect to descriptive representation at the parliamentary level, one might conclude that these scenarios are equivalent. However, the principle of representation is radically different—cross-party (integrated) in Legislature II and pillarized (segregated) in Legislature VI. We suspect this has important ramifications.

Table 1: *Illustrative scenarios.*

Legislatures (N)	**I** (12)			**II** (12)			**III** (12)			**IV** (12)			**V** (14)			**VI** (12)		
Ethnic groups (N)	**A** (12)			**A** (4)	**B** (4)	**C** (4)	**A** (8)	**B** (4)		**A** (10)	**B** (1)	**C** (1)	**A** (8)	**B** (4)	**C** (2)	**A** (4)	**B** (4)	**C** (4)
Parties (N)	**P1** (4)	**P2** (4)	**P3** (4)	**P1** (4)	**P2** (4)	**P3** (4)	**P1** (6)	**P2** (4)	**P3** (2)	**P1** (5)	**P2** (5)	**P3** (2)	**P1** (4)	**P2** (4)	**P3** (6)	**P1** (4)	**P2** (4)	**P3** (4)
MPs	*a* *a* *a* *a*	*a* *a* *a* *a*	*a* *a* *a* *a*	*a* *b* *c* *d*	*a* *b* *c* *d*	*a* *b* *c* *d*	*a* *a* *a* *a* *b* *b*	*a* *a* *a* *a*	*b* *b*	*a* *a* *a* *a* *a*	*a* *a* *a* *a*	*b* *c*	*a* *a* *a* *a*	*a* *a* *a* *a*	*b* *b* *b* *b* *c* *c*	*a* *a* *a* *a*	*b* *b* *b* *b*	*c* *c* *c* *c*
Cleavage Index	0			0			0.22			0.28			0.44			0.66		

1.2 An Ethnic Cleavage Index

The extreme scenarios laid out in Legislatures I, II, and VI mark two ends of a continuum. In between lie an infinite variety of intermediate settings, a few of which are illustrated by Legislatures III-V.

To provide a sensitive metric that expresses all possible points on this continuum of integration and segregation, we focus on the magnitude of the transformations required for a legislature to

[6] In this respect, we follow Hamza and Maeda (2025), though the latter is focused on ethnic voting rather than elites.
[7] Tellingly, Horowitz (1985) does not provide a definition for a non-ethnic party, a point noted by Elischer (2013).

reach perfect integration. This scale ranges from 0 (where no changes are required) to a value that approaches 1 asymptotically (where nearly everything needs to change in order to achieve integration). We refer to this as an *ethnic cleavage index*.

To put this notion into motion, we begin with the observed distribution of parties by ethnic group in a legislature. We then measure the distance between this observed arrangement and a hypothetical arrangement under perfect integration. In this ideal scenario, the marginal distribution of groups and parties is preserved. Within that constraint, there is a uniform allocation of members across groups.

To quantify distance from perfect integration, we employ techniques from optimal transport theory (Lott and Villani, 2009). This mathematical framework provides a rigorous method for measuring the distance between two probability distributions. It also melds the party- and ethnic-centered approaches to ethnic cleavages introduced at the outset.

In this index, we ask a simple question: what is the smallest fraction of legislative seats that would have to be reassigned across parties—while keeping each MP's ethnicity fixed—to make every party's delegation mirror the chamber's overall ethnic makeup? We cast this as a discrete optimal-transport problem (ibid.). Each cell in the groups,×, parties table is a "location." Leaving an MP in the same cell costs 0; moving an MP to any other cell costs 1. Because the "perfect-integration" target preserves each group's total number of MPs (row sums), any optimal reassignment can be implemented *within* ethnic groups only; conceptually, changing an MP's ethnicity carries an infinite penalty and is therefore ruled out. With this 0–1 cost structure, the transport objective equals the *total variation distance* (half the ℓ_1 difference) between the observed joint distribution and the integration target. The resulting score lies in $[0, 1]$ and has a direct interpretation: it is the minimal *share of seats* that must be reallocated across parties (within groups) to achieve perfect integration.

To get a feel for this way of measuring cleavages, let us return to the examples sketched in Table 1. Scenarios I and II exemplify perfect integration, so these legislatures receive a perfect score of 0 (no segregation) in the bottom row.

Scenario VI exemplifies perfect segregation. Nearly everything—but not quite everything—must change in order to achieve perfect integration across parties. Specifically, one would have to reassign all but one MP per party in order to achieve integration. This is why it is an asymptotic value. In cases of purely ethnic parties, the value of the segregation index approaches 1 as the number of groups becomes larger.

1.3 An Ethnic Party Index

The same logic may be extended to scoring for individual parties. Indeed, the score for a legislature is simply the weighted average of the segregation scores of all the parties in the legislature.

Consider the example illustrated in Table 2. Here, we see a legislature with sixteen MPs, three parties, and two ethnic groups. The majority group, A, comprises 3/4 of the legislature while the minority group, B, comprises 1/4.

Table 2: *Ethnic representation across parties. Groups A and B are represented in varying proportions.*

	P1								*P2*						*P3*	
MPs	*b*	*b*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*a*	*b*	*b*

Perfect integration for a party means that the distribution of ethnicities in its delegation mirrors the distribution of ethnicities in the legislature as a whole. This is the case for Party 1, which receives a score of 0, indicating perfect integration.

Where a party deviates, one must calculate the degree of deviation, which may be understood loosely as the magnitude of changes required to achieve perfect integration.

Although Party 2 is composed exclusively of one ethnic group, A happens to be the largest group in the legislature (by far). Accordingly, it requires only a small change in composition for Party 2 to mirror the legislature, rendering a modest ethnic party score. This exemplifies the situation of "ethno-nationalist" parties such as Fidesz in Hungary, United Russia in Russia, or the BJP in India (Rydgren, 2007).

Party 3 is also composed of a single ethnic group. However, this group composes a small minority of the legislature. As such, its membership must be thoroughly transformed in order to mirror the ethnic characteristics of the legislature. This exemplifies the situation of "ethnic parties," introduced at the outset.

Now, let us consider a legislature incorporating three equal-sized ethnic groups distributed across three political parties, as shown in Table **??**. This setting exemplifies an enduring debate about how to measure ethnic cleavages, as previously discussed. If cleavages are determined by the ethnic purity of each party, Parties 1 and 2 should receive the highest score. If, on the other hand, cleavages are determined on the basis of ethnic voting, one might grant the highest score to groups B and C, whose devotion to Party 3 is complete.

Our approach melds these two perspectives into a single score. In this case, Parties 1 and 2 receive higher segregation scores than Party 3 because a greater transformation of these parties is required in order to achieve perfect integration.

However, in another scenario — where group A is much larger than groups B and C — Parties 1 and 2 would lie closer to integration than Party 3, which would require a more thoroughgoing transformation to reach perfect integration. The general point is that the ethnic segregation of a party can be evaluated only in terms of the distribution of ethnicities across the entire legislature. Where the latter changes, the former is bound to change.

Another general point to note is that as the number of parties shrinks, the expected segregation of parties and the legislature at-large also shrinks, approaching zero in the case of a single-party system. Where the number of parties = 1, segregation = 0 for the party and for the legislature, by construction. Likewise, as the number of ethnic groups shrinks, so does the expected segregation score of political parties and the legislature. Where the number of ethnic groups = 1, segregation = 0 for parties and for the legislature, by construction. These compositional effects play a crucial role in our explanatory framework, as outlined in Section 4.

1.4 An Ethnic Cohesion Index

The third elaboration of social cleavages concerns the extent to which members of an ethnic group are concentrated in a single party or dispersed across multiple parties. To fix ideas, Table 3 explores three ethnic groups, each with three MPs. Group A is concentrated on one party. Group B is spread across two parties. And Group C is dispersed across all three parties.

To provide a precise measure of cohesion, we again leverage optimal transport, now from the group perspective. For a given ethnic group with observed distribution across parties, we compute the distance to a reference distribution where its members are allocated proportionally to overall party seat shares—a counterfactual of perfect dispersion. Using the 0-1 cost structure (0 to stay, 1 to switch parties), this distance equals the total variation between the conditional party distribution

Table 3: *Ethnic representation across parties. Groups A, B, and C shown seat-by-seat.*

	P1						*P2*		*P3*
MPs	*a*	*a*	*a*	*b*	*b*	*c*	*b*	*c*	*c*

for the group and the marginal party distribution, interpretable as the minimal share of the group's MPs requiring reassignment to achieve dispersion.

2 Data Collection and Coding

Having laid out our approach to measuring ethnic cleavages, ethnic parties, and ethnic cohesion, we turn to the empirical study of these matters. Data collection and coding is much more than an afterthought as it involves solving difficult challenges of conceptualization and measurement.

We begin with ethnicity, about which so much has been written (Abdelal et al., 2009; Marquardt and Herrera, 2015). For present purposes, this concept encompasses any ascriptive identity, i.e., any set of group characteristics understood as inherited, including customs, language, religion, race, region, and various combinations of the foregoing. We reserve the word "ethnic" for the most salient of these dimensions as understood in a particular society at a particular point in time. This recognizes the constructed nature of ethnicity as well as its capacity for change.

Since our intention is to assess how ethnic groups are represented, not whether they are represented (at all), our purview is limited to groups that gain some political representation at national levels. Very small ethnic groups or groups that face intense discrimination or are denied citizenship are likely to be excluded. This is in keeping with common understandings of the concept of a political *cleavage*, which refers to a relationship between political parties and constituencies that are allowed to participate and to attain public office. A racial political cleavage in the American South appears only when blacks were enfranchised, for example.

Likewise, we do not consider whether the representation of groups is proportional to their population. This is, of course, an important issue (Gerring, Jerzak, and Öncel, 2024), but it is orthogonal to the concept of cleavage structures. A party whose leaders are drawn exclusively from a single ethnic group is no more or less ethnic if that group is over- or under-represented.

2.1 Expert Coding

To code the ethnicity of MPs, we rely primarily on country experts enlisted for the Global Leadership Project (ibid.). Their judgments rest on cues about each MP drawn from names, birthplaces, and photos—often contained in parliamentary websites. Bear in mind that we are interested in how MPs represent themselves to their constituents. For present purposes, the ethnicity of Representative X is whatever X says or implies it to be—their public presentation of self (Goffman, 1959). We do not concern ourselves with whether X is really (authentically) Christian or Muslim, Croat or Serb.

Despite this caveat, we acknowledge that different judgments about how to define ethnic groups in a given country might lead to different conclusions. The same is true for extant codings of ethnic voting and ethnic parties (reviewed above), which depend upon prior judgments about what an ethnic group is and how it should be operationalized in a given context—often, a fraught exercise subject to ongoing revision (Csata, Hlatky, and Liu, 2021). We see no way around this conundrum. Readers may survey the decisions reached by our coders, listed in Appendix ??, and decide for themselves.

2.2 Coding from an AI Search Agent

Expert coding is supplemented with a subset of high-confidence party, ethnicity, and other leader-specific values from an LLM-based agent that analyzes Wikipedia and search engine results, and bases its prediction of the relevant missing value on the search content found.

More specifically, we deploy a retrieval-augmented AI search agent (ASA) to infer missing attributes from publicly available digital sources. The agent operates under a country-specific, closed codebook provided by our experts and may either assign a single label from that set or abstain from doing so.

As visualized in Figure 1, for each MP the agent queries parliamentary portals, official biographies, reputable news, and encyclopedic entries; extracts corroborating statements; and returns a structured prediction (label, one-sentence rationale, and source citations). The agent is instructed to prioritize precision over recall and to abstain when evidence is weak or conflicting sources are present. All queries, snippets, and links are archived to ensure auditability and replication. This provenance-preserving design expands coverage while limiting measurement error and maintaining cross-national comparability.

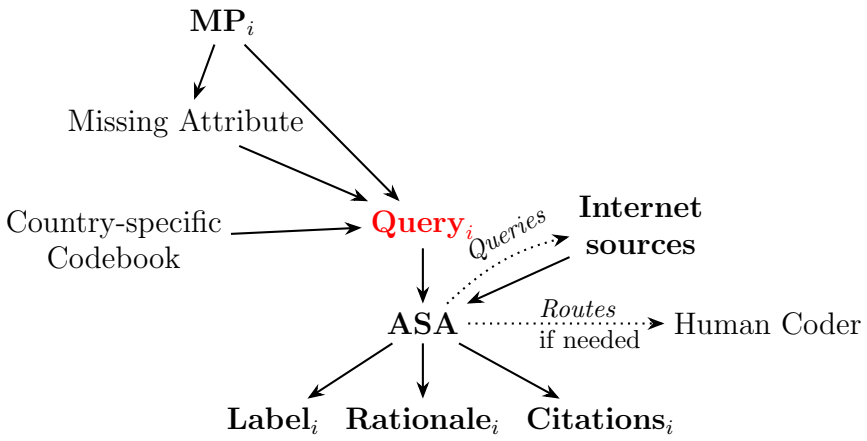

Figure 1: *AI Search Agent (ASA) visualization.*

Agent Performance. To gauge the performance of this protocol we focus on missing party affiliations rather than ethnicity because it is objectively verifiable through public records and exhibits substantially higher rates of missing data, providing a more robust test.

Using the subset of records for which expert codes are available, this verifiable search agent attains high agreement on party labels. Across $N = 114$ countries and $34,618$ matched leader–records covering $1,209$ distinct party labels, the overall high–confidence accuracy is 0.860, with a mean per–country high–confidence accuracy of 0.879 (Figure 2). Relative to a simple country–majority baseline of 0.536, this yields an average uplift of 0.343 points on a 0–1 scale, ranging from -0.602 in Gambia to 0.761 in Finland. Point estimates improve in later election cohorts, consistent with expanding web coverage and corpus quality (Figure 3). If we ignore the confidence gate and score all predictions, overall accuracy is 0.751.

Limitations, however, remain. On average across countries, a share of 0.250 of records is flagged "Low" and withheld from downstream analysis, yielding a conservative working sample. Performance is predictably uneven across label frequency: mean high–confidence accuracy for small/minority party labels is 0.692, compared to 0.820 for plurality parties. In practice, disagree-

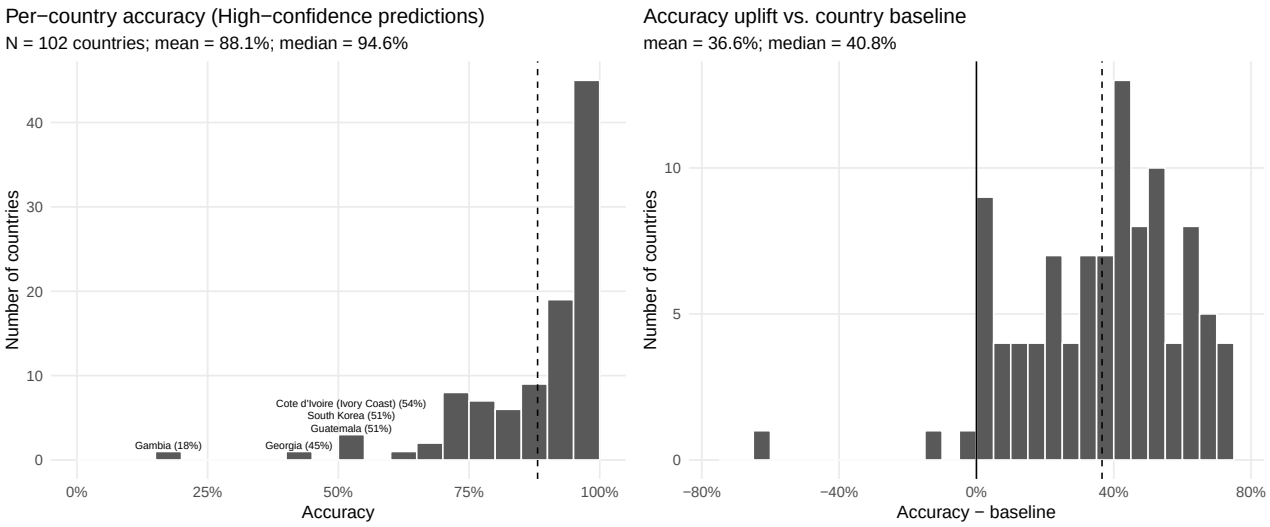

Figure 2: *Agent performance in predicting party, across countries having at least 20 agent codes. Median country-level performance is above 90%. The baseline refers to the accuracy of predicting that every politician in a given country takes on the dominant characteristic being considered.*

ments and abstentions concentrate in (i) very small or newly formed parties, (ii) earlier election years, and (iii) cases with sparse or non-standardized online biographies.

Overall, this provenance-preserving search agent substantially expands usable coverage by 20.8% (12,898 observations with previously missing party labels) at low marginal cost while maintaining high out-of-sample agreement with expert codings (Figure 2). The approach is transparent in that every assigned label is backed by stored queries and citations—portable across countries because predictions are constrained by closed codebooks. Our downstream estimates will employ the high-confidence entries from the agent. For results using only expert coding, see Appendix A; for additional details about the agent architecture, see Appendix B. See Table 7 for a sample of the sources obtained by the agent.

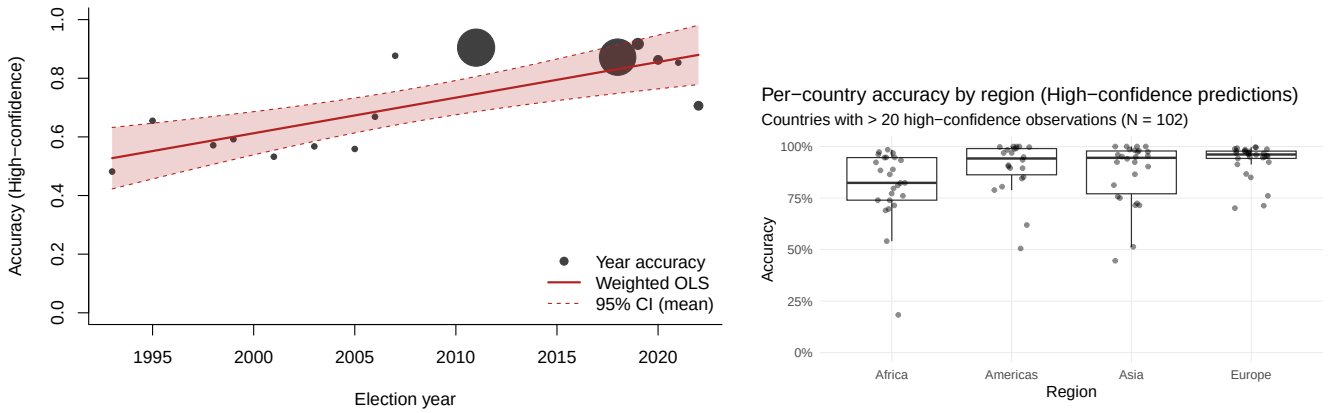

Figure 3: *Agent performance over time; point size corresponds to the number of observations in each (approximate) year (*Left*) and space (*Right*).*

3 Ethnic Cleavages Across the World

Having outlined our protocol for data collection, we turn to a survey of ethnic cleavages across the world. Table 4 offers a list of all countries in our dataset (N=153), ordered by the strength of their ethnic cleavage. Multiple observations for each country are averaged to obtain a single estimate for each country.

In the upper tier of Table 4 are countries where party competition closely tracks ascriptive communities—often settings with institutionalized communal representation, pillarized party histories, and regionally concentrated electorates, e.g., Lebanon, Bosnia and Herzegovina, Malaysia, Belgium, India.

A broad middle comprises varied systems in which explicitly ethnic parties coexist with cross-ethnic, programmatic, or catch-all organizations; this stratum includes several multilingual democracies and diverse federations where regional or linguistic cleavages are politically salient but not uniformly determinative, e.g., Spain, Canada, New Zealand, and parts of sub-Saharan Africa.

At the lower end cluster three sorts of cases: (i) non-competitive or hegemonic-party regimes that prevent the expression of ethnic cleavages (e.g., North Korea, Venezuela); (ii) relatively homogeneous polities where ethnic differences are muted (e.g., Japan, Norway) and (iii) systems whose principal conflicts are ideological rather than ascriptive (e.g., much of Western Europe).

Table 4: *Cleavage index by country.*

Country	Index	Country	Index	Country	Index
Lebanon	0.449 (1)	El Salvador	0.084 (53)	Azerbaijan	0.009 (105)
Bosnia & Herz.	0.435 (2)	Lithuania	0.084 (53)	Burundi	0.009 (105)
Malaysia	0.364 (3)	Guyana	0.077 (54)	Nigeria	0.009 (105)
Guinea	0.359 (4)	Slovakia	0.072 (55)	Fiji	0.008 (108)
Guinea-Bissau	0.354 (5)	Djibouti	0.070 (56)	Austria	0.008 (108)
Peru	0.351 (6)	Burkina Faso	0.063 (57)	Italy	0.008 (108)
Benin	0.345 (7)	Morocco	0.060 (58)	Zimbabwe	0.007 (111)
Solomon Is.	0.339 (8)	Turkey	0.058 (59)	Luxembourg	0.007 (111)
Belgium	0.306 (9)	Senegal	0.054 (60)	Armenia	0.007 (111)
India	0.287 (10)	Gabon	0.053 (62)	Venezuela	0.006 (112)
Latvia	0.283 (11)	Jamaica	0.053 (62)	Puerto Rico	0.005 (117)
Mali	0.282 (12)	Ecuador	0.052 (63)	Greece	0.005 (117)
Pakistan	0.279 (13)	Kosovo	0.048 (64)	Cambodia	0.005 (117)
Niger	0.277 (14)	Romania	0.047 (65)	Kyrgyzstan	0.005 (117)
Congo	0.273 (15)	Czech Rep.	0.045 (66)	Cape Verde	0.005 (117)
Netherlands	0.232 (17)	Taiwan	0.044 (67)	France	0.004 (119)
Moldova	0.232 (17)	Serbia	0.043 (68)	Poland	0.004 (119)
Mauritius	0.231 (18)	Singapore	0.042 (69)	Tajikistan	0.003 (121)
Afghanistan	0.227 (19)	Thailand	0.038 (72)	Chile	0.003 (121)
Congo	0.222 (21)	Liberia	0.038 (72)	Yemen	0.002 (126)
Haiti	0.222 (21)	Madagascar	0.038 (72)	Belarus	0.002 (126)
Macedonia	0.191 (22)	Philippines	0.036 (73)	Uruguay	0.002 (126)
Bolivia	0.184 (23)	Botswana	0.035 (75)	Egypt	0.002 (126)
Myanmar	0.182 (24)	Uzbekistan	0.035 (75)	Honduras	0.002 (126)
Kenya	0.172 (25)	Dom. Rep.	0.034 (77)	Portugal	0.001 (130)
Ethiopia	0.166 (26)	Cuba	0.034 (77)	Albania	0.001 (130)
Iraq	0.149 (27)	Sweden	0.032 (78)	Tunisia	0.001 (130)
Algeria	0.146 (28)	Mongolia	0.030 (79)	Bangladesh	0.001 (130)
Namibia	0.145 (29)	Iran	0.029 (83)	Bahrain	0.000 (153)
Israel	0.134 (30)	Cote d'Ivoire	0.029 (83)	Costa Rica	0.000 (153)
Sierra Leone	0.131 (32)	Kazakhstan	0.029 (83)	Cyprus	0.000 (153)
Ukraine	0.131 (32)	Nicaragua	0.029 (83)	Iceland	0.000 (153)
Zambia	0.128 (33)	United Kingdom	0.027 (87)	Japan	0.000 (153)
Indonesia	0.127 (34)	United States of America	0.027 (87)	Jordan	0.000 (153)
C. Afr. Rep.	0.126 (35)	Denmark	0.027 (87)	Korea	0.000 (153)
S. Africa	0.125 (36)	Germany	0.027 (87)	Kuwait	0.000 (153)
Spain	0.124 (37)	China	0.024 (88)	Lesotho	0.000 (153)
Montenegro	0.121 (38)	Hungary	0.023 (89)	Libya	0.000 (153)
Timor-Leste	0.120 (39)	Finland	0.022 (91)	Mauritania	0.000 (153)
New Zealand	0.119 (41)	Croatia	0.022 (91)	Norway	0.000 (153)
Panama	0.119 (41)	Georgia	0.020 (93)	Oman	0.000 (153)
Paraguay	0.114 (42)	Viet Nam	0.020 (93)	Palestinian Territory	0.000 (153)
Gambia	0.109 (43)	Turkmenistan	0.019 (94)	Qatar	0.000 (153)
Guatemala	0.104 (44)	Ireland	0.018 (96)	Saudi Arabia	0.000 (153)
Ghana	0.100 (45)	Mexico	0.018 (96)	Somalia	0.000 (153)
Rwanda	0.099 (47)	Brazil	0.017 (97)	Somaliland	0.000 (153)
Russia	0.099 (47)	Colombia	0.014 (98)	South Korea	0.000 (153)
Switzerland	0.098 (48)	Malta	0.013 (99)	Sri Lanka	0.000 (153)
Canada	0.095 (49)	Slovenia	0.012 (101)	Suriname	0.000 (153)
Trin. & Tob.	0.089 (50)	Estonia	0.012 (101)	Syria	0.000 (153)
Malawi	0.088 (51)	Bulgaria	0.011 (102)	United Arab Emirates	0.000 (153)

Figure 4 displays a histogram of the data, where each data point represents a national legislature. Here, we see a markedly right-skewed distribution: most legislatures cluster near zero (weak ethnic segregation), with a long tail comprised of more segmented systems.

Box plots in the left panel of Figure 4 focus on different regions of the world. Here, we find minimal differences across medians. However, Asia and especially Africa have considerably higher variability, conforming to standard views in the literature (cited above).

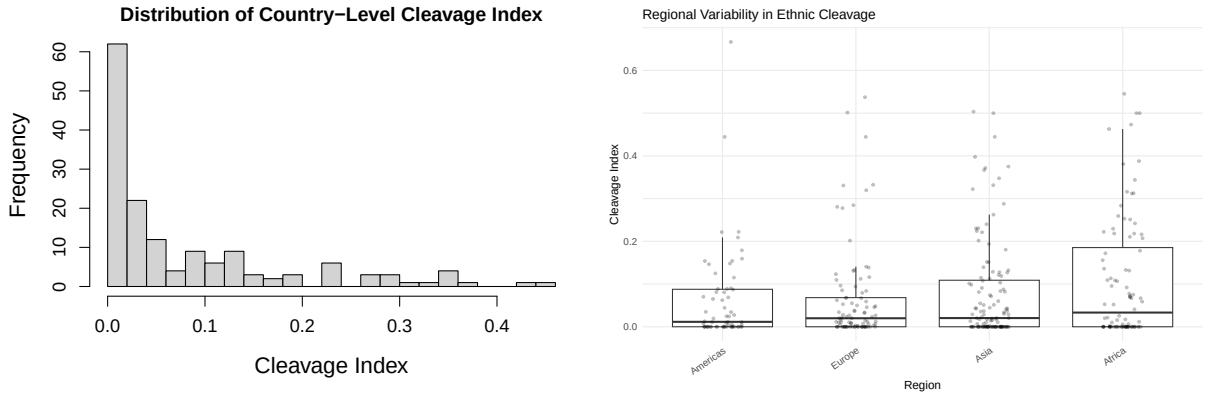

Figure 4: LEFT. *Distribution of the cleavage index across country-elections in the sample. The index is bounded between 0 (perfect integration) and 1 (perfect segregation); the distribution is heavily skewed to the right, indicating that most national legislatures in the dataset exhibit low levels of ethnic segregation among their party delegations.* RIGHT. *Distribution by region.*

Figure 5 shows variability over time in ethnic cleavages for a handful of countries for which we have more than two data points. (Typically, each observation observes the status of a legislature in between elections, so four data points might be separated by three elections.) We can see that ethnic cleavages have increased in New Zealand and Panama, decreased in the Russian Federation, and remained relatively flat in Thailand and Ukraine over the period of observation. Across the entire sample, the mean absolute change in cleavage index from one period (election) to the next is 0.078, signaling considerable stability.

3.1 Ethnic Parties and Ethnic Cohesion

Recall that our approach to measurement renders an overall index of ethnic cleavages (presented above) as well as indices focused on the degree to which parties are ethnically defined and the degree to which ethnic groups are politically defined (cohesive).

Table ?? shows descriptive statistics for these three indices. Figure ?? shows histograms of the data for these indices. [Perhaps move the histogram for ethnic cleavage from Figure 4 to here so they can be shown together...]

[Comment on the mechanical relationship between the three indices. How much of the variability in the ethnic party and ethnic cohesion indices is within-country?]

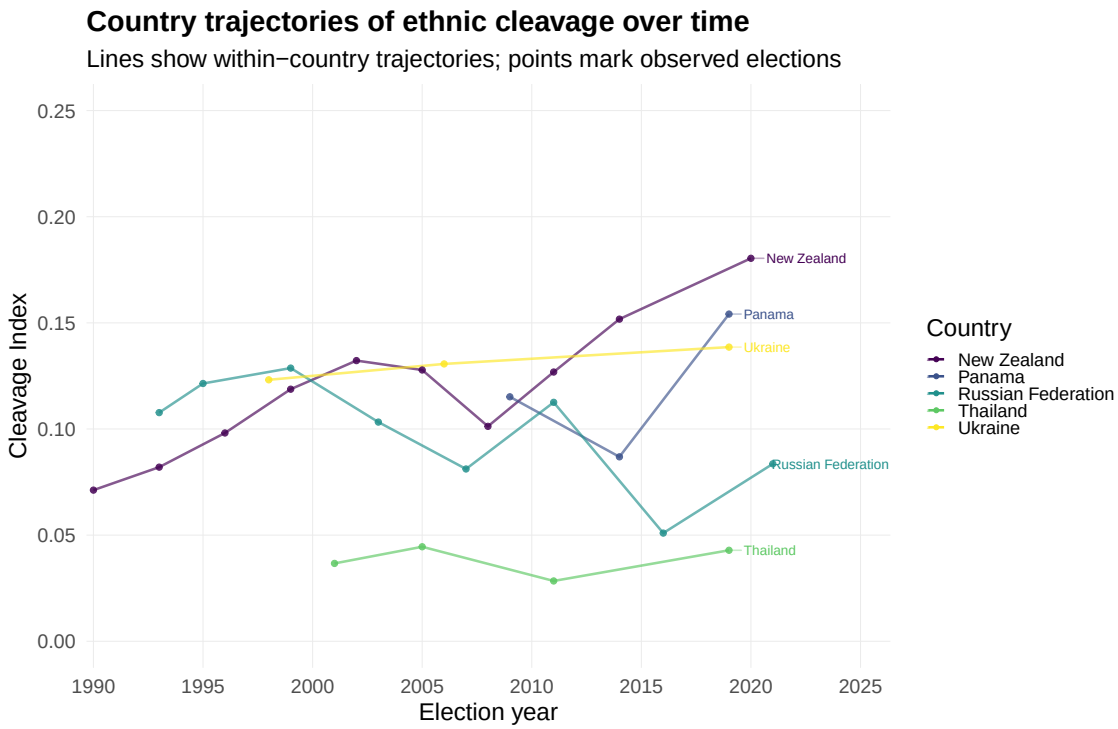

Figure 5: *Over time variability in cleavage structures.*

3.2 Convergent Validity

In this section, we compare our indices with extant work. As discussed, most studies judge the ethnic character of politics primarily on the basis of mass political behavior, i.e., on patterns of ethnic voting. Since our measure rests on the ethnic identity of elites, we are comparing two angles on the same general subject. We expect to find some degree of concordance, but not perfect concordance. Where correlations are weak, this may indicate measurement error or perhaps theoretically interesting disjunctures between politics at mass and elite levels.

To make these comparisons, we draw on studies of ethnic voting and ethnic parties with the broadest coverage, e.g., Houle, Park, and Kenny (2019), Fraenkel and Crawley (2025), Lublin (2014). We may also enlist the V-Party dataset Düpont et al. (2022), where the orientation of individual parties is coded, allowing for a direct comparison with our Ethnic party index. (Lublin's dataset may also allow for this sort of party-centered comparison.)

4 Explanations

With descriptive patterns in hand, we turn from measurement to explanation. Recall that an ethnic cleavage is defined here as the degree of alignment between ethnic groups and party delegations. This means that there is a mechanical connection to the distribution of ethnic groups and political parties. At the lower limit, where there is only one ethnic group or one political party, there can be no ethnic cleavages. As ethnic groups and political parties multiply, the space available for ethnic cleavages expands. Following previous work, we refer to this as a "compositional" relationship since it depends upon the fit between sociological and institutional configurations (Gerring, Jerzak, and Öncel, 2024).

To demonstrate the compositional elements of this puzzle,...[Place simulations here.]

To test these relationships with our data, we focus on the dispersion of groups and parties. Dispersion may be operationalized in a variety of ways, e.g., by a Herfindahl index, an entropy index, a largest share index (the size of the largest group or party), and so forth. All are highly correlated and, unsurprisingly, all produce very similar results. We adopt the Herfindahl index because of its familiarity and ease of interpretation.

In Model 1, we test the relationship of our ethnic cleavage index to the degree of ethnic and party fractionalization in each election across the entire sample, clustering standard errors by country. In Model 2, we interact the two variables. In Model 3 we add two-way fixed effects. It will be seen that over half of the variability in ethnic cleavage structures is explained by the first model and nearly three-quarters by the second and third models, which incorporate the interaction term.

In Model 4, we add a variety of additional predictors including the Polyarchy index of electoral democracy (Teorell et al., 2019), per capita GDP, PPP, log (World Bank, 2024), urbanization (ibid.), population, log (ibid.), state history (Borcan, Olsson, and Putterman, 2018), electoral system rules (Coppedge et al., 2025). We also include V-Dem indicators measuring the distribution of resources — power, services, state jobs, and business opportunities — across social groups (ibid.), understood as a measure of social discrimination. This analysis is replicated in a LASSO regression (TBA), which shows the same general results.

None of these demographic, sociological, and economic factors does a very good job of predicting ethnic cleavages, and the total model-fit is scarcely improved. Accordingly, we conclude that most of the variability in ethnic cleavages is explained by the intersection of demography and party systems, and is therefore largely compositional, as theorized. This leaves several questions dangling.

The first concerns the interaction between ethnic and party fractionalization. Extant work sug-

	Model 1	Model 2	Model 3	Model 4
Party Fractionalization	0.16 (8.54)*	0.01 (1.12)	0.02 (0.66)	0.02 (1.56)
Group Fractionalization	0.25 (11.60)*	0.00 (0.14)	0.03 (0.58)	0.00 (-0.04)
Party Times Group Fractionalization		0.65 (14.03)*	0.60 (5.41)*	0.64 (11.63)*
Polyarchy Index				0.03 (1.26)
Power Distributed by Social Group				-0.01 (-1.64)
Access to Public Services by Social Group				0.00 (0.07)
Access to State Jobs by Social Group				0.00 (-0.13)
Access to State Business Opportunities by Social Group				0.00 (-0.24)
log(Population)				0.00 (-1.77)
log(GDP per capita (PPP))				-0.01 (-1.59)
Country, Percent of Pop. Urbanized				0.00 (1.01)
Country FE			✓	
Year FE			✓	
Other statistics				
Countries	153	153	153	139
Observations	372	372	316	288
Adjusted R-squared	0.54	0.74	0.73	0.73

Table 5: *Outcome: Cleavage Index. Estimator: OLS with clustered standard errors by country.* $*$ *indicates* $p < 0.05$*; t-statistics are in parentheses.*

gests that identities influence party affiliations (Ordeshook and Shvetsova, 1994) and party politics influences identities (Eifert, Miguel, and Posner, 2010). The two are very weakly correlated (Pearson's $R = 0.17$), so there is little risk of collinearity. However, we cannot discount the possibility that they are mutually constituted – a feature that our interaction model (Model 2) may help to account for.

Second, there are questions about deeper causes, i.e., what lies prior to ethnic and party fractionalization. One might speculate about modernization, regime type, electoral rules, geography, and various other structural factors. However, initial tests show that these factors, collectively, explain only a very small portion of the variance once the two proximal factors are removed (R^2 <0.10).

A final question concerns the variance remaining in our benchmark model. Evidently, there is a fair bit of variation in ethnic cleavages among polities with multiple (sizeable) ethnic groups and political parties. [Let's look for most-similar cases - where ethnic and party fract are similar but cleavage index scores are different.]

5 Do Cleavages Matter?

Before concluding, it seems important to return to a question raised briefly at the outset. Do elite-level ethnic cleavages matter?

As noted, there is considerable disagreement among scholars about the role of ethnic diversity in outcomes such as civil conflict, corruption, democracy, and growth. Many recent studies cast doubt on the dangers of diversity. This could be because extant studies usually focus on demographic characteristics or on voting behavior.

It is often observed that latent cleavages become problematic only when politicized. So long as ethnic groups regard their separate identities as a background feature, political cleavages are likely to form along economic or ideological lines. These may be less fractious and more susceptible to compromise.

However, when ethnic identities are politicized, governance often becomes more difficult, perhaps even impossible. Even rich countries with long democratic histories, such as Belgium, may experience the strain of ethnically based political cleavages (De Winter and Baudewyns, 2013).

Our elite-level measure of ethnic cleavage, resting on the character of party delegations in the national legislature, may represent the most politically relevant expression of ethnicity. Here, we might have strong reasons to suppose that heightened ethnic cleavages would be associated with worse governance performance.

At present, our data are unable to test the causal direction of this relationship in a rigorous fashion. Our goal, therefore, is to explore correlations. If those correlations are fairly strong, we may be justified in concluding that ethnic cleavages are either a cause or a consequence of undesirable outcomes. If the correlations are weak or in unexpected directions, we may be justified in concluding that ethnic cleavages, as measured here, are irrelevant for the overall quality of governance in a country.

[For conflict, see UCDP. For conflict coded as ethnic see ETHNIC ARMED CONFLICT, VERSION 3, www.awimmer.com/data. For corruption see CPI and V-Dem. Limit analysis to the past two decades.]

6 Conclusion

Our analysis advances the study of ethnic politics on three fronts. Conceptually, we recast elite cleavages as a continuous alignment problem and show that an optimal-transport formulation yields a single, bounded, and interpretable scale—the minimal share of seats that would have to be reassigned across parties (holding MPs' identities fixed) to achieve perfect integration. The same logic provides party-level and group-level diagnostics, allowing researchers to see not only *whether* a system is segmented but also *where* segmentation resides.

Empirically, we pair this measure with a global dataset of parliamentary delegations assembled from expert codings and an auditable retrieval-augmented search agent that privileges precision and preserves provenance. This pipeline recovers canonical cases in the literature while revealing fine-grained within-country and within-party variation that binary typologies and survey coverage gaps obscure.

Substantively, we find robust evidence for a simple demand–supply account: elite ethnic segregation rises with underlying social heterogeneity and is amplified—rather than created—by party fragmentation; institutional moderators behave in broadly intuitive ways but are less stable than the demand fundamentals. Taken together, these contributions furnish a replicable baseline for cross-national and over-time comparisons, enable targeted evaluation of electoral reforms and power-sharing arrangements, and open a path to cumulative tests linking elite segmentation to downstream outcomes—accountability, clientelism, public-goods provision, and conflict.

The AI search agent approach outlined here is also cost-effective, and in that sense, has the potential to help democratize research on political elites. While one round of the original GLP coding entailed costs of upwards of $50,000 due to payments towards RAs and expert coders, one round of data here can be coded for under $50 USD (2025 dollars) per feature, representing a hundred-fold decrease in cost for a theoretically targeted dataset of, e.g., 10 variables. Looking

forward, one might hope that these logarithmic reductions in the cost of obtaining high-quality data on political elites will accelerate the pace of discovery in this field and enable researchers from smaller institutions, developing countries, and those with limited funding to conduct comprehensive studies that were previously accessible only to well-funded research teams.

Limitations remain—most notably the definitional choices that any coding of ethnicity entails and our focus on represented groups—but the quantitative framework is flexible: transport costs in the cleavage measure can incorporate geography, multi-membership identities can be accommodated, and the design extends naturally to other ascriptive cleavages (religion, language, region) and other arenas (cabinets, local councils). □

References

Abdelal, Rawi et al., eds. (2009): *Measuring Identity: A Guide for Social Scientists*. Cambridge: Cambridge University Press.

Adida, Claire et al. (2017): "Overcoming or reinforcing coethnic preferences? an experiment on information and ethnic voting". In: *Quarterly Journal of Political Science*, no. 4, vol. 12, pp. 437–477.

Alesina, Alberto and Eliana La Ferrara (2005): "Ethnic diversity and economic performance". In: *Journal of economic literature*, no. 3, vol. 43, pp. 762–800.

Alford, Robert R (1962): "A suggested index of the association of social class and voting". In: *Public Opinion Quarterly*, pp. 417–425.

Arriola, Leonardo R (2012): *Multi-ethnic coalitions in Africa: Business financing of opposition election campaigns*. Cambridge University Press.

Berman, Bruce, Dickson Eyoh, and Will Kymlicka (2004): *Ethnicity and democracy in Africa*. Ohio University Press.

Birnir, Jóhanna Kristín (2006): *Ethnicity and electoral politics*. Cambridge University Press.

Borcan, Oana, Ola Olsson, and Louis Putterman (2018): "State history and economic development: evidence from six millennia". In: *Journal of Economic Growth*, no. 1, vol. 23, pp. 1–40.

Chandra, Kanchan (2005): "Ethnic parties and democratic stability". In: *Perspectives on politics*, no. 2, vol. 3, pp. 235–252.

— (2007): *Why ethnic parties succeed: Patronage and ethnic head counts in India*. Cambridge University Press.

— (2011): "What is an ethnic party?" In: *Party Politics*, no. 2, vol. 17, pp. 151–169.

Converse, Philip E (1958): "The shifting role of class in political attitudes and behavior". In: *Readings in social psychology*, vol. 3.

Coppedge, Michael et al. (2025): *V-dem codebook v15*. Tech. rep. Varieties of Democracy Project.

Csata, Zsombor, Roman Hlatky, and Amy H Liu (2021): "How to head count ethnic minorities: validity of census surveys versus other identification strategies". In: *East European Politics*, no. 3, vol. 37, pp. 572–592.

De Winter, Lieven and Pierre Baudewyns (2013): "Belgium: Towards the breakdown of a nation-state in the heart of Europe?" In: *Pathways from Ethnic Conflict*. Routledge, pp. 20–44.

Dixit, Avinash and John Londregan (1996): "The determinants of success of special interests in redistributive politics". In: *the Journal of Politics*, no. 4, vol. 58, pp. 1132–1155.

Düpont, Nils et al. (2022): "A global perspective on party organizations. Validating the varieties of party identity and organization dataset (V-Party)". In: *Electoral Studies*, vol. 75, p. 102423.

Easterly, William and Ross Levine (1997): "Africa's growth tragedy: policies and ethnic divisions". In: *The quarterly journal of economics*, pp. 1203–1250.

Eifert, Benn, Edward Miguel, and Daniel N Posner (2010): "Political competition and ethnic identification in Africa". In: *American journal of political science*, no. 2, vol. 54, pp. 494–510.

Elischer, Sebastian (2013): *Political parties in Africa: Ethnicity and party formation*. Cambridge University Press.

Fraenkel, Jon and Sam Crawley (2025): "Does Ethnic Fractionalization Lead to Polarization? Introducing a new Dataset on Communal Party Affiliations". In: *Ethnopolitics*, pp. 1–23.

Geertz, Clifford (1963): "The integrative revolution: primordial sentiments and civil politics in the new states." In: *Old societies and new states: The quest for modernity in Asia and Africa*, pp. 105–157.

Gerring, John, Connor T Jerzak, and Erzen Öncel (2024): "The composition of descriptive representation". In: *American Political Science Review*, no. 2, vol. 118, pp. 784–801.
Goffman, Erving (1959): *The presentation of self in everyday life*. Doubleday.
Gould, Lewis (2007): *Grand Old Party: A History of the Republicans*. Random House.
Hamza, Ahmed and Ko Maeda (2025): "Beyond binary: A new measure of party ethnicization". In: *Party Politics*, no. 0, vol. 0, p. 13540688251376485. DOI: 10.1177/13540688251376485.
Horowitz, Donald L (1985): *Ethnic groups in conflict*. Univ of California Press.
Houle, Christian (2018): "Does ethnic voting harm democracy?" In: *Democratization*, no. 5, vol. 25, pp. 824–842.
Houle, Christian, Chunho Park, and Paul D Kenny (2019): "The structure of ethnic inequality and ethnic voting". In: *The Journal of Politics*, no. 1, vol. 81, pp. 187–200.
Huber, John D (2012): "Measuring Ethnic Voting: Do Proportional Electoral Laws Politicize Ethnicity?" In: *American Journal of Political Science*, no. 4, vol. 56, pp. 986–1001.
Ishiyama, John (2009): "Do ethnic parties promote minority ethnic conflict?" In: *Nationalism and ethnic politics*, no. 1, vol. 15, pp. 56–83.
— (2012): "Explaining ethnic bloc voting in Africa". In: *Democratization*, no. 4, vol. 19, pp. 761–788.
Kitschelt, Herbert (2007): *Patrons, Clients and Policies: Patterns of Democratic Accountability and Political Competition*. Cambridge University Press.
Layman, Geoffrey (2001): *The great divide: Religious and cultural conflict in American party politics*. Columbia University Press.
Lemarchand, René (1972): "Political clientelism and ethnicity in tropical africa:* competing solidarities in nation-building". In: *American political science review*, no. 1, vol. 66, pp. 68–90.
Lijphart, Arend (1969): "Consociational democracy". In: *World politics*, no. 2, vol. 21, pp. 207–225.
— (1977): *Democracy in plural societies: A comparative exploration*. Yale University Press.
Lipset, Seymour Martin and Stein Rokkan (1967): *Party systems and voter alignments: cross-national perspectives*. International yearbook of political behavior research.
Liu, Amy H and Jacob I Ricks (2022): *Ethnicity and politics in Southeast Asia*. Cambridge University Press.
Lott, John and Cédric Villani (2009): "Ricci curvature for metric-measure spaces via optimal transport". In: *Annals of Mathematics*, pp. 903–991.
Lublin, David (2014): *Minority Rules: Electoral Systems, Decentralization, and Ethnoregional Party Success*. Oxford University Press.
Madrid, Raúl L (2012): *The rise of ethnic politics in Latin America*. Cambridge University Press.
Marks, Gary et al. (2023): "The social bases of political parties: A new measure and survey". In: *British Journal of Political Science*, no. 1, vol. 53, pp. 249–260.
Marquardt, Kyle L and Yoshiko M Herrera (2015): "Ethnicity as a variable: an assessment of measures and data sets of ethnicity and related identities". In: *Social Science Quarterly*, no. 3, vol. 96, pp. 689–716.
Ordeshook, Peter C and Olga V Shvetsova (1994): "Ethnic heterogeneity, district magnitude, and the number of parties". In: *American journal of political science*, pp. 100–123.
Page, Scott (2008): *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press.
Posner, Daniel N (2004): "Measuring ethnic fractionalization in Africa". In: *American journal of political science*, no. 4, vol. 48, pp. 849–863.

Rabushka, Alvin and Kenneth A Shepsle (1972): *Politics in plural societies*. Charles E. Merrill.
Rae, Douglas W. and Michael Taylor (1970): *The analysis of political cleavages*. Yale University Press.
Reilly, Benjamin (2001): *Democracy in Divided Societies: Electoral Engineering for Conflict Management*. Cambridge University Press.
— (2021): "Cross-ethnic voting: An index of centripetal electoral systems". In: *Government and Opposition*, no. 3, vol. 56, pp. 465–484.
Rustow, Dankwart A (1970): "Transitions to democracy: Toward a dynamic model". In: *Comparative politics*, no. 3, vol. 2, pp. 338–63.
Rydgren, Jens (2007): "The sociology of the radical right". In: *Annu. Rev. Sociol.* No. 1, vol. 33, pp. 241–262.
Silbey, Joel H and Allan G Bogue (1978): *The history of American electoral behavior*. Princeton University Press.
Strijbis, Oliver and Michal Kotnarowski (2015): "Measuring the electoral mobilization of ethnic parties: Towards comparable indicators". In: *Party Politics*, no. 3, vol. 21, pp. 456–469.
Teorell, Jan et al. (2019): "Measuring polyarchy across the globe, 1900–2017". In: *Studies in Comparative International Development*, vol. 54, pp. 71–95.
Van Cott, Donna Lee (2007): *From movements to parties in Latin America: The evolution of ethnic politics*. Cambridge University Press.
World Bank (2024): *World development indicators 2024*. International Economics Department. Development Data Group and World Bank.

7 Appendix A

7.1 Results, Expert Only

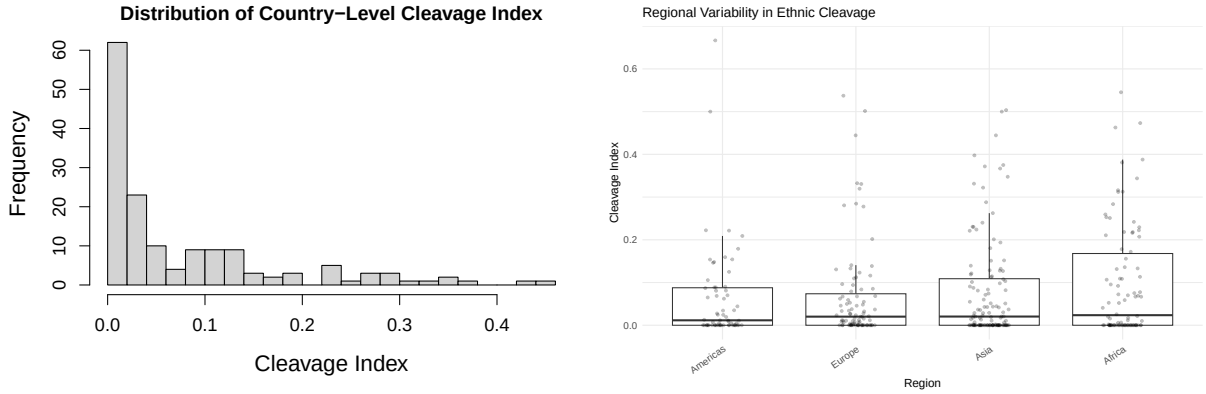

Figure 6: LEFT. *Distribution of the cleavage index across country-elections in the sample. The index is bounded between 0 (perfect integration) and 1 (perfect segregation); the distribution is heavily skewed to the right, indicating that most national legislatures in the dataset exhibit low levels of ethnic segregation among their party delegations.* RIGHT. *Distribution by region.*

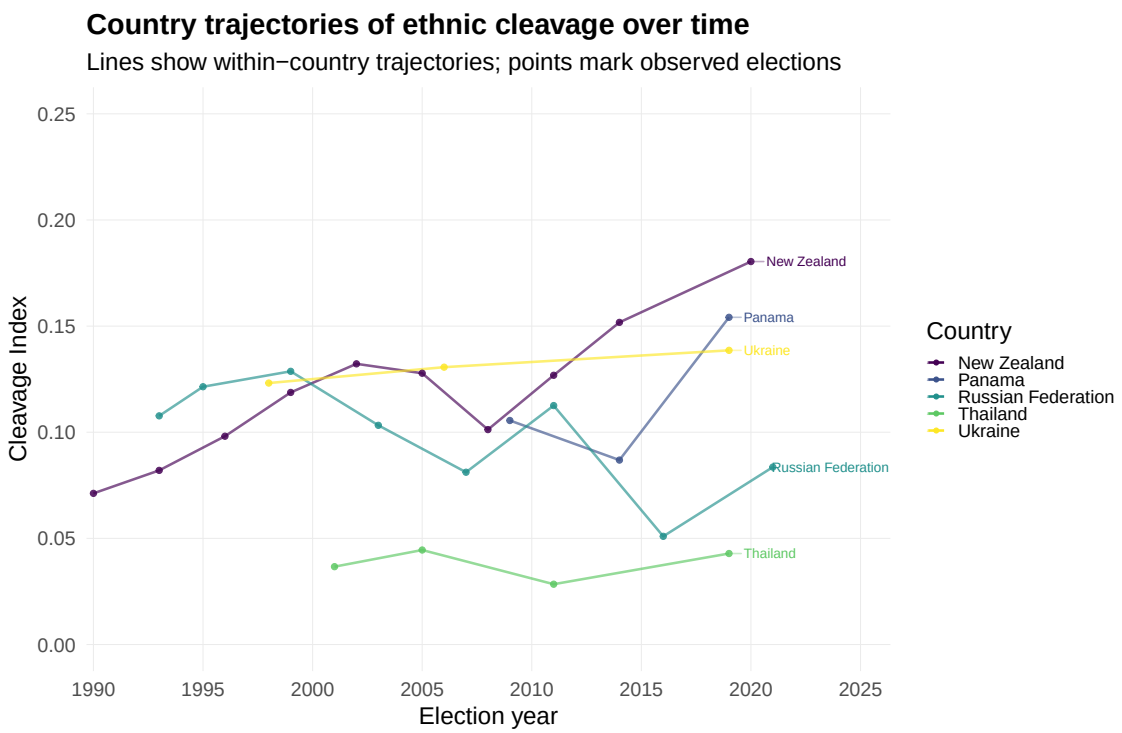

Figure 7: *Over time variability in cleavage structures.*

	Model 1	Model 2	Model 3	Model 4
Party Fractionalization	0.16 (8.75)*	0.01 (1.07)	-0.01 (-0.25)	0.02 (1.29)
Group Fractionalization	0.24 (11.74)*	0.00 (-0.04)	0.07 (1.32)	0.00 (0.00)
Party Times Group Fractionalization		0.65 (14.07)*	0.60 (5.75)*	0.64 (11.97)*
Polyarchy Index				0.03 (1.06)
Power Distributed by Social Group				-0.01 (-1.29)
Access to Public Services by Social Group				0.00 (-0.21)
Access to State Jobs by Social Group				0.00 (-0.06)
Access to State Business Opportunities by Social Group				0.00 (-0.14)
log(Population)				0.00 (-1.45)
log(GDP per capita (PPP))				-0.01 (-1.14)
Country, Percent of Pop. Urbanized				0.00 (1.00)
Country FE			✓	
Year FE			✓	
Other statistics				
Countries	153	153	153	139
Observations	368	368	312	284
Adjusted R-squared	0.55	0.75	0.77	0.74

Table 6: *Outcome: Cleavage Index. Estimator: OLS with clustered standard errors by country.* $*$ *indicates* $p < 0.05$*; t-statistics are in parentheses.*

8 Appendix B

8.1 AI Agent Details

Design goals. The verifiable AI search agent (ASA) is engineered to expand coverage of missing labels while preserving cross-national comparability and auditability. Three principles guide the design: (i) *precision over recall* via conservative prediction and a confidence gate that withholds uncertain cases; (ii) *closed-world classification* against country-specific codebooks supplied by experts; and (iii) *provenance preservation*, archiving all queries, snippets, and Internet links for replication.

System architecture. The ASA is a retrieval-augmented ReAct agent implemented with `langgraph`. A small instruction-tuned LLM (GPT5-nano, `temperature` = 1) orchestrates tool use and produces a one-shot JSON output containing a label, a one-sentence rationale with citations, and a confidence flag (High/Medium/Low).

The agent has access to two read-only connectors: a Wikipedia retriever and a general web search tool (DuckDuckGo). All decisions are constrained to a *closed codebook* (per country) delivered in-prompt; the model is explicitly instructed to abstain from guessing and to prioritize verifiable sources over known or implicit heuristics.

Data flow and reproducibility. Input candidate records (GLP leaders) are materialized in `SQLite`. For each row, the agent executes a short search session and returns a structured prediction. Results–including raw tool traces—are appended to a `SQLite` store. Post-hoc analysis, scoring, and figure generation are performed in `R`.

Country-closed matching and normalization. To guard comparability across countries and reduce typographic drift, the pipeline applies a two-stage, codebook-guided normalization to the model's raw label string:

1. **Strict closed-set match (country scope).** If the predicted string exactly matches a member of the country's codebook, it is accepted.

2. **Conservative fuzzy match.** Otherwise, a relaxed mapper computes a similarity score s combining (a) Jaro–Winkler similarity on a normalized label, (b) token overlap coverage, and (c) acronym equality. Let m denote the runner-up score. Accept as a match if if

$$s \geq 0.92 \quad \text{or} \quad \Big(s \geq 0.85 \;\&\; s - m \geq 0.08\Big).$$

This mapping process produces a normalized label used for evaluation and aggregation while preserving the original string for audit. It corrects innocuous variants (pluralization, punctuation, acronyms) without introducing new classes. The mapper's conservatism is intentional: tightening thresholds reduces false positives at the cost of additional abstentions.

Confidence gating and abstention policy. The agent emits a categorical confidence estimate; records flagged `Low` or `Medium` are withheld from downstream analyses. This implements an *abstention* layer that trades coverage for precision and is especially important where web evidence is sparse, parties are newly formed, or transliterations vary. As documented in the main text, low-confidence rates vary by country and cohort and are higher for small/minority parties, earlier periods, and cases lacking standardized biographies.

Task prompt. For the party prediction task, see below for the prompt sent to the AI search agent:

```
TASK OVERVIEW:
You are an advanced search-enabled Large Language Model (LLM) specializing in party affiliation inference.
Your role is to determine the most likely political party of well-known political leaders or other notable individuals,
strictly from a provided list of possible parties. Your answers are based on:

1. Publicly available background data and official information (obtained via search tools),
2. Contextual evidence such as party membership records, voting history, public statements, or credible news sources.

ACCESS TO SEARCH:
You are REQUIRED to first use query search tools to research the name in question.
Search for any authoritative or credible sources referencing the individual's
official party membership or widely recognized affiliation.
If you find clear, credible information on the person's party, rely on it.
If information is conflicting or indeterminate, then fall back on contextual inference
grounded in legislative records, news coverage, or the individual's own statements.

TARGET INDIVIDUAL:

* Name: <PERSON-NAME>
* Country: <COUNTRY>
* Approximate year: <YEAR>
* Potential Parties in this Country (PARTIES\_OF\_COUNTRY): {<PARTIES-OF-COUNTRY>}

CONSTRAINTS FOR "pol\_party":

1. You MUST choose exactly ONE party from the above list for "pol\_party".
2. You must NOT introduce any party that is not in the list.
3. You must preserve EXACT spelling, capitalization, and punctuation
   for the chosen party as it appears in the list (INCLUDING ABBREVIATIONS).
4. All explanations should be written in English.

RESPONSE FORMAT:
Your output must follow this precise JSON structure (and nothing else):
{
"justification": "A concise one-sentence justification citing either the external source findings or,
      if no consensus, contextual inference from legislative or news records.",
"pol\_party": "A party from PARTIES\_OF\_COUNTRY with exact same spelling, capitalization/punctuation/abbreviation style.",
"pol\_party\_relaxed": "One party, best prediction, no constraints.",
"confidence": "Confidence in your answer (High, Medium or Low)."
}

IMPORTANT REQUIREMENTS:

* Do NOT include additional text or commentary beyond the JSON object.
* If you find verifiable sources confirming the individual's party, reference them EXPLICITLY
  within your single-sentence justification.
* If no definitive sources exist, clearly state that your choice is based on contextual inference.
* If there are remaining ambiguities, select the MOST likely
  based on searched content and public records.

FINAL TASK STEPS:

1. Use search tools to verify the individual's publicly acknowledged party membership.
2. If confirmed, select that party from the PARTIES\_OF\_COUNTRY list.
3. If conflicting or no direct sources, apply best-effort sources-first, context-second inference.
4. Output the strict JSON block:
   {
   "justification": "...",
   "pol\_party": "...",
   "pol\_party\_relaxed": "...",
   "confidence": "..."
   }

WARNINGS:

* Under NO circumstances produce any output outside the JSON format.
* Any deviation from this exact JSON structure risks rejection.
* Justification must be ONE sentence only.
* Party must match EXACTLY the spelling in the list.
```

Hyperparameters and their consequences. Key choices and expected effects are as follows:

1. **LLM decoding.** A near-deterministic setting reduces variance in structured outputs and improves JSON validity, at the cost of slightly lower recall on ambiguous cases.

2. **Fuzzy-match thresholds.** Country thresholds (0.92; , 0.85+0.08) and stricter global thresholds (0.95; , 0.90+0.12) suppress false alignments across near-synonyms, benefiting cross-national comparability while increasing abstentions where parties have overlapping or evolving brands.

3. **Confidence gate.** Withholding `Low` and `Medium` confidence predictions curbs label noise—especially for minority and emergent parties—but mechanically reduces coverage; this is the intended precision-first behavior for a reference-quality dataset.

Languages and programs. The agentic layer is implemented in **Python** with `langgraph` (ReAct) and OpenAI-compatible chat bindings; retrieval uses Wikipedia and general web search connectors. Orchestration is done in R. Caching and data interchange use **SQLite** with `DBI`/`RSQLite`.

Applicability and limitations. The design is most reliable for *verifiable attributes* (e.g., party membership), where authoritative sources exist and closed codebooks are stable. Performance is predictably uneven for (i) very small or new parties, (ii) early cohorts with sparse web footprints, and (iii) languages/alphabets with heterogeneous transliteration conventions. The relaxed mapper mitigates harmless surface variation without altering class boundaries, and the abstention gate limits error propagation. While the same architecture extends to other attributes (e.g., ethnicity, religion), attributes lacking public self-identification should be treated with additional caution and stricter confidence gating.

Summary. In sum, the ASA combines a *langgraph* ReAct controller, source-first prompts, conservative similarity thresholds, and a confidence gate to deliver auditable, high-precision labels at scale. The resulting pipeline is portable across countries (closed codebooks), reproducible (stored traces and caches), and well-suited to augment expert codings with verifiable, citation-backed inferences.

Table 7: *Sample agent traces. Text content truncated for readability (and may contain typograical errors as present in native source). Links are clickable. Full traces contain many more sources.*

Field	Content
Entry 1: Syleiman Abusaidovich Kerimov — Russian Federation (1999)	
Country	Russian Federation
Year	1999
Person	Syleiman Abusaidovich Kerimov
Wikipedia	Page: Ashot Egiazaryan Summary: Ashot Gevorkovich Egiazaryan (Russian: ; Armenian: ; born...
Search 1	In the spring of 1998, Yeltsin dismissed Chernomyrdin as head of government and in1999Yeltsin's administration backed a newly formedparty,Un...
URL 1	https://en.wikipedia.org/...
Search 2	OURHOMEISRUSSIAPARTYOurHomeIsRussia(Nash Dom—Rossiya, or NDR) was a sociopolitical movement and a rulingpartyfrom 1996 to 1998. Source for i...
URL 2	https://www.encyclopedia.com/...
Entry 2: Jasminka Stanojevic — Serbia (2018)	
Country	Serbia
Year	2018
Person	Jasminka Stanojevic
Wikipedia	Page: Supreme Court (Serbia) Summary: The Supreme Court (Serbian: , romanized: Vrhovni sud) is the court of last resort in Serbia...
Search 1	This article lists political parties inSerbia, including parties that existed in the Kingdom ofSerbiabetween the early 1860s and 1918. A kol...
URL 1	https://en.wikipedia.org/...
Search 2	Imali su dve i četiri godine kad smo izbegli iz Knina. Kad bi neko pokucao na vrata, vikali bi: „Tata, tata". Tri godine nakon progona sazna...
URL 2	https://www.kurir.rs/...
Entry 3: Mihai STROE — Romania (2011)	
Country	Romania
Year	2011
Person	Mihai STROE
Wikipedia	Page: Adrian Stroe Summary: Adrian Stroe (born 24 October 1959), known as The Taxi Driver of Death, is a Romanian serial killer responsible ...
Search 1	Născut în Bucureşti şi cu origini în comuna argeşeană Morăreşti, fost medaliat cu aur la olimpiada internaţională de informatică,MihaiStroe(...
URL 1	https://adevarul.ro/...
Search 2	MihaiSTROEParliamentary activity in legislature 2008-2012 DEPUTY Constituency no.38 TULCEA, uninominal college no.2 Membru al PDL, deputatul...
URL 2	https://www.cdep.ro/...
Entry 4: Matsie Angelina Motshekga — South Africa (2018)	
Country	South Africa
Year	2018
Person	Matsie Angelina Motshekga
Wikipedia	Page: Angie Motshekga Summary: Matsie Angelina "Angie" Motshekga (born 19 June 1955) is a South African politician and educator who is curre...
Search 1	MatsieAngelina"Angie"Motshekga(born 19 June 1955) is a SouthAfricanpolitician and educator who is currently serving as the Minister of Defen...
URL 1	https://en.wikipedia.org/wiki/Angie_Motshekga
Search 2	Motshekgawas elected thenationalpresident of theAfricanNationalCongressWomen's League (ANCWL) in 2008, defeating the League's secretary-gene...
URL 2	https://www.sahistory.org.za/people/matsie-angelina-motshekga-angie-motshekga