

Supplementary Information

1 Analytical workflow overview

To systematically examine how higher-order interactions among five major factors—writing systems, official use, geographical distribution, historical contexts, and economic condition—shape global language endangerment, we employed a three-step analytical workflow: (1) data sources and preparation, (2) network analysis, and (3) random forest regression analysis.

Fig. S1 illustrates the three-step workflow, showing the input datasets, the procedures for data preparation and analysis, and the resulting outputs (tables, figures, and data) for each step. In STEP 1, we prepare analytical datasets from Ethnologue and World Development Indicators. In STEP 2, we conduct community detection, strength analysis, and edge-based percolation simulation to examine tertiary interactions among writing system, official use, and one contextual factor—geographical distribution, historical contexts, or economic conditions. In STEP 3, we conduct random forest regression analysis and following analyses to empirically identify patterns in the effects of the five factors and identify languages overlooked by current frameworks for evaluating language endangerment.

2 Data sources and preparation

Language extraction criteria

From Ethnologue’s original 7,615 languages (spoken and sign; existing and extinct), we extracted 7,010 languages using three criteria. The column names and values indicated in quotation marks refer to the database’s column names and values:

1. Indigenous status (“Is_Indigenous” == “T”): Capturing languages that originated within the country, including pidgins and creoles at their places of contact, and excluding recent immigrant languages.
2. Spoken-language status (“Language_Family” != “Sign Language”): Restricting the analysis to spoken languages. Sign languages were excluded: this study examines language endangerment as it arises from language shift between spoken languages, and a shift in modality between spoken and sign languages falls outside the scope of the present framework. This criterion reflects language modality (spoken vs. sign), not current frequency of use.
3. Non-extinct status (“Extinct” == 0): Excluding only those languages classified under EGIDS 10, which represents languages that have fallen completely out of use with no remaining associated ethnic identity.

Writing system: Definition and Operationalisation

A language has Ethnologue-coded written status if it is coded as written in Ethnologue’s “Is_Written” column (“Is_Written” == “T”). However, this binary variable requires explicit caveats regarding its interpretation and encompasses several complexities:

- A language may possess multiple orthographies
- An orthography may be used only by specialists rather than the broader community
- An orthography may be recently developed
- A writing system may not be widely taught
- A community member may be literate primarily in another language
- A language may retain strong oral vitality and intergenerational transmission independently of widespread written representation

This binary variable is treated strictly as a structural indicator only for the purpose of stratifying the network into analytical multilayers and serving as a predictor in exploratory regression analysis. It does not measure:

- Literacy rates or the strength and breadth of literacy in the language community
- Orthographic stability, standardisation, or active daily use
- Language vitality, intergenerational transmission status, or community practice

3 Network analysis

Official use: Definition and Operationalisation

Following Ethnologue's framework, a language has Ethnologue-coded governmental working-language status if it has an official function within a country or is specifically recognised in legislation. Ethnologue's fifteen official recognition categories are constructed from three parameters: (1) legal status (statutory or de facto), (2) nature of official use (working language versus symbolic language of identity), and (3) geopolitical scope (national or provincial level).

From these fifteen categories, we selected eight categories in which a language functions as the working language for national or provincial governments or administrations:

- Statutory national language
- Statutory national working language
- De facto national language
- De facto national working language
- Statutory provincial language
- Statutory provincial working language
- De facto provincial language
- De facto provincial working language

We deliberately excluded categories where a language is recognised primarily as a symbol of national or provincial identity without functioning in active government business:

- Statutory language of national identity
- De facto language of national identity
- Statutory language of provincial identity
- De facto language of provincial identity

We also excluded categories representing other forms of statutory recognition not tied to government function:

- Recognized language
- Provincially recognized language
- Language of recognized nationality

This binary variable was constructed for network stratification and serves as a predictor in exploratory regression analysis. However, this operationalisation is treated strictly as an administrative-function indicator only and does not measure or capture:

- Community vitality or speaker demographics
- Everyday intergenerational transmission or transmission security
- Educational use or implementation
- Public-service use or implementation
- Local administrative practice
- Protection against language endangerment

Language (ISO 639-3)	a. United States	b. Canada	c. Mexico
1. English (eng)	Layer 1	Layer 1	Layer 2
2. French (fra)	Layer 2	Layer 1	Layer 2
3. Spanish (spa)	Layer 1	—	Layer 1
4. Hawaiian (haw)	Layer 1	—	—
5. Plains Cree (crk)	Layer 2	Layer 2	—
6. Yaqui (yaq)	Layer 2	—	Layer 2
7. Assiniboine (asb)	Layer 3	Layer 3	—
8. Kumiai (dih)	Layer 3	—	Layer 3
9. Central Nahuatl (nhn)	—	Layer 2	—
10. North Alaskan Inuktitut (esi)	Layer 2	Layer 2	—

Table S1. Subset of 10 languages and 3 countries used for network construction and percolation simulation.

Matrix calculations for network construction and percolation simulation

To explain the specific procedures of network construction and percolation simulation, we conducted concrete calculations, using a subset of 10 languages and 3 countries presented in Table S1.

Layer 1 comprises a bipartite network with language nodes $U_1 = \{1, 2, 3, 4\}$ and country nodes $V_1 = \{a, b, c\}$, where the edge set is defined as $E_1 = \{(1, a), (1, b), (2, b), (3, a), (3, c), (4, a)\}$. The corresponding incidence matrix $B^{(1)}$ is

$$B^{(1)} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad (\text{rows } 1, 2, 3, 4; \text{ cols } a, b, c). \quad (1)$$

Layer 2 comprises a bipartite network with language nodes $U_2 = \{1, 2, 5, 6, 9, 10\}$, country nodes $V_2 = \{a, b, c\}$, where the edge set is defined as $E_2 = \{(1, c), (2, a), (2, c), (5, a), (5, b), (6, a), (6, c), (9, b), (10, a), (10, b)\}$. The corresponding incidence matrix $B^{(2)}$ is

$$B^{(2)} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \quad (\text{rows } 1, 2, 5, 6, 9, 10; \text{ cols } a, b, c). \quad (2)$$

Layer 3 comprises a bipartite network with language nodes $U_3 = \{7, 8\}$, and country nodes $V_3 = \{a, b, c\}$, where the edge set is defined as $E_3 = \{(7, a), (7, b), (8, a), (8, c)\}$. The corresponding incidence matrix $B^{(3)}$ is

$$B^{(3)} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad (\text{rows } 7, 8; \text{ cols } a, b, c). \quad (3)$$

For each layer, the language and country projections are

$$\mathbf{W}_L^{(1)} = \mathbf{B}^{(1)}(\mathbf{B}^{(1)})^T = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 2 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \quad \mathbf{W}_C^{(1)} = (\mathbf{B}^{(1)})^T \mathbf{B}^{(1)} = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad (4)$$

$$\mathbf{W}_L^{(2)} = \mathbf{B}^{(2)}(\mathbf{B}^{(2)})^T = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 1 & 2 & 0 & 1 \\ 0 & 1 & 2 & 1 & 1 & 2 \\ 1 & 2 & 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 1 & 2 \end{bmatrix}, \quad \mathbf{W}_C^{(2)} = (\mathbf{B}^{(2)})^T \mathbf{B}^{(2)} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 0 \\ 2 & 0 & 3 \end{bmatrix}, \quad (5)$$

$$\mathbf{W}_L^{(3)} = \mathbf{B}^{(3)}(\mathbf{B}^{(3)})^T = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{W}_C^{(3)} = (\mathbf{B}^{(3)})^T \mathbf{B}^{(3)} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \quad (6)$$

Diagonal entries correspond to self-loops and are set to zero for analysis.

In this modified percolation simulation, we apply percolation to the Layer 3 bipartite network, rather than to the projected network, and observed the resulting changes in the projected Layer 3 language network. In this subset, suppose that Kumiai (dih) spoken in the United States has become extinct. The resulting matrix is denoted $\tilde{\mathbf{B}}^{(3)}$, and the corresponding projected language network matrix is as follows, respectively:

$$\tilde{\mathbf{B}}^{(3)} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{W}_L^{(3)} = \tilde{\mathbf{B}}^{(3)}(\tilde{\mathbf{B}}^{(3)})^T = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}. \quad (7)$$

As shown above, the extinction of Kumiai (dih) in the United States has led to a loss of connections in Layer 3.

Community detection and strength analysis

For 121 languages in the largest connected component of the Layer 1 language network, Infomap detected 2 first-level communities comprising 15 second-level subcommunities. L1–C1 contained 13 subcommunities, and L1–C2 contained 2 subcommunities. Strength scores were also computed for these 121 languages. Supplementary Data 1 provides the community/subcommunity number, strength score, and ranking of each language. The columns are as follows: “ISO 639” is the three-letter identifier assigned by the ISO 639-3 standard; “Language Name” is the corresponding name of each language; “Community” is the community number to which the language belongs; “Subcommunity” is the subcommunity number within the assigned community; “Community–Subcommunity” is the combined label of the community and subcommunity; “Strength” is the strength score of the language; and “Strength Rank” is the ranking of the language based on its strength score in decreasing order among all languages in the largest connected components of the Layer 1 language network.

For 3,962 languages in the largest connected component of the Layer 2 language network, Infomap detected 5 communities. Subcommunity detection was performed but not reported. Communities L2–C2, C3, C4, C5 exhibited local connections, with subcommunity structures corresponding approximately to groupings by individual country. L2–C1 exhibited long-range connections, and its subcommunity structure separated languages in South America from the rest. In all cases, subcommunities were closely connected primarily by national boundaries, providing no additional informative structural patterns relevant to the main analysis. Supplementary Data 2 provides strength scores and rankings for languages in the largest connected components of the Layer 2 language network. The columns are as follows: “ISO 639” is a three-letter identifier assigned by the ISO 639-3 standard; “Language Name” is the corresponding name of each language; “Community” is the community number to which the language belongs; “Strength” is the strength score of the language; and “Strength Rank” is the ranking of the language based on its strength score in decreasing order among all languages in the largest connected components of the Layer 2 language network.

For the Layer 3 language networks, Infomap was applied independently to each of the five largest connected components. However, its results are not reported because community detection revealed that languages were closely connected primarily by national boundaries with no additional informative structural patterns. Strength analysis was also applied but not reported because it similarly showed minimal variation: strength scores were predominantly 1, with a strength score of 2 only for languages spoken in two countries.

High-risk area selection

High-risk areas were identified within each Layer 2 community and Layer 3 connected component by ranking countries according to their number of endangered and dormant languages. Within each community or connected component, the top three countries were prioritised. A country was retained if its count of endangered and dormant languages reached approximately 60% or more of that of the preceding rank, ensuring comprehensive coverage of representative areas; countries falling below this threshold, or ranking below third, were excluded to avoid non-representativeness. In addition, countries with a small absolute number of endangered and dormant languages were excluded even when they met the proportional threshold, so that the selected areas reflect substantive concentrations of endangerment.

In Layer 2, communities L2–C1 (48.9%), L2–C2 (30.5%), and L2–C4 (26.5%) were retained, whereas L2–C3 (10.8%) and L2–C5 (10.9%) were excluded owing to their low proportions of endangered and dormant languages. In Layer 3, all five connected components were retained. Within each retained community or component, high-risk countries were selected as follows:

- L2–C1: Mexico (74 endangered and dormant languages), the United States (72), and Canada (42) were retained. Russia (38) and lower-ranked countries were excluded as they ranked below third.
- L2–C2: Papua New Guinea (132) and Indonesia (100) were retained. Malaysia (30) fell below 60% of the preceding rank and was excluded.
- L2–C4: India (61), Nepal (43), and China (26) were retained. Vietnam (19) and lower-ranked countries were excluded as they ranked below third.
- L3–CC1: Indonesia (325), Papua New Guinea (180), and Australia (155) were retained. China (106) and lower-ranked countries were excluded as they ranked below third.
- L3–CC2: Nigeria (113) and Cameroon (59) were retained. Sudan (22) fell below 60% of the preceding rank and was excluded.
- L3–CC3: The United States (106), Mexico (50), and Canada (22) were retained.
- L3–CC4: Brazil (81) was retained. Colombia and lower-ranked countries were excluded as they fell below 60% of Brazil.
- L3–CC5: Côte d'Ivoire was retained. Ghana (7) and lower-ranked countries were excluded owing to their small absolute numbers of endangered and dormant languages.

Percolation simulation of the Layer 2 language network

As a preliminary analysis parallel to Layer 3, we examined how national wealth level and income inequality are associated with language endangerment among unofficially used languages with writing systems. Following the same procedure, we applied percolation to the Layer 2 bipartite network and observed the resulting changes in the five communities of the projected Layer 2 language network (L2–C1 to L2–C5). At each percolation stage, we (1) measured the sizes of the communities in the resulting language networks and (2) recalculated country-level economic indicators—the GDP per capita and the Gini index—based on the countries that remained connected. Figure S2 illustrates the progressive reduction in community sizes, with the vertical axis showing the percentage of languages in the community and the horizontal axis representing the EGIDS thresholds. Figure S3 presents the economic composition at each percolation stage; panels (a–f) plot the Gini index (x-axis) against log GDP per capita (y-axis) for the countries remaining in the network, with colours denoting L2–C1 to L2–C5 and lines indicating the mean and range for each variable.

Structurally, the communities exhibited a similar pattern of progressive reduction and persisted until later percolation stages (Fig. S2). Languages with long-range connections among geographically distant countries (L2–C1) detached to the greatest extent, with its component size decreasing to approximately 50%. Economically, the percolation simulation revealed almost no change in economic composition (Fig. S3). Although the community sizes decreased through the percolation simulation, the number of constituent countries declined by only a few, and the economic indicators remained nearly constant across the percolation stages. The mean GDP per capita stayed within a narrow range (approximately 14,200–14,500 US\$), and the mean Gini index remained around 37, with no change in their minimum or maximum values. In contrast to Layer 3, no early detachment of high-GDP countries was observed. These results indicate that, for unofficially used languages with writing systems, neither national wealth level nor income inequality showed a clear association with language endangerment. This contrasts with the Layer 3 findings, where higher GDP per capita was associated with language endangerment.

Speaker population as an additional demographic factor

To evaluate the contribution of demographic scale to language endangerment, we extended the binary factor-incidence matrix by adding speaker population as a sixth factor category alongside the five factor categories. Although EGIDS level 10 corresponds to zero speakers, creating a potential definitional circularity when population is used as a predictor, EGIDS is not solely determined by population size, but reflects broader institutional and functional conditions. While EGIDS and log-transformed speaker population are strongly negatively correlated ($r = -0.699$, Supplementary Fig. S4), the relationship is not one-to-one. Therefore, including the speaker population as an additional factor is justified. For the 6,589 languages with available data, the population values were log-transformed (mean $\mu = 3.97$, standard deviation $\sigma = 1.35$) and categorised into six bins using the boundaries $\mu - 2\sigma$, $\mu - \sigma$, μ , $\mu + \sigma$, and $\mu + 2\sigma$. These correspond to approximate speaker-population thresholds of 18, 415, 9,358, 211,120, and 4,762,975.

Binary factor-incidence matrix

For subsequent analyses, we constructed a binary factor-incidence matrix covering all 7,010 languages (Supplementary Data 3), encoding the presence or absence of binary variables across six factor categories: writing system, official use, geographical distribution, historical contexts, economic conditions, and speaker population. The ISO 639-3 identifier and language name are omitted in accordance with Ethnologue's data policy.

Writing system (Factor 1) and official use (Factor 2) are inherently binary, each represented by a single column indicating presence (1) or absence (0).

Geographical distribution (Factor 3) is represented by multiple columns corresponding to the network structures identified in each layer: community and subcommunity numbers for Layer 1, community numbers for Layer 2, and connected-component numbers for Layer 3. Each language receives a value of 1 in the column(s) corresponding to the community, subcommunity, or connected component to which it belongs. A language appearing in both Layers 1 and 2 therefore receives 1 in the relevant columns of both layers. Isolated languages that belong to no community or connected component receive 0 across all geographical-distribution columns.

Historical contexts (Factor 4) comprise six columns, each indicating the presence (1) or absence (0) of a historical context associated with the countries where each language is spoken. These contexts are interpretive categories that link historical processes to the contemporary institutional conditions represented in the multilayer network, rather than assertions of direct causation:

- Extractive colonisation: Languages spoken in countries where English, French, Spanish, Portuguese, Dutch, Russian, or Standard German established long-range institutional connections through colonial administration, excluding settler-colonial countries and the imperial languages themselves.
- Religious and imperial-administrative expansion: Languages spoken in countries where Spanish, Portuguese, or Modern Standard Arabic became and remain an official language through religious expansion—Spanish and Portuguese via colonial-ecclesiastical administration primarily in the Americas, and Modern Standard Arabic via the institutional infrastructure established by the expansion of Islam across Northern Africa and Western Asia.
- Monarchy: Languages spoken in Central and Eastern European countries associated with the monarchical administration of the Habsburg territories and surrounding regions.
- Migration: Languages spoken in countries where Chinese varieties (Mandarin, Yue, Hakka) established regional presence through overseas migration.
- Settler colonisation: Indigenous languages spoken in settler-colonial countries (United States, Canada, New Zealand, Australia, South Africa, Algeria, Liberia, Kenya, Zimbabwe, Israel, Argentina, and Chile), excluding the major imperial languages.
- Creole formation contexts: Creole and pidgin languages in the Caribbean and Pacific regions.

Economic conditions (Factor 5) are based solely on GDP per capita (US\$), which was found to be associated with language endangerment, whereas the Gini index was not used. GDP per capita values were log-transformed and divided into three bins: $10^2 - 10^3$, $10^3 - 10^4$, $10^4 - 10^5$. This factor applies only to languages spoken in the countries represented in Layer 3; languages spoken elsewhere receive 0 across all economic-condition columns.

Speaker population (Factor 6) is represented by six binned columns derived from log-transformed speaker population, as described above, indicating whether a language falls within a given population range (within the range = 1; outside the range = 0).

4 Random forest regression analysis

Hyperparameter tuning and model validation

We trained random forest regression models to predict each language's EGIDS endangerment level from its binary factor indicators, comparing a five-factor model (38 variables, 7,010 languages) and a six-factor model (44 variables, 6,589 languages). For each model, the data were split into training (80%) and independent test (20%) sets. Hyperparameters were tuned on the training set via grid search (GridSearchCV) over 72 parameter combinations with five-fold cross-validation, optimising the cross-validated R^2 . The selected hyperparameters were, for the five-factor model, $n_estimators = 300$, $max_depth = 20$, $max_features = 'sqrt'$, $min_samples_split = 2$, $min_samples_leaf = 2$; and for the six-factor model, $n_estimators = 200$, $max_depth = 15$, $max_features = 'sqrt'$, $min_samples_split = 5$, $min_samples_leaf = 1$.

To assess robustness, each model was trained under five random seeds (42, 123, 456, 789, 999), and performance was evaluated using cross-validated R^2 , independent test-set R^2 and RMSE, and out-of-bag (OOB) estimates. Results were stable across seeds. The five-factor model achieved a cross-validated R^2 of 0.562 ± 0.001 (RMSE = 0.942 ± 0.023) and a test-set R^2 of 0.576 ± 0.001 (RMSE = 0.922 ± 0.001), with an OOB R^2 of 0.566 ± 0.001 . The six-factor model achieved higher performance, with a cross-validated R^2 of 0.670 ± 0.003 (RMSE = 0.749 ± 0.024) and a test-set R^2 of 0.666 ± 0.001 (RMSE = 0.751 ± 0.001), with an OOB R^2 of 0.674 ± 0.001 . Figure S5 shows predicted against observed EGIDS for both models, with the six-factor model exhibiting closer alignment to the identity line, particularly at intermediate endangerment levels.

Overfitting was assessed through the gap between training and test performance. The five-factor model showed a negligible train-test gap (0.003), indicating no meaningful overfitting. The six-factor model exhibited a slightly larger but still modest gap (0.035), consistent with its higher overall fit. Across both models, the close agreement among cross-validated, test-set, and OOB estimates confirmed that the models generalised well to unseen data.

Feature importance analysis

To identify the factor variables most strongly associated with EGIDS, we quantified feature importance using two complementary methods. Mean Decrease in Impurity (MDI) measures each variable's contribution to reducing prediction error within the trained trees, but tends to inflate the apparent importance of high-cardinality variables. Permutation importance, which measures the decrease in test-set performance when a variable's values are randomly shuffled, provides a complementary and less biased assessment. Permutation importance was computed on the independent test set across five random seeds (42, 123, 456, 789, 999) with ten permutation repeats each, and the results were averaged.

For the five-factor model, both methods identified official use and writing system as by far the two most important variables, together accounting for the majority of predictive contribution (Fig. S6). MDI ranked official use first (0.246) and writing system second (0.189), whereas permutation importance ranked writing system first (0.223) and official use second (0.209); this reordering of the top two variables reflects their comparable magnitudes rather than a substantive disagreement. Among the remaining variables, historical-context indicators (particularly extractive and settler colonisation), economic-condition bins, and geographical-distribution communities appeared consistently in the upper ranks under both methods. Geographical-distribution variables, which have higher cardinality, ranked somewhat lower under permutation importance than under MDI, consistent with the known tendency of MDI to overstate the importance of high-cardinality features.

For the six-factor model, both methods again identified official use (MDI = 0.209, permutation = 0.185) and writing system (MDI = 0.119, permutation = 0.109) as the two leading variables, with identical rankings across methods (Fig. S7). The binned speaker-population variables occupied several of the subsequent positions under both methods, confirming the substantive contribution of demographic scale identified in the model-validation analysis. Extractive colonisation also ranked consistently among the top variables.

Overall, the close agreement between MDI and permutation importance—especially for the leading variables—indicates that the identified importance rankings are robust to the choice of method and not artefacts of variable cardinality.

Ablation analysis

To assess the individual and joint contributions of the six historically interpretive context categories, we conducted a leave-one-out ablation analysis. Each category was removed in turn from the full model, and the resulting decrease in cross-validated R^2 was recorded. The analysis was performed for both the five-factor and six-factor models.

In the five-factor model (baseline CV $R^2 = 0.562$), the six historical-context categories together contributed 0.021 R^2 points. Extractive colonisation had the largest individual effect (R^2 drop = 0.012, 2.1%), followed by settler colonisation and religious and imperial-administrative expansion (both 0.004, 0.7%) and migration (0.002, 0.4%). Creole formation contexts and monarchy showed negligible individual contributions (R^2 drop ≈ 0.000). The six-factor model (baseline CV $R^2 = 0.670$) showed the same overall pattern: extractive colonisation again had the largest individual effect (R^2 drop = 0.009, 1.4%), and creole formation contexts and monarchy again contributed negligibly on their own (Fig. S8).

The negligible—or even slightly negative—individual contributions of creole formation contexts and monarchy do not imply that these categories are redundant, because leave-one-out ablation cannot capture joint effects among categories. We therefore examined their joint contribution by removing both categories simultaneously and comparing the resulting R^2 drop against the sum of their individual drops. Although neither creole formation contexts nor monarchy contributed on its own (individual R^2 drops ≈ 0 or slightly negative), their joint removal produced a positive R^2 drop in both models (five-factor: joint drop 0.0002 versus summed individual drops -0.0003; six-factor: 0.0003 versus -0.0004). That is, removing both categories together degrades model performance even though removing either one alone does not, indicating that they contribute jointly rather than independently.

Extractive colonisation showed the largest individual effect on cross validated R^2 , while categories with negligible individual effects—creole formation contexts and monarchy—showed a measurable effect only when removed jointly, rather than individually. This pattern is consistent with retaining the full set of six categories in the model, though it does not itself validate the categories as historically accurate or causally distinct. As with the feature-importance analysis, these ablation results describe the relative predictive contributions of categories within the fitted model and should not be interpreted as confirmatory evidence of the underlying theoretical framework.

Residual analysis and outlier identification

To identify languages whose endangerment levels deviate from model predictions, we analysed the residuals of each fitted model, defined as $e = \text{observed EGIDS} - \text{predicted EGIDS}$. In both models, residuals were approximately normally distributed and centred near zero (five-factor: $\mu = -0.009$, $\sigma = 0.924$; six-factor: $\mu = -0.002$, $\sigma = 0.722$; Fig. S9), indicating no systematic prediction bias.

Outlier languages were defined as those with $|e| \geq 3\sigma$. Upper outliers ($e > 3\sigma$) are more endangered than the model predicts, whereas lower outliers ($e < -3\sigma$) are less endangered than predicted. The five-factor model identified 58 outliers (22 upper, 36 lower), and the six-factor model identified 80 outliers (22 upper, 58 lower).

To confirm that outlier identification was not an artefact of a particular random seed, we repeated the analysis across five seeds (42, 123, 456, 789, 999). The residual statistics were highly stable (five-factor: σ ranging 0.9239-0.9244; six-factor: σ ranging 0.7213-0.7220), and the number of outliers varied only marginally (five-factor: 58-62; six-factor: 78-81). Of these, 58 languages in the five-factor model and 80 in the six-factor model were flagged consistently across all five seeds, confirming that the identified outliers are robust to seed variation.

SHAP-based interpretation of outlier languages

Outlier languages identified from residuals ($|e| \geq 3\sigma$) were analysed with SHAP (SHapley Additive exPlanations) to determine the contribution of each factor to over- or under-prediction. For the five-factor model (Fig. S10), writing-system presence exerted the strongest influence among the 22 upper outliers—languages more endangered than predicted—indicating that the model underestimated endangerment when scripts existed, whereas literacy and functional use were minimal. One typical case includes classical or liturgical languages, which retain writing systems and symbolic value but have limited use in everyday communication. Another typical pattern involves languages with unique scripts and low literacy among speakers, illustrating cases where script invention or preservation occurred without broad literacy or institutional embedding.

In contrast, lower outliers—safer than predicted languages—also exhibited strong protective signals from writing-system presence; however, the model overestimated their endangerment because other risk indicators (e.g., specific historical contexts, belonging to the long-range community L2–C1, and lack of working-language status) were not fully offset. One typical pattern involves languages that are officially recognised despite lacking official function as working languages; such legal recognition brings indirect protection that the model did not fully capture. A second pattern involves regional lingua francas that sustain broad intergroup communication but lack formal status, for which relatively low EGIDS levels may have been assigned despite their wide functional reach.

For the six-factor model (Fig. S11), similar tendencies persisted, with writing system and extractive colonisation remaining the leading contributors. In addition, mid-sized speaker-population bins emerged as additional sources of bias among both outlier groups. Among the upper outliers, languages with large absolute populations were over-predicted to be stable, as their population size masked rapid intergenerational decline concentrated among older speakers. Among the lower outliers, officially or regionally used mid-sized languages were often under-predicted as endangered, indicating that population scale can both inflate and obscure vitality depending on demographic composition.

The stability of these SHAP-based explanations was assessed across five random seeds (42, 123, 456, 789, 999). The leading features explaining both the upper and lower outlier groups were identified consistently in all five seeds, and the overall SHAP-based feature rankings were highly correlated across seeds (mean Spearman $\rho = 0.996$, minimum $\rho = 0.994$ for both models). These results confirm that the SHAP-based interpretations of outlier languages are robust and not dependent on a particular random seed.

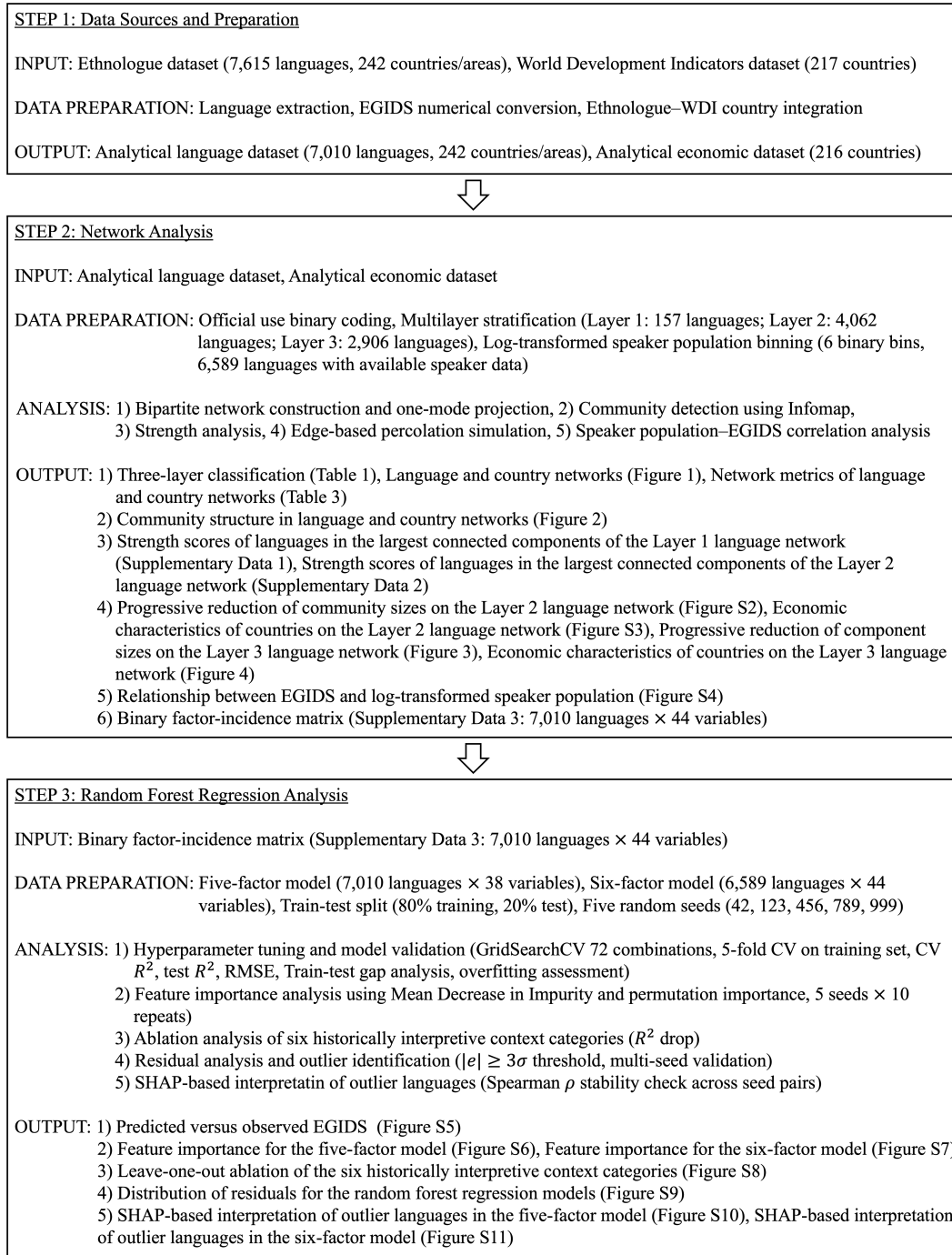


Figure S1. Analytical workflow overview. STEP 1 (Data sources and preparation) shows the integration of language datasets from Ethnologue and economic datasets from World Development Indicators, resulting in analytical datasets. STEP 2 (Network analysis) constructs and analyses multilayer bipartite networks through community detection, strength analysis, and edge-based percolation simulation, then encodes the results into a binary factor-incidence matrix. STEP 3 (Random forest regression analysis) conducts random forest regression analysis with feature importance analysis and residual analysis with SHAP to examine factor contributions and identify languages overlooked by current evaluation frameworks.

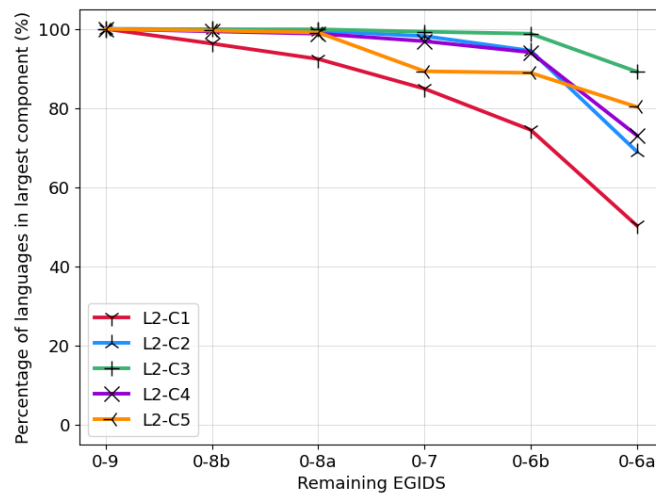


Figure S2. Percolation simulation on the Layer 2 language network: progressive reduction of component sizes. The curves show the share of languages in the five communities (y-axis) as edges incident to endangered and dormant languages are removed in the descending EGIDS thresholds (x-axis: 9→8b→8a→7→6b→6a). Each colour identifies each of the five communities (L2-C1 to L2-C5). The five communities exhibited a similar pattern of reduction and persisted until later percolation stages. L2-C1, the community characterised by long-range connections among geographically distant countries, showed the greatest decline, with its community size decreasing to approximately 50%.

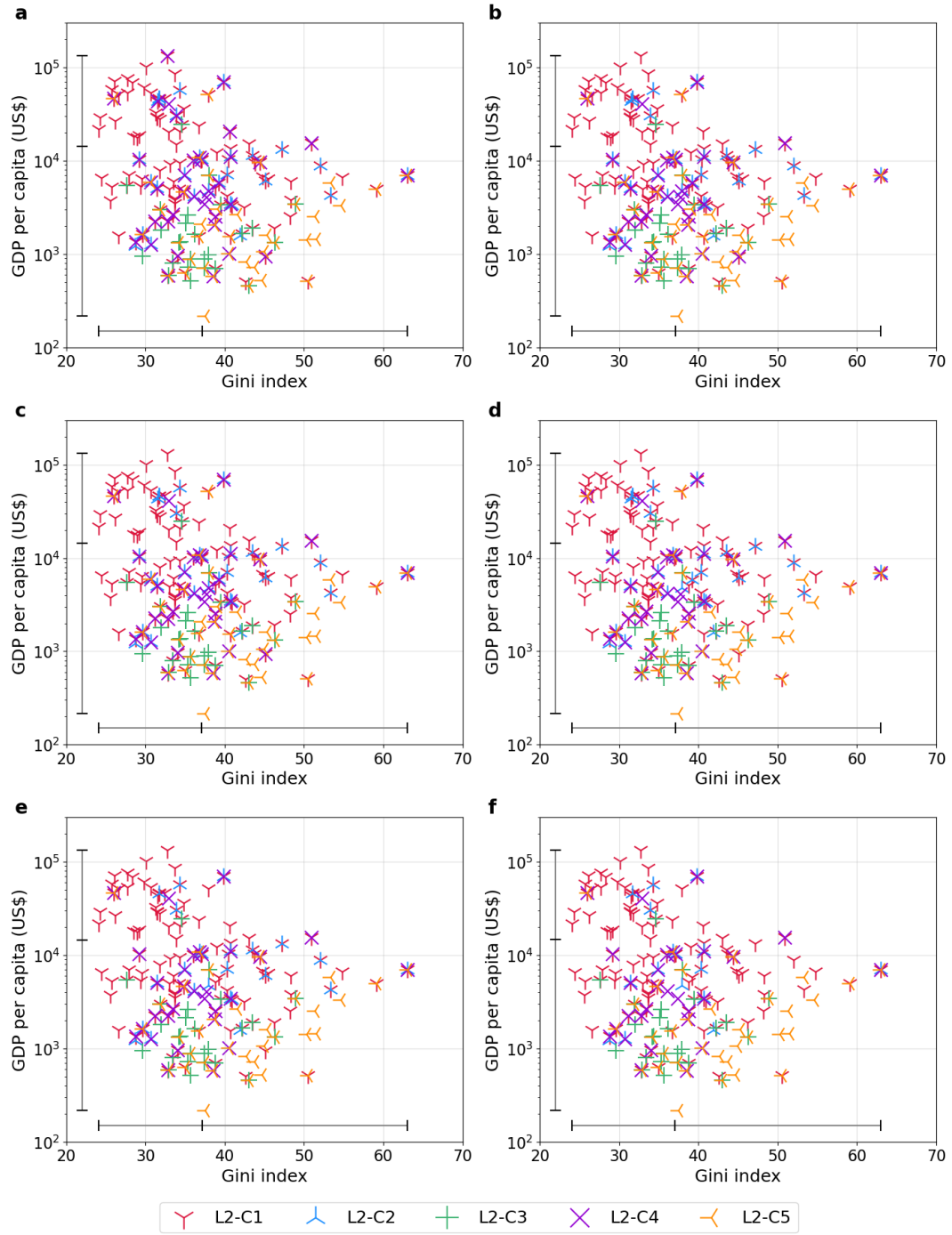


Figure S3. Percolation simulation on the Layer 2 language network: economic characteristics of countries. Panels **a-f** correspond to EGIDS thresholds 0-9, 0-8b, 0-8a, 0-7, 0-6b, and 0-6a. Each point is a country retained at that stage; the x-axis is the Gini index and the y-axis is log GDP per capita. Colours denote the five communities in Layer 2 (L2-C1 to L2-C5); because languages spoken in the same country may belong to different communities, plots for a given country can overlap across colours. Vertical and horizontal lines mark the range and the mean of each variable. Although the size of communities decreased through the percolation simulation, the number of constituent countries declined by only a few, and the economic indicators showed almost no change across the percolation stages.

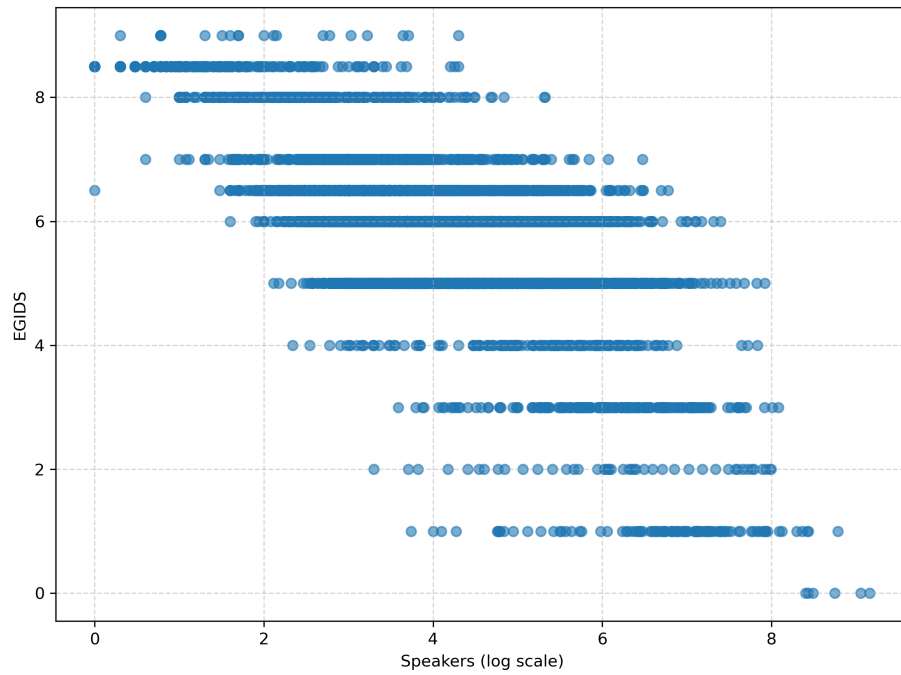


Figure S4. Relationship between EGIDS and speaker population (log scale). Each point represents one of the 6,589 languages, plotted using the EGIDS endangerment level (vertical axis) and the log-transformed number of speakers (horizontal axis). A strong negative correlation is observed ($r = -0.699$), supporting the inclusion of speaker population as a sixth explanatory variable despite potential definitional circularity.

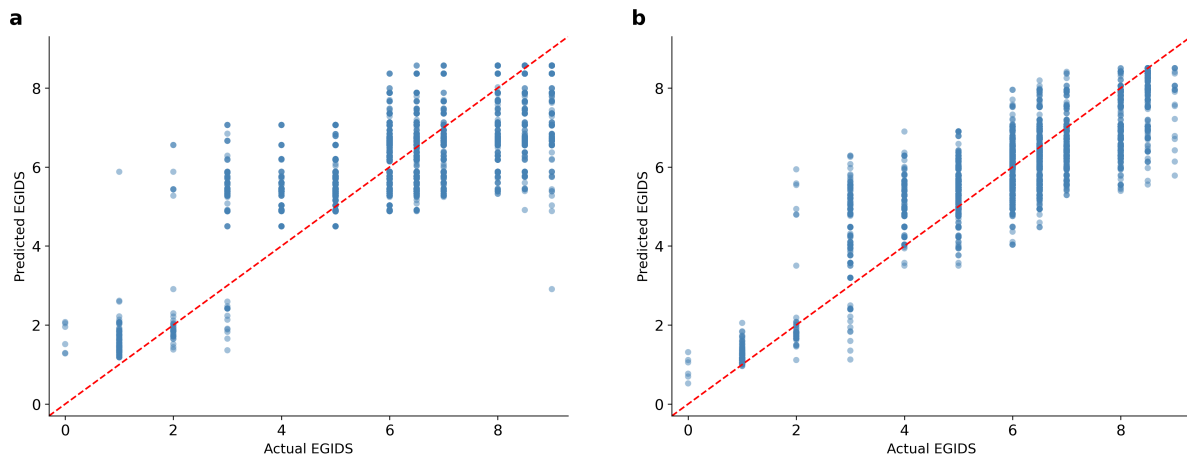


Figure S5. Predicted versus observed EGIDS for the random forest regression models. Each point represents one language, plotted by its observed EGIDS (x-axis) and the value predicted by the model (y-axis); the red dashed line indicates perfect agreement. Panel **a** shows the five-factor model (7,010 languages) and panel **b** shows the six-factor model (6,589 languages). The six-factor model, which incorporates speaker population, shows closer alignment with the identity line, particularly at intermediate EGIDS.

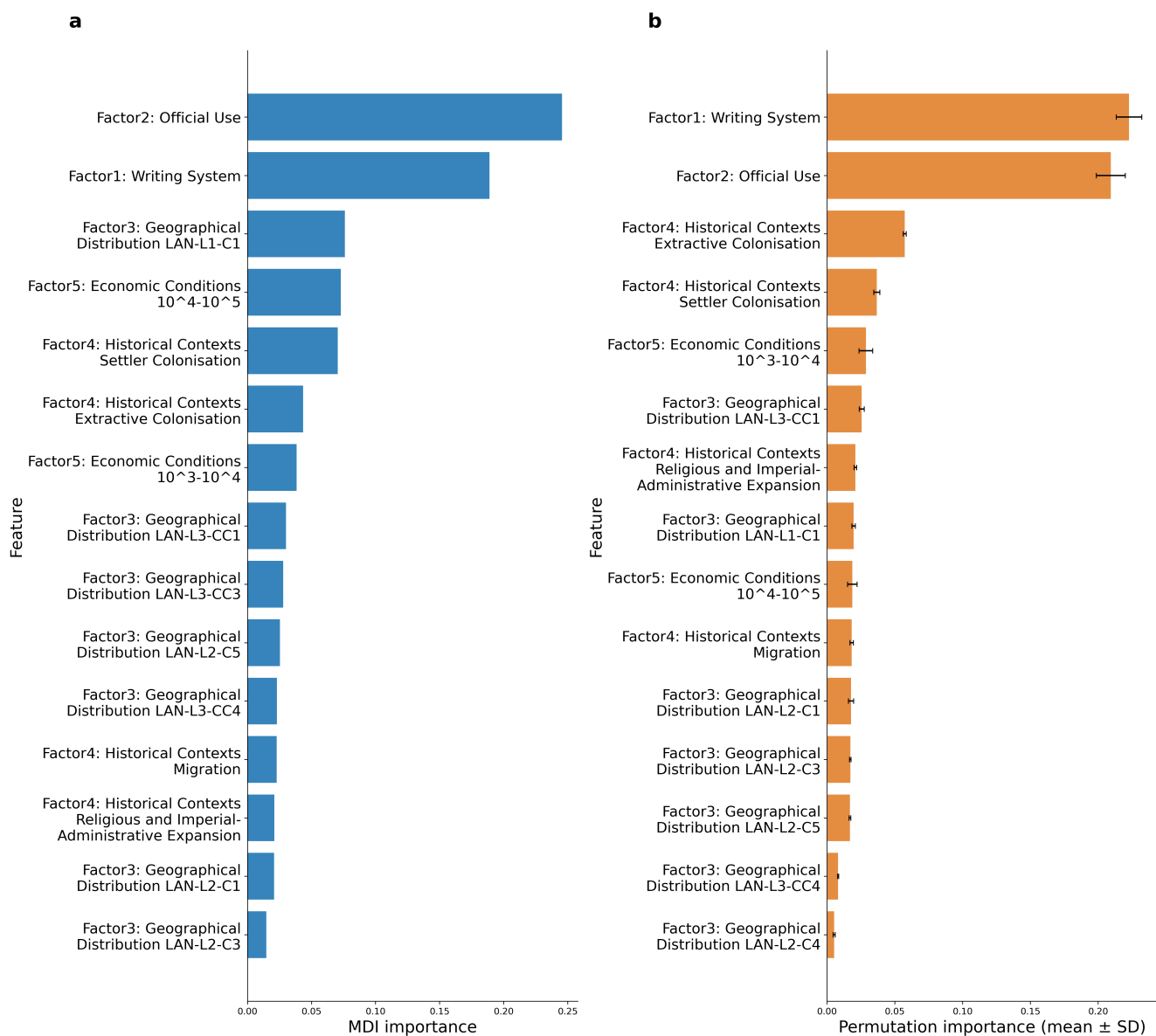


Figure S6. Feature importance for the five-factor model (7,010 languages). The top 15 variables are shown, ordered by decreasing importance. Panel **a** shows Mean Decrease in Impurity (MDI), and panel **b** shows permutation importance computed on the independent test set (mean $\pm$ standard deviation across five random seeds, ten repeats each). Official use and writing system are the two leading variables under both methods; their order is reversed between the two methods, reflecting their comparable magnitudes. Geographical-distribution variables, which have higher cardinality, rank lower under permutation importance than under MDI.

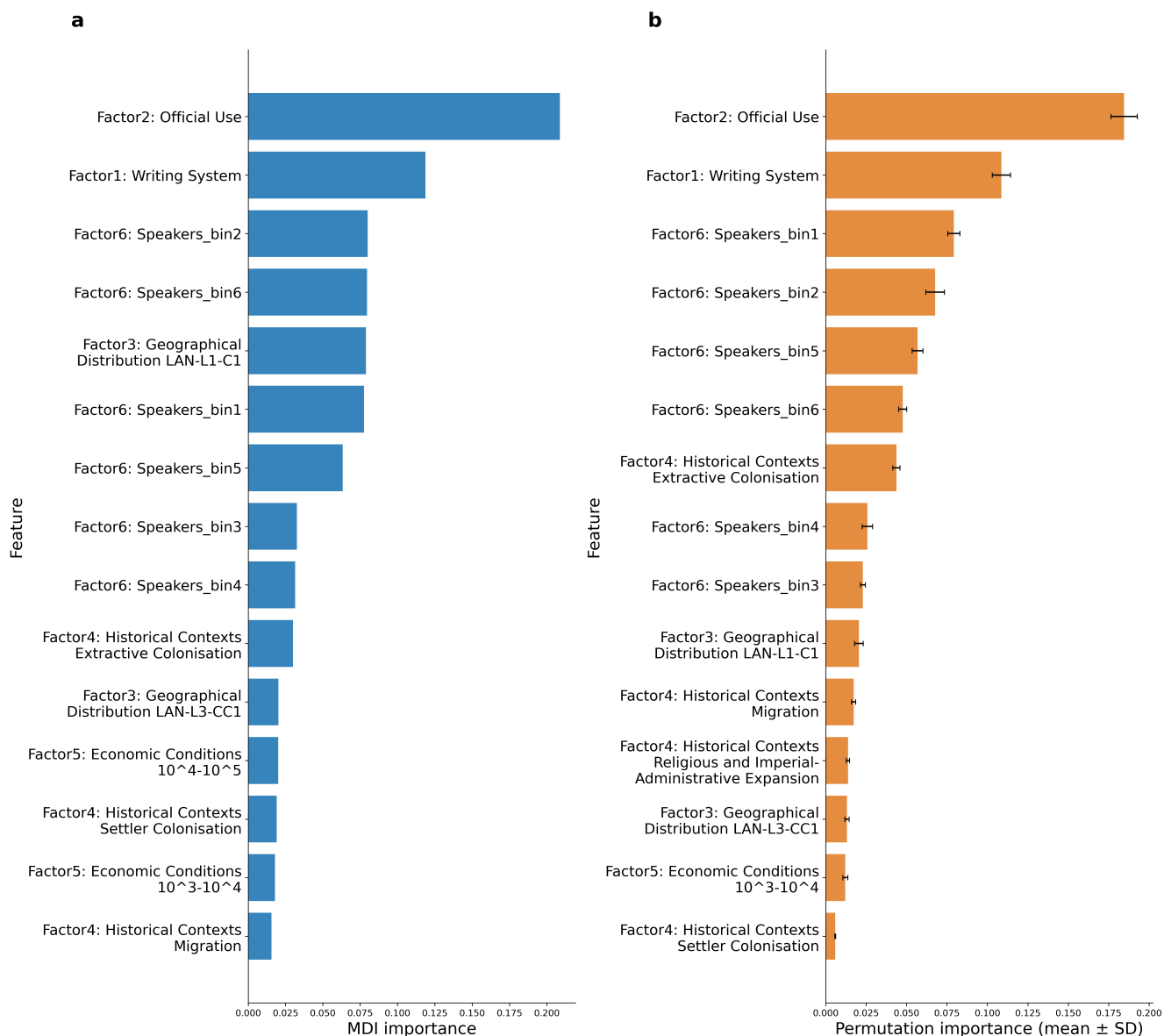


Figure S7. Feature importance for the six-factor model (6,589 languages). The top 15 variables are shown, ordered by decreasing importance. Panel **a** shows Mean Decrease in Impurity (MDI), and panel **b** shows permutation importance computed on the independent test set (mean $\pm$ standard deviation across five random seeds, ten repeats each). Official use and writing system are the two leading variables under both methods, with identical top rankings. The binned speaker-population variables occupy several of the subsequent positions, reflecting the contribution of demographic scale.

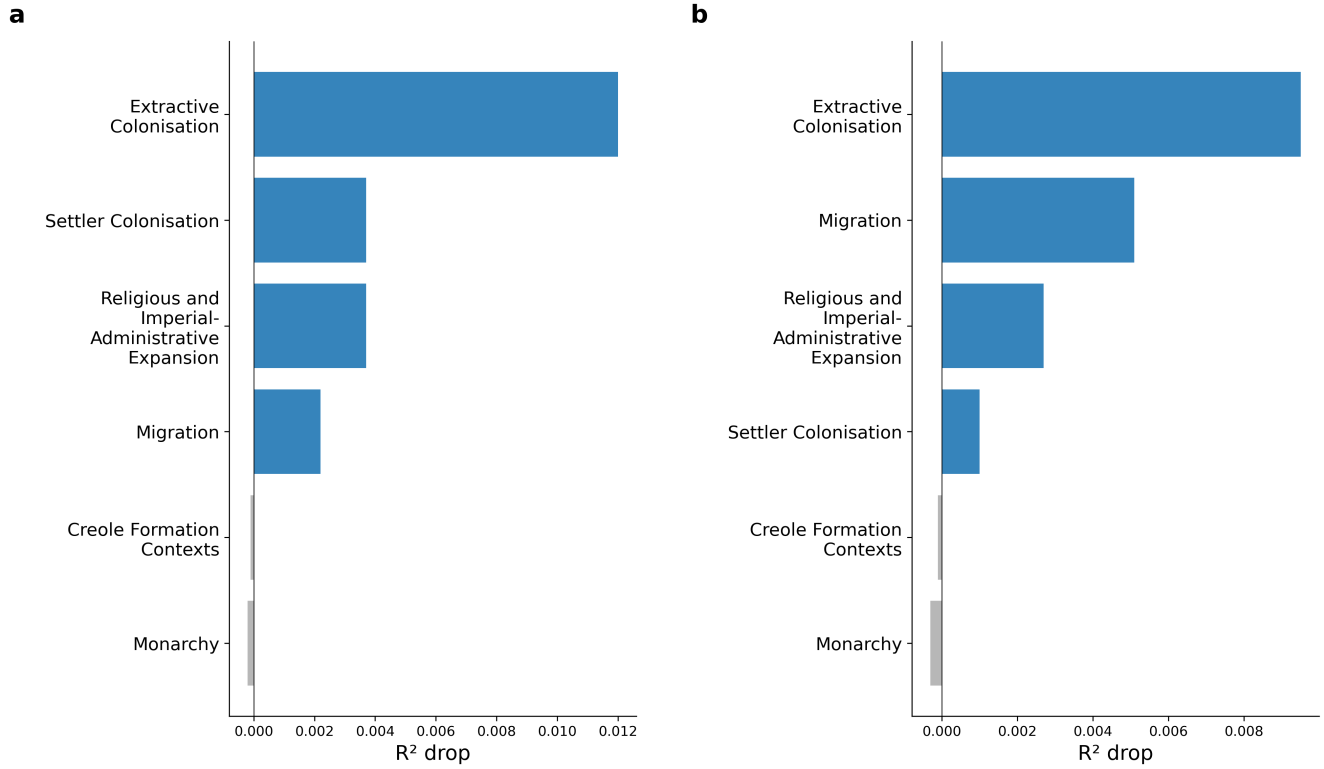


Figure S8. Leave-one-out ablation of the six historically interpretive context categories. Each bar shows the decrease in cross-validated R^2 when a single category is removed from the full model, ordered by decreasing R^2 drop. Panel **a** shows the five-factor model (7,010 languages) and panel **b** shows the six-factor model (6,589 languages). Blue bars indicate positive R^2 drops; grey bars indicate the negligible or slightly negative individual contributions of creole formation contexts and monarchy. Extractive colonisation shows the largest individual effect in both models.

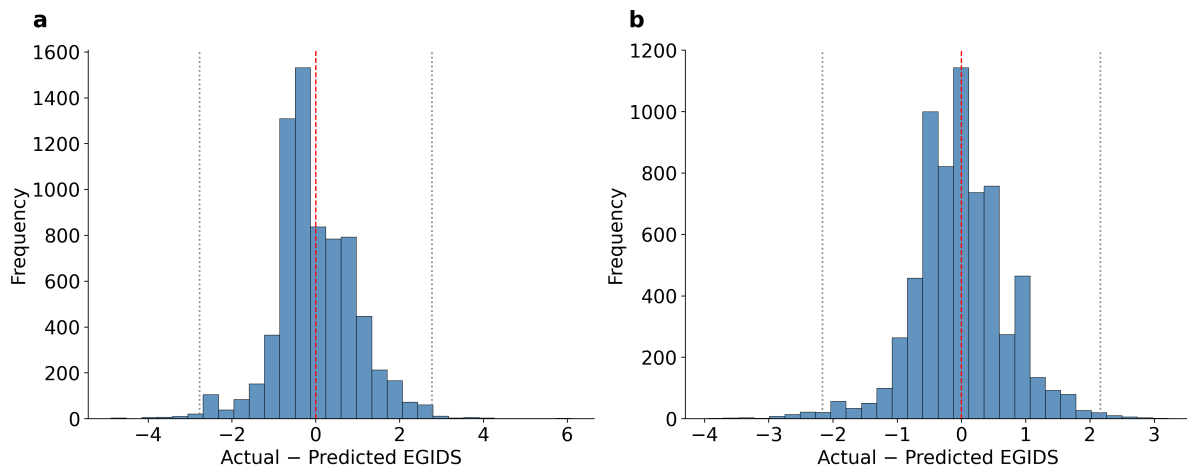


Figure S9. Distribution of residuals (observed - predicted EGIDS) for the random forest regression models. Panel **a** shows the five-factor model (7,010 languages) and panel **b** shows the six-factor model (6,589 languages). The red dashed line marks zero residual, and the grey dotted lines mark the $\pm 3\sigma$ thresholds used to define outlier languages. In both models, residuals are approximately normally distributed and centred near zero, indicating no systematic prediction bias.

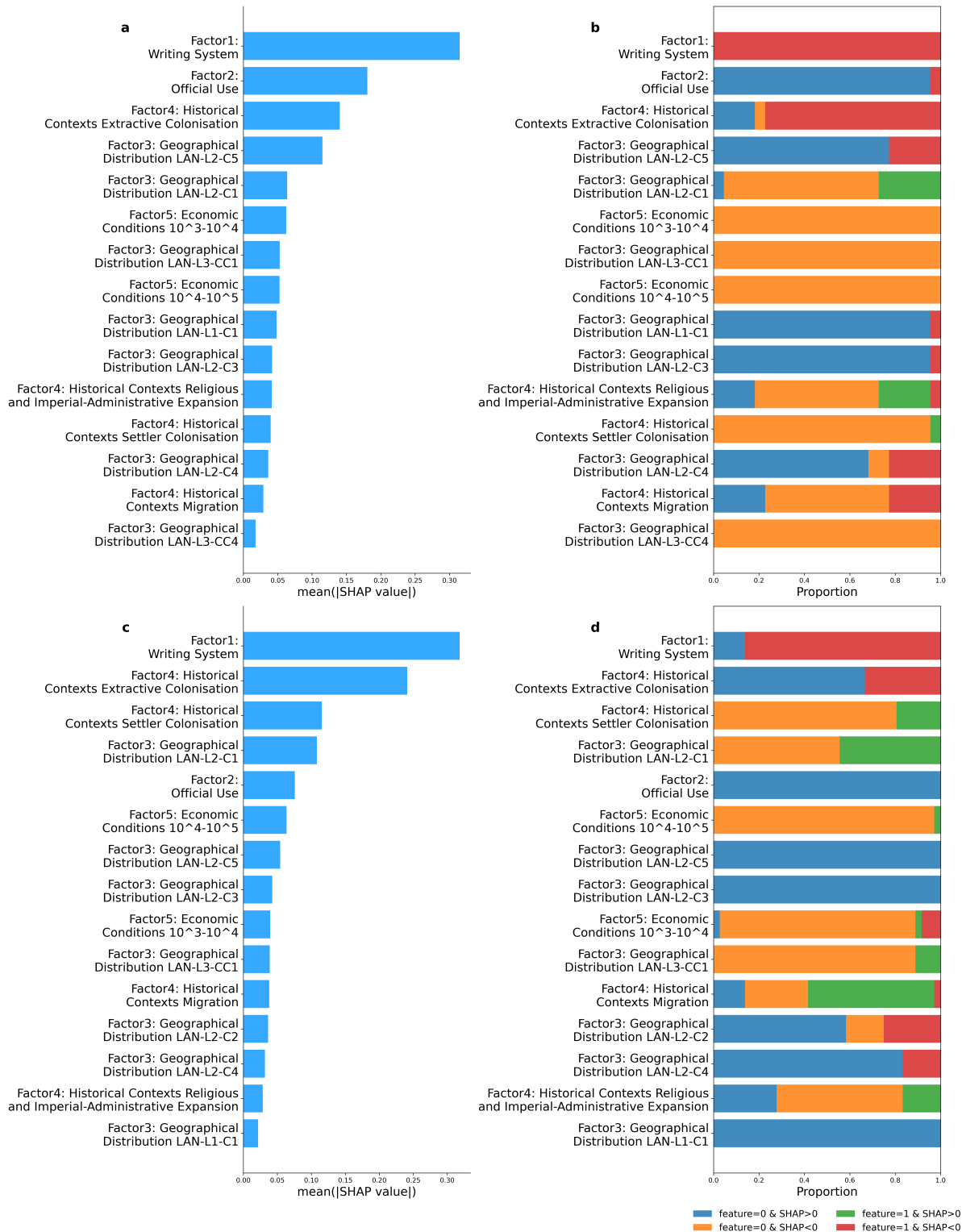


Figure S10. SHAP-based interpretation of outlier languages in the five-factor model. Panels **a** and **b** show the 22 upper outliers (more endangered than predicted), and panels **c** and **d** show the 36 lower outliers (less endangered than predicted). Panels **a** and **c** show the top 15 features ranked by mean absolute SHAP value. Panels **b** and **d** show, for the same features, the proportion of languages in each of four categories: feature absent with a positive SHAP contribution (feature= 0 & SHAP> 0), feature absent with a negative contribution (feature= 0 & SHAP< 0), feature present with a positive contribution (feature= 1 & SHAP> 0), and feature present with a negative contribution (feature= 1 & SHAP< 0). Writing system is the leading contributor in both outlier groups; among the upper outliers, its presence consistently lowers the predicted endangerment level, contributing to the underestimation of these languages' actual endangerment.

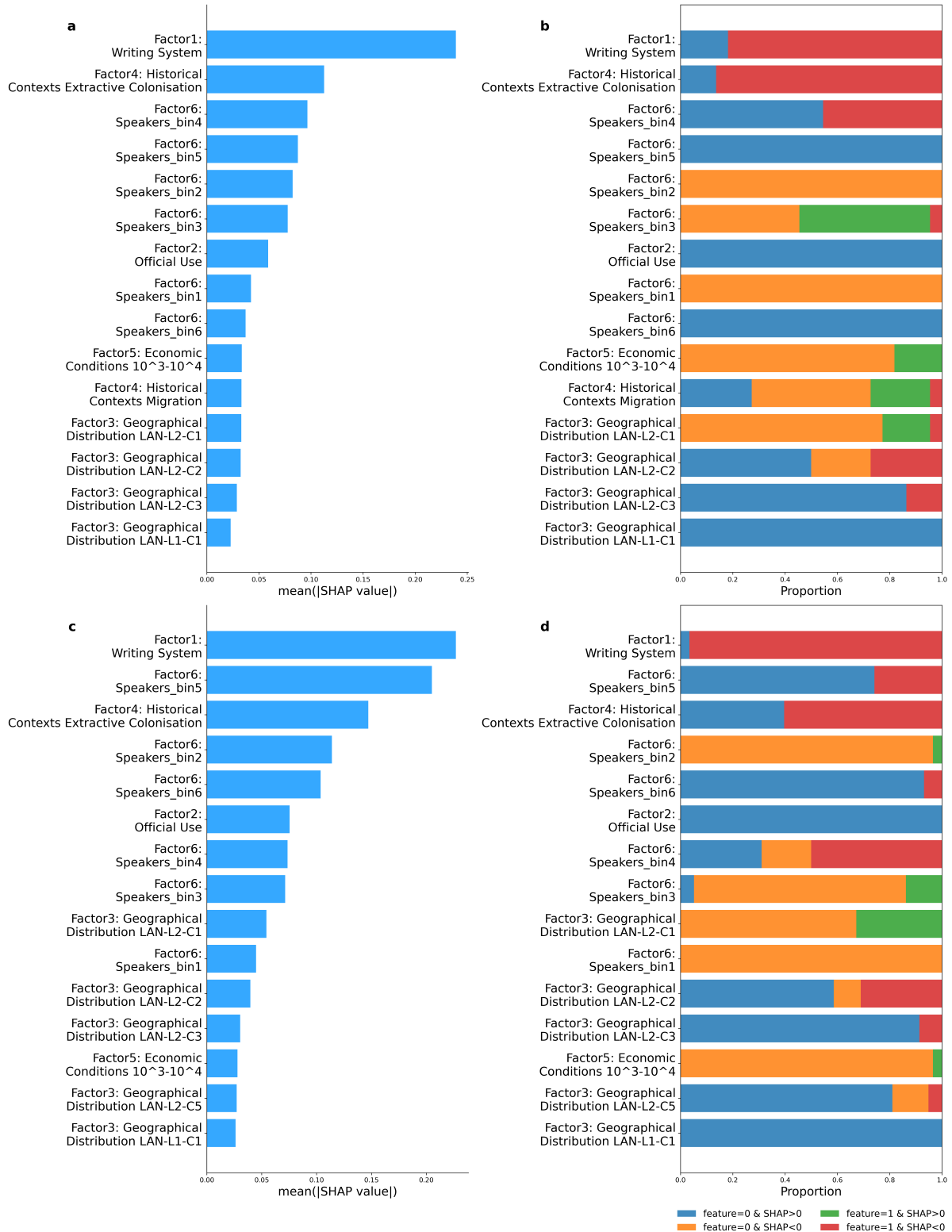


Figure S11. SHAP-based interpretation of outlier languages in the six-factor model. Panels **a** and **b** show the 22 upper outliers (more endangered than predicted), and panels **c** and **d** show the 58 lower outliers (less endangered than predicted). Panels **a** and **c** show the top 15 features ranked by mean absolute SHAP value. Panels **b** and **d** show, for the same features, the proportion of languages in each of four categories: feature absent with a positive SHAP contribution (feature=0 & SHAP>0), feature absent with a negative contribution (feature=0 & SHAP<0), feature present with a positive contribution (feature=1 & SHAP>0), and feature present with a negative contribution (feature=1 & SHAP<0). As in the five-factor model, writing system is the leading contributor, with extractive colonisation and binned speaker-population variables also featuring prominently among the top contributors.