Lexical Register, Not Cultural Proximity: Cross-Population Generalization of Schwartz Value Classifiers

Anonymous ACL submission

Abstract

The Schwartz system of universal human values (containing e.g., Achievement, Security, Tradition) has been designed and shown to be applicable across cultures. Yet classifiers trained to detect these values in texts show degraded performance when applied to texts from populations different from the training data. This paper investigates the reasons for the mismatch between theoretical universality and practical specialization using a series of cross-training experiments on four different new datasets of work goal descriptions. We show that classifiers can generalize across cultures when descriptions originate from similar contexts–in our case, generic employee samples collected from Indians (in India) or Asian-Americans. Conversely, we show that generalization can be poor even among descriptions from people of the same nationality–in our case, employees of a nonprofit organization and educators from an ultra-Orthodox community. Our focused error analysis reveals how these different populations use similar goal- and impact-oriented language, typically associated with Achievement value, to express several other values. Using the special case of work goals, this paper thus provides first evidence that for constructing training datasets for human value classifiers, the cross-cultural diversity is much less important than a diversity in life choices of the people behind the dataset texts.

1 Introduction

Pretrained value classifiers are typically evaluated on data drawn from the same distribution they were trained on.[1] Whether their competence transfers to other populations — speakers from other cultures, employees of organizations with their own linguistic conventions, members of communities whose value vocabulary differs from a generic training corpus — is much less clear. This question of cross-population performance has direct practical implications: value-detection tools are increasingly applied to free-text survey data in cross-cultural and organizational settings, where misaligned classifiers can silently distort inferred value profiles.

To make value classification performance across populations testable, we describe a generic setup for cross-population evaluation of value classifiers and then instantiate it for the Adam-Smith model of Schroter et al. (2023) — the top-performing system in SemEval-2023 Task 4 (ValueEval) — applied to short free-text responses from four populations chosen to vary along both *cultural* (geography, religion, language community) and *organizational* (workplace, religious-educational, lay-survey) axes: employees from India (**Indian**), Asian-American respondents (**Asian**), employees of an Israeli Jewish humanitarian nonprofit (**IJH**), and female educators from the Israeli ultra-Orthodox Jewish community (**Ultra**). The setup — per-population and combined fine-tuning, a complete cross-population transfer matrix, and a population-aware error analysis on the Schwartz circumplex — is reusable: others studying value detection in their own populations can plug their data in without designing a new evaluation framework.

We use *lexical register* to refer to the population-typical vocabulary and phrasing through which speakers articulate their work goals: workplace-strategic in the IJH nonprofit, communal-religious in the Ultra educators, brief and aspirational in the Asian and Indian survey respondents. Register is the surface signal a text classifier sees, and it tracks the *organizational* axis — the life choices that shape a speaker's daily linguistic environment — more closely than the *cultural* axis. Our research questions below test exactly this split: shared register without shared culture (Indian vs. Asian) and shared culture without shared register (IJH vs. Ultra).

We use the setup to ask two parallel questions

[1] Code: https://anonymous.4open.science/r/value-detection-adamsmith-eval-8513. Data: https://doi.org/10.5281/zenodo.20324552.

and find sharply different answers.

(1) Can classifiers fine-tuned on the Indian sample generalize to the Asian-American sample, and vice versa? They can: bilateral transfer reaches within 0.09–0.15 macro F1 of the in-domain ceiling, without fine-tuning to the target population.

(2) Can classifiers fine-tuned on the IJH sample generalize to the Ultra sample, and vice versa? They cannot: transfer collapses to indistinguishable from the off-the-shelf model in both directions, despite shared Israeli nationality. A targeted error analysis on each non-transferring population exposes *theoretically interpretable, mirror-image attractor patterns*: the model misreads IJH's prosocial motivations as Achievement, and Ultra's achievement-oriented motivations as Benevolence or Tradition. Both patterns cross Schwartz's higher-order axes in opposite directions; together they identify the underlying axis that matters for transfer as *lexical register* rather than cultural proximity. Finally, we show that fine-tuning on the heterogeneous union of all four populations yields a single model competitive with each population-specific specialist on its test set, providing a practical mitigation for the register-diversity gap surfaced by the error analysis.

2 Related Work

Schwartz value theory and value detection in NLP. Our work builds on Schwartz's (1992) theory of basic human values, which posits a set of cross-culturally recognizable motivational categories arranged on a circumplex along two higher-order dimensions (self-enhancement vs. self-transcendence, openness to change vs. conservation). The refined theory of Schwartz (2012) subdivides these into 19 finer-grained values and adds two new categories (Face and Humility), yielding a 12-value coarse scheme that subsumes the original 10. Computational identification of these values from text was operationalized at scale by SemEval-2023 Task 4 (Kiesel et al., 2023), in which systems classified short arguments according to the 20 fine-grained values of the refined theory. The Adam-Smith system of Schroter et al. (2023) achieved the top performance on that task using DeBERTa-based ensembles, and is the model whose cross-population behavior we study here. Subsequent value-detection work has continued to evaluate primarily on argumentation corpora and on probing language-model values directly (Arora et al., 2023; Yao et al., 2023; Ren et al., 2024); the question of how such classifiers generalize to free-text value statements from non-argumentative populations has received less attention. We use the circumplex structure to interpret classifier errors as well as the values themselves: the cross-axis distance between a model's predicted and annotated value (Section 6) is a theoretically grounded measure of how badly the model's mistake *misses*.

Cross-cultural and cross-population NLP. A growing body of work documents how NLP models behave differently for different populations of speakers and how this matters for downstream use. Hovy and Yang (2021) argue that NLP systems systematically under-model the social factors (demographic, cultural, organizational) that shape language production, and that this under-modeling produces measurable performance disparities. Hershcovich et al. (2022) extend this to a cross-cultural framing, identifying multiple axes (linguistic, common-sense, aboutness, objectives) along which NLP resources fail to transfer across cultures and recommending evaluation protocols that make these failures visible. Our experiments are in this spirit: we treat the four populations as distinct evaluation targets and report a complete cross-population transfer matrix rather than a single aggregate score. Where most cross-cultural NLP work expects cultural proximity to predict transfer, we find the opposite: transfer succeeds across nations when lexical register is shared, and fails within a nation when it is not (Sections 4–5).

Combined-data training as a robustness strategy. Gururangan et al. (2020) showed that exposing pre-trained models to a broader set of in-domain text consistently improves downstream performance, and that the gains generalize across tasks. The combined-data fine-tuning result we report (Section 7) is a finer-grained instance of the same pattern: training on a heterogeneous mixture of *populations* (rather than a single specialist) yields a model that is competitive with each specialist on its own data, at the cost of being best on none. This is a useful default precisely when the deployment population is unknown.

3 Testing Cross-Population Generalization of Value Classifiers

Pre-trained value classifiers are typically evaluated on a single corpus and a single annotation regime, so we have little systematic evidence about how they

behave when the population that produced the texts changes. Our setup has four ingredients: (i) multiple populations producing short, value-bearing free-text responses annotated under the same coarse Schwartz scheme; (ii) an off-the-shelf base classifier already trained on value detection (here, Adam-Smith on the SemEval-2023 Task 4 corpus); (iii) per-population fine-tuning of the base classifier on each population separately, plus a combined fine-tuning on the union of all populations; and (iv) a cross-population transfer matrix whose rows are training sources and columns are evaluation targets, every cell filled by running the corresponding fine-tuned model on the corresponding test set. Where transfer collapses, a population-aware error analysis examines the confusion structure on the Schwartz circumplex and the population-typical distinctive vocabulary that drives the errors. We next describe our instantiation of this procedure.

Corpus. We study short, free-text responses collected from four populations that differ in cultural background and organizational context: *Indian* (employees from India; the largest sample), *Asian* (Asian-American employees), *IJH* (employees of an Israeli Jewish humanitarian nonprofit), and *Ultra* (female educators from the Israeli ultra-Orthodox Jewish community). Together they yield 2,699 labeled instances (Anonymous Authors, 2025). The populations vary along both *cultural* (geography, religion, language community) and *organizational* (workplace, religious-educational, lay-survey) axes. In all four populations, each participant first completed an adapted Schwartz Values Survey (Schwartz, 1992) and then wrote up to three short sentences describing their work goals; the work-goal sentences are what we annotate and study (full survey instruments and elicitation prompt in Appendix F). Asian and Indian respondents wrote in English; IJH and Ultra responses were originally written in Hebrew and translated to English for modeling (see Limitations). Responses are short (median 3 words overall; range 2–5 across populations), and the label distribution is heavily skewed: Achievement, Benevolence, and Universalism together account for roughly two thirds of all annotations, with Achievement dominating Asian and Indian, Benevolence dominating Ultra, and Universalism most frequent in IJH. For full transparency, Appendix E shows the per-population sizes, label distributions, and length distributions of the data. The studies were approved by the ethics review board of the lead author's institution (anonymized for review).

Model, splits, and protocol. We use Adam-Smith (Schroter et al., 2023), the top-performing system in SemEval-2023 Task 4 (ValueEval), with no architectural modifications. Annotations use a 12-category coarse Schwartz scheme; Adam-Smith outputs over 20 fine-grained values, which we project to coarse at inference and use to define a multi-positive supervision target during fine-tuning (mechanics, hyperparameters, parameter count, and compute in Appendix A). A team of expert annotators trained in Schwartz value theory developed a detailed annotation guide; each goal was independently coded by two annotators and a curator resolved disagreements, with inter-annotator agreement on the dominant value ranging from 67% (Asian) to 88% (Indian; 75% for IJH and 83% for Ultra). For each population we form an 80/20 train/test split, stratified by coarse label, with a fixed random seed: Asian (321/81), Indian (928/233), IJH (332/83), Ultra (576/145). The *Combined* split is the union of the four (2,157/542), so combined evaluation respects the same per-instance partition as the per-population evaluations. We run five fine-tuning experiments using standard fine-tuning hyperparameters without hyperparameter search. The IJH and Combined runs (the lowest in-domain ceiling and the proposed robust default) are repeated with three random seeds (42, 43, 44) and reported as mean $\pm$ std; the three remaining single-population runs (Asian, Indian, Ultra) use seed 42 only. We report macro F1 (primary, given heavy class imbalance) and micro F1 in the 12-category coarse space, computed from top-1 predictions at the final fine-tuning epoch with no model selection on the test set. The Combined per-epoch learning curve is shown in Figure 3 (for completeness, Appendix B shows the four per-population curves). Applied without fine-tuning, the base Adam-Smith model reaches overall top-1 coarse accuracy of 53.3% and macro F1 ranging from 0.19 (IJH) to 0.40 (Asian). Appendix D provides the full per-population base-model breakdown.

4 Generalization Across the Asian and Indian Samples

Following the cross-cultural NLP literature, our prior is that classifiers fine-tuned on one population do not generalize to another: a model fine-tuned on Indian responses should perform substantially

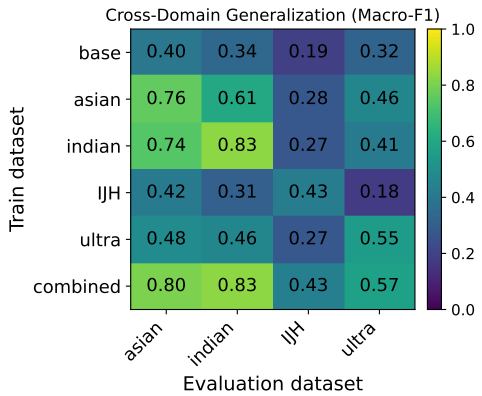

Figure 1: Cross-population transfer matrix. Rows are training sources (the *base* row is the un-tuned Adam-Smith checkpoint); columns are evaluation targets (the per-population 20% test sets). Cells report macro F1 at the final fine-tuning epoch (seed 42 for the Asian, Indian, and Ultra rows; mean over seeds 42/43/44 for the IJH row; Section 3).

worse on Asian-American responses than on its own held-out test set, and vice versa. The Indian and Asian samples are a useful test because their respondents differ in nationality but produce structurally similar texts (short, English, work-goal descriptions; Section 3).

Both populations attain strong in-domain macro F1 (Figure 1 diagonal). Indian reaches 0.83 by epoch 10 with stable training and no overfitting visible (Figure 5). Asian reaches 0.76 at epoch 10, peaking at 0.84 around epochs 5–6 and decaying slightly thereafter (Figure 4). Both are roughly 2× the corresponding base-row entries (Asian 0.40, Indian 0.34; Table 1).

Restricting to the Asian–Indian sub-matrix: a model fine-tuned on Indian achieves 0.74 macro F1 on Asian, within 0.09 of its in-domain performance (0.83). The reverse direction is weaker: a model fine-tuned on Asian achieves 0.61 on Indian, 0.15 below its own in-domain performance (0.76). Both directions remain well above the base entries for their columns (0.40 Asian, 0.34 Indian), and exceed every other off-diagonal cell in the matrix.

Discussion: hypothesis rejected. Bilateral transfer between Asian and Indian is substantial: a model trained on one transfers to the other within 0.09–0.15 macro F1 of the in-domain ceiling, despite no exposure to the target population during fine-tuning. The directional asymmetry — Indian → Asian is stronger than Asian → Indian — is consistent with the size difference between the two training sets (Asian n=321 vs. Indian n=928): the smaller Asian-trained model has not seen enough vocabulary to cover Indian's larger and more diverse responses, but the larger Indian-trained model has effectively encompassed Asian's. The two populations therefore share enough *lexical register* for a classifier to learn one and apply it to the other, even though they differ in nationality and demographic composition. The standard cross-cultural-transfer pessimism does not hold here.

5 Generalization Across the IJH and Ultra Samples

By the same reasoning as in Section 4, we expect classifiers fine-tuned on one population to fail on another. The IJH and Ultra samples are a useful test in the opposite direction from Section 4: their respondents share a nationality (both are Israeli, with original responses in Hebrew) but produce structurally and lexically very different texts: workplace-strategic versus communal-religious. If transfer is driven by cultural proximity, transfer here should be substantial.

IJH reaches macro F1 0.43 ± 0.06 at epoch 10 across three seeds, despite near-perfect training F1 within a few epochs; the in-domain test gap therefore reflects properties of the dataset, not under-training (Figure 6). The seed-to-seed std (0.06) is roughly 4× the Combined run's variability (Section 7), making IJH the most unstable of the four configurations. Ultra reaches 0.55 at epoch 10, peaking at 0.72 around epoch 4 and decaying steadily thereafter, the most pronounced overfitting trajectory among the four populations (Figure 7).

On the IJH–Ultra sub-matrix of Figure 1: IJH → Ultra is 0.18 ± 0.01, lower than the base-row entry for the Ultra column (0.32); training on IJH *hurts* Ultra performance relative to the off-the-shelf model. The reverse, Ultra → IJH, is 0.27, 0.16 below IJH's own in-domain mean (0.43) and indistinguishable from the base-row entry for the IJH column (0.19) within measurement error. IJH's off-diagonal transfer is similarly flat for the other two populations (0.42 ± 0.03 on Asian, 0.31 ± 0.02 on Indian; IJH's own in-domain is 0.43 ± 0.06), and the base-row entries for the IJH and Ultra columns (0.19 and 0.32 respectively) are the two lowest in the entire matrix.

Discussion: hypothesis accepted. Transfer between IJH and Ultra is essentially absent in both directions, and shared Israeli nationality has bought

the two populations no cross-population competence. The failure is not a model-capacity or data-size effect: IJH is hard as a *target* from every training source in the matrix, and Ultra's specialist learns something that transfers nowhere. The two populations behave as if they were annotating different value schemes, despite using the same coarse Schwartz labels. We argue in Section 6 that this is a consequence of each population speaking the value scheme through a distinct, population-typical lexical register — workplace-strategic for IJH and communal-religious for Ultra — and that the registers are far enough apart on the Schwartz circumplex that a model learned on one systematically mislabels the other.

6 Error Analysis

Section 5 established that fine-tuning plateaus on IJH and that no training source in the matrix transfers well to either IJH or Ultra. We now examine why. We focus first on the IJH in-domain test set under the seed-42 epoch-10 fine-tuned checkpoint, the configuration whose mean macro F1 of 0.43 ± 0.06 across three seeds is the lowest in-domain mean in the transfer matrix. On this seed the model is correct on 56 of 83 instances (overall accuracy 67.5%; weighted F1 0.64; macro F1 0.35 over the 12 coarse Schwartz categories, or 0.47 if averaged only over the 9 classes with non-zero support (matching the matrix's 9-class macro-F1 convention; the matrix's 0.43 for this cell is the corresponding multi-seed mean). The large gap between the weighted and macro averages reflects strong performance on a few well-represented classes and total failure on several minority ones.

Achievement is a systematic attractor. The dominant error pattern is a one-way confusion into Achievement (Figure 2): nine Universalism instances, three Self-Direction, three Security, two Benevolence, and one each of Power and Stimulation are predicted as Achievement (19 of 27 total errors). The reverse direction is rarer: only three Achievement instances escape, two into Universalism and one into Self-Direction. Sample errors illustrate the mechanism: *"To encourage growth in people"* (BE → AC), *"Influence on policy"* (UN → AC), *"Promoting fairness in public systems"* (UN → AC), *"Trust-based organization"* (BE → AC), *"Promote significant social change, especially for populations in need"* (UN → AC). In each case, a prosocial or growth-oriented motivation is expressed in performance- or advancement-oriented language. The errors are not local on the Schwartz circumplex: 18 of the 19 collapse-into-Achievement errors cross at least one of Schwartz's higher-order axes (most often self-transcendence → self-enhancement, occasionally conservation → self-enhancement), and across all 27 errors the mean circumplex distance between annotated and predicted class is 3.6 out of a possible 6, with 92.6% of errors crossing a higher-order axis. Readers interested in the performance of the base model (which shows the same Achievement-attractor) are referred to Appendix D. For completeness, all 27 errors are tabulated in Appendix G.

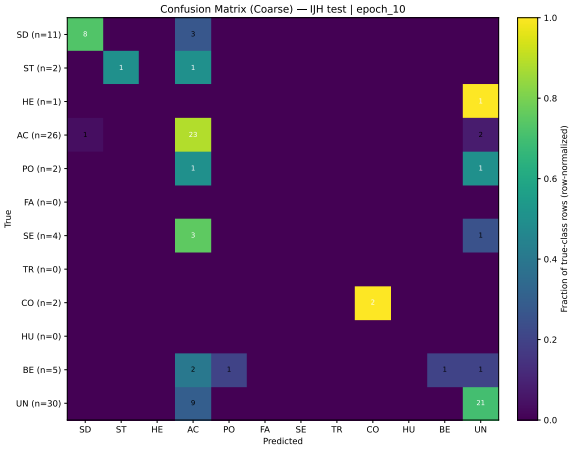

Figure 2: Confusion matrix on the IJH in-domain test set at fine-tuning epoch 10. Achievement (AC) absorbs misclassifications from semantically adjacent classes; Face (FA) and Humility (HU) have no support.

Direct vocabulary ablation. We directly test whether a compact set of achievement-associated tokens (e.g., *achieve*, *impact*, *improve*, *advance*, with morphological variants; 22 tokens total) causally drives the model's over-prediction of Achievement. Across all five settings on the full 2,699-instance corpus (the base model and each of the four per-population fine-tuned checkpoints), masking this vocabulary with the tokenizer's `[MASK]` token produces only marginal shifts in Achievement prediction share ($|\Delta| \leq 2.1$pp; base -1.4pp, Asian -0.9, Indian -0.6, IJH -0.1, Ultra -2.1). The largest shift, on the Ultra checkpoint, is still an order of magnitude smaller than the cross-population spread of Achievement-prediction share (IJH FT: 56.5% vs. Ultra FT: 21.4%) and well below the magnitude of the IJH AC-attractor at the error level. On the base model, masking also reduces Achievement recall (Δ

−3.8pp) and overall F1, indicating that these tokens function as genuine informative cues rather than spurious shortcuts. IJH FT is the *least* sensitive of the five settings (Δ AC share −0.1pp), consistent with IJH FT having internalized the population-typical pattern at a level beyond any specific token (cf. Figure 8 in Appendix C for a visual impression). Taken together, the five-way ablation rules out a simple lexical explanation for the Achievement attractor: the bias must arise from broader semantic and distributional factors that survive the removal of the most obvious achievement-coded tokens. The corresponding result on the Ultra side (whether masking achievement vocabulary affects the TR/BE attractor) is by definition smaller still, since the Ultra attractor is on the opposite circumplex pole; a parallel ablation of caring / values vocabulary on Ultra is left to future work.

The mirror attractor on Ultra. The same population-aware analysis applied to Ultra's in-domain test set at seed-42 epoch 10 exposes a structurally similar but *topographically inverted* attractor pattern (cf. Figure 9 in Appendix C for a visualization). Of 54 misclassifications, 30 (56%) collapse into the Tradition + Benevolence pair: 16 into Tradition (mostly from BE and AC) and 14 into Benevolence (mostly from AC). Compared to IJH's single-class collapse into Achievement, Ultra's attractor is dual-class: a Tradition arm absorbs errors from BE, AC, UN, and SE; a Benevolence arm absorbs errors from AC, SE, SD, and ST. Cross-axis confusion is again the dominant pattern: 85.2% of Ultra errors cross at least one of Schwartz's higher-order axes, and the mean circumplex distance between annotated and predicted class is 3.3 out of 6 (vs. IJH's 3.6). The two attractor patterns sit on *opposite poles of the Schwartz circumplex*: IJH pulls errors into the self-enhancement quadrant (AC) from across the circle, while Ultra pulls errors into the conservation / self-transcendence quadrants (TR and BE) from across the circle. The same model, applied to the same value scheme, makes systematically opposite errors depending on which population produced the text.

A population-typical lexical register drives the Ultra attractor. We verify the register-level explanation directly by extracting the most distinctive content tokens of each (population, value) cell relative to the same value class in the other populations, using log-odds with an informative Dirichlet prior (Monroe et al., 2008). The full table is reported in Appendix I; the relevant pattern for Ultra is that the BE class and the TR class share five distinctive content tokens (*children, education, love, students, values*), even though their non-overlapping distinctive vocabularies are otherwise very different (Ultra-TR additionally surfaces *torah, jewish, god, virtues, girls*, none of which appear in the BE list). Concretely, the BE-annotated instances that collapse into TR draw on this shared subvocabulary (e.g., *"To educate in the right values"*, *"Instilling values and virtues"*, *"Educate the next generation to be good people today"*, *"Strengthen their values and resilience"*); all seven items are annotated BE but predicted TR at circumplex distance 3. The model interprets the conservative-religious vocabulary as evidence for Tradition rather than for the underlying caring motivation. The reverse direction — AC-annotated responses predicted BE — shows a complementary pattern in which achievement-oriented motivations are articulated through education / caring vocabulary (e.g., *"Significant knowledge"*, *"Professional knowledge"*, *"Be effective and promote children"*, *"That they will succeed in the matriculation exams"*). Both directions are tabulated in Appendix H.

The same mechanism drives IJH's collapse into Achievement. The IJH UN and IJH AC cells share three of the top distinctive content tokens (*promoting, significant, social*; Appendix I), and the remaining top IJH-AC tokens are themselves achievement-coded (*impact, influence, creating, processes, organization, increasing*). The boundary the model has to draw between IJH-UN and IJH-AC therefore runs through a vocabulary where prosocial advocacy and personal-achievement framing are partially indistinguishable at the surface, and the model resolves the ambiguity toward Achievement.

Cross-population synthesis. The same model exhibits two systematic, theoretically interpretable failure modes on different populations: IJH's prosocial language is misread as Achievement, and Ultra's achievement-oriented language is misread as Benevolence or Tradition. Both failure modes traverse the Schwartz circumplex in the cross-axis direction predicted by the population's typical lexical register. The implication for training-data construction is direct: cross-cultural diversity is not the relevant axis for value-classifier robustness; *lexical-register* diversity is. A training set that contains many cultures but only one register will not generalize across populations whose registers differ, while a training

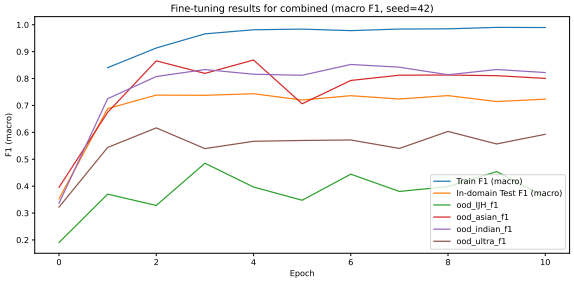

Figure 3: Fine-tuning on the combined dataset (seed 42). Training F1 saturates early; in-domain test stabilizes near 0.73. Out-of-domain F1 on the four constituent populations is competitive with their respective per-population specialists. Epoch 0 is the un-tuned base Adam-Smith model.

set that spans the registers will (Section 7).

7 Combined-Population Training as a Robust Default

If population-typical lexical registers are what break cross-population transfer (Sections 5, 6), the simplest mitigation is to train the classifier on a heterogeneous mixture of populations and let it see all of the registers at once. We fine-tune a single model on the union of all four populations (2,157 train / 542 test) and evaluate on every per-population test set.

Across three seeds, Combined achieves 0.80 ± 0.02 on Asian, 0.83 ± 0.04 on Indian, 0.43 ± 0.09 on IJH, and 0.57 ± 0.02 on Ultra. Compared to each population's specialist on its own test set, Combined *exceeds* the specialist for Asian (0.80 vs. 0.76) and Ultra (0.57 vs. 0.55), *matches* it for IJH (0.43 vs. 0.43), and is essentially tied for Indian (0.83 vs. 0.83). Combined's variance is also consistently tight (std 0.02–0.04 on Asian, Indian, and Ultra), except where every training source is unstable (IJH, std 0.09). The combined-training trajectory is shown in Figure 3.

The Combined model is therefore a strong default when the deployment population is unknown: it is competitive with every population-specific specialist on its own test set, and the standard cross-cultural-transfer pessimism (that a single model cannot serve multiple populations) does not hold once the union covers the populations' distinct lexical registers. The remaining gap on IJH (which Combined matches but does not exceed) is consistent with IJH's high seed variance and the structural difficulty discussed in Section 6: no source of training data, including the union, fully bridges IJH's workplace-strategic register.

8 Discussion

Two findings cut across our experiments. First, value expression is population-specific in ways that limit cross-population transfer; second, this specificity appears to arise largely from *vocabulary in context* rather than from the underlying value categories themselves, with consequences both substantive and practical.

The cross-domain heatmap (Figure 1) shows that transfer is bilateral only between the Asian and Indian samples (macro F1 0.61–0.74, within 0.09–0.15 of the in-domain ceiling in each direction), the two populations whose responses are short, English, and ambition-laden. The IJH and Ultra samples each transfer poorly in both directions despite all four populations annotating the same Schwartz scheme. This pattern is not well explained by "cultural distance" alone: the Asian and Indian samples are themselves culturally heterogeneous, and IJH and Ultra are both Israeli. A more parsimonious account is that each population's responses reflect a distinctive *lexical register*: workplace-strategic in IJH, communal-religious in Ultra, brief-aspirational in Asian and Indian. Transfer is strongest where lexical registers align, regardless of the demographic label on the underlying sample.

The IJH and Ultra cases are particularly informative because their failure modes mirror each other on the Schwartz circumplex (Section 6). IJH's prosocial and security-oriented motivations are frequently expressed in goal- and impact-oriented language, leading the model to systematically favor Achievement; Ultra's achievement-oriented motivations are frequently expressed in values-keeping and educational language, leading the model to systematically favor Benevolence or Tradition. Each population's error structure is theoretically interpretable as a cross-axis collapse driven by a population-typical lexical register, and the two registers sit on opposite poles of the circumplex.

Interpretation in Schwartz's framework. The mirror attractor structure has a natural reading in value theory. Schwartz's universal claim is about the *presence and structure* of values — that any population can be described along the same 12-category circumplex with the same two higher-order axes (Schwartz, 2012) — not about the *language* through which any particular value is articulated. Within the circumplex, adjacent values share motivational

goals (a compatibility relation), and circumplex-opposite values represent motivational conflict. Our cross-axis attractor patterns traverse precisely these conflict relations: IJH's UN → AC collapse pulls self-transcendence motivations into the self-enhancement quadrant, and Ultra's AC → BE/TR collapses pull self-enhancement motivations into the self-transcendence and conservation quadrants. The classifier is not making local, adjacency-level confusions; it is blurring theoretically distinct motivations along the very axes Schwartz's theory uses to organize them.

This is also consistent with well-known population-level value-priority differences in the Schwartz literature: religious and traditional populations score lower on self-enhancement and higher on conservation, while institutional contexts (e.g., nonprofit workplaces) reshape how self-transcendence is articulated because the organizational logic is itself performance-based. From this angle, the attractors are less a pure model failure than the classifier accurately picking up on *how* a given population realizes the universal value scheme, doing so against the annotators' intent because the annotators encoded the underlying motivation rather than the surface vocabulary. The label disagreement is an interaction between two valid signals: the population's lexical register, and the value-theoretic content the annotators encoded.

Implication: register-diverse training data. The diagnosis points to a concrete solution that follows directly from the theoretical interpretation: *training data for a value classifier must span the lexical registers the deployed model is expected to handle*, not merely the cultures or organizations. A model fine-tuned on a single population becomes a specialist in that population's register and inherits a register-specific decision boundary; a model fine-tuned on the heterogeneous union of populations approaches each population's specialist on its own test set (Section 7) at the cost of being best at none. When the deployment population is known, specialist training is justified; when it is uncertain (typical in cross-cultural or cross-organizational deployments of value-detection tools), combined training is the more robust strategy precisely because it spans the registers. We expect the register-diversity principle to generalize beyond Schwartz value detection: any classifier whose label space is defined by motivational or attitudinal content (rather than by surface form) is likely to inherit a register-specific decision boundary from its training data and to fail in the cross-axis direction predicted by the source-target register difference.

9 Conclusion

We evaluated a state-of-the-art value classifier on free-text responses from four culturally and organizationally distinct populations and examined how performance changes under fine-tuning and cross-population transfer. We find that transfer is not well explained by cultural proximity but aligns with similarities in lexical register: populations that express values using similar linguistic patterns transfer well, while those with distinct registers do not.

Two complementary, theoretically interpretable failure modes emerge. IJH's prosocial responses are absorbed by Achievement (self-enhancement pole); Ultra's achievement-oriented responses are absorbed by Benevolence and Tradition (conservation / self-transcendence pole). Both patterns cross Schwartz's higher-order axes in opposite directions, consistent with each population realizing the universal value scheme through a population-typical register that the classifier inherits from training.

The practical implication is direct: *training data for value classifiers must span the lexical registers of the deployment populations*, not merely their cultures or organizations. We show that fine-tuning on the heterogeneous union of populations is a simple, strong default that matches each population's specialist on its own test set: a register-diversity mitigation rather than a model-architecture one. The natural next steps are to validate the register-level diagnosis in contexts beyond work-goal descriptions (the generic setup of Section 3 is reusable for this), to test whether population-aware architectures can better preserve cross-circumplex distinctions when register diversity in the training data is limited, and to investigate whether the cross-axis attractor effect is primarily driven by class imbalance, embedding-space geometry, or broader semantic overlap. The headline insight, we believe, generalizes beyond Schwartz value detection: when a label space is defined by motivational content but realized through population-typical language, register diversity in the training data is what determines cross-population robustness.

Limitations

- **Translation effects.** Responses from two populations (IJH and Ultra) were originally

authored in Hebrew and translated into English using a machine translation pipeline with manual quality control. Translation may alter lexical cues relevant for value classification and introduce noise that is difficult to disentangle from genuine population-specific linguistic patterns. This is particularly relevant for interpreting the lower performance observed on the IJH dataset.

- **Limited per-class support.** Several value categories have very small representation in some populations, particularly in the IJH test set. As a result, macro F1 is sensitive to a small number of predictions, and per-class performance in the long tail should be interpreted with caution.
- **Partial multi-seed coverage.** The IJH and Combined runs are reported as mean $\pm$ std across three random seeds (42, 43, 44); the remaining three single-population runs (Asian, Indian, Ultra) use seed 42 only. Differences below $\sim$0.05 macro F1 in single-seed runs may not be statistically reliable; for the multi-seeded runs the std (0.01–0.09) is reported and should bound interpretation directly.
- **Label space alignment.** Although we align coarse annotations with fine-grained model outputs via a hierarchical mapping, this introduces an additional modeling assumption. Alternative formulations (e.g., predicting directly in the coarse space or using hierarchical losses) may yield different performance characteristics.

Ethical Considerations

Ethics approval. All four studies were approved by the institutional ethics review board of the lead author's institution. The committee name and per-study approval numbers are withheld for anonymous review and will be disclosed in the camera-ready version.

Recruitment, consent, and compensation. Participation was voluntary, anonymous, and obtained with informed consent in all four samples. The Asian-American (N=180; 44% female; mean age 28) and Indian (N=782; 33% female; mean age 33) samples were recruited through professional survey panels (Prolific and C-Voter, respectively) and compensated at the platforms' standard rates; an attention check excluded inattentive respondents. IJH respondents (N=238, a 73% response rate from a 328-employee organization; 86% female; mean age 47) were recruited via internal email and on-site QR codes; participation was incentivized through personalized feedback on participants' value profiles and a lottery for small in-kind prizes. Ultra-Orthodox respondents (N=400, all Haredi women employed in education; mean age 34) were recruited via snowball sampling through a digital magazine distributed by email to women in the ultra-Orthodox sector and incentivized through a lottery for a gift voucher (200 NIS). Both Hebrew-speaking samples were administered in Hebrew, the dominant working language of the respective populations; translation to English for modeling is described in Section 3 and revisited in the Limitations.

Annotator team. Annotation was conducted by a team of four to six researchers (a mix of graduate students, post-doctoral researchers, and faculty) with specialized knowledge of the Schwartz value theory; all were Hebrew-native bilingual English speakers, allowing fluent access to both the English source materials (Asian and Indian samples) and the Hebrew originals (IJH and Ultra samples). The team developed a detailed annotation guide collaboratively and refined it iteratively across the four populations. Each goal was independently coded by two annotators for the dominant Schwartz value expressed, and a curator reviewed all annotations and resolved disagreements by consulting the guide and the broader team (Section 3).

Screening for personally-identifying or offensive content. By design, the elicitation prompt asks participants to describe professional priorities rather than personal experiences; responses are correspondingly short (median 3 words; Section 3) and concern work goals rather than personal narrative. Because each goal was independently reviewed by two trained annotators and a curator during annotation, this review pass also functioned as a screening pass for personally-identifying information (names, email addresses, identifiable workplaces) and offensive content; no such cases were flagged during review.

Sensitive populations and potential harms. Two of the four populations we study have lexical registers that are under-represented in the argumentation data on which off-the-shelf value classifiers are trained. Ultra is an ultra-Orthodox minority community whose communal-religious register is

rarely seen in standard NLP corpora; IJH is distinctive in being the workforce of an Israeli Jewish humanitarian nonprofit, whose institutional register (workplace-strategic, mission-driven) likewise differs from typical training data. Our central empirical finding, that cross-population transfer of value classifiers tracks lexical register rather than cultural proximity (Section 8), has a direct ethical implication: applying a classifier trained predominantly on argumentative or workplace-strategic registers to populations with different registers (e.g., communal-religious, ultra-Orthodox, or other under-represented minorities) will systematically distort the inferred value profile of those populations in the cross-axis direction we document. The mirror-image attractor patterns of Section 6 make this concrete: communal-religious responses are systematically misread as Achievement, and achievement-oriented responses from religiously-conservative populations are systematically misread as Benevolence or Tradition. Practitioners should not deploy value classifiers cross-culturally without explicit register coverage in the training data, and aggregate cross-population value comparisons drawn from a single-register classifier should be treated as unreliable.

Translation. IJH and Ultra responses were originally written in Hebrew and translated to English via Google Translate with manual quality control prior to modeling. Translation may erase or distort culturally-specific lexical cues, particularly for communal and religious vocabulary, and is treated as a methodological limitation in the Limitations section.

Data release. The annotated corpus is released on Zenodo (DOI: `https://doi.org/10.5281/zenodo.20324552`) under CC-BY-4.0 for research use. The release contains the 2,699 annotated work-goal responses, the 12-category coarse Schwartz labels, and the 80/20 per-population train/test splits used in this paper. Population identifiers in the released data are anonymized (*Indian*, *Asian*, *IJH*, *Ultra*); the IJH and Ultra responses are distributed as the English translations used for modeling (Hebrew originals are retained by the authors and not redistributed). The two-annotator + curator review pass functioned as a screen for personally-identifying information (Section above); no such cases were flagged.

Use of pre-existing artifacts. The Adam-Smith model (Schroter et al., 2023), released on HuggingFace as `tum-nlp/Deberta_Human_Value_Detector` under the OpenRAIL++ license, and the SemEval-2023 Task 4 ValueEval corpus (Kiesel et al., 2023), released on Zenodo (record 7550385) under the CC-BY-4.0 license, are used in accordance with their respective licenses. Our use, evaluating cross-population generalization for value classification, is consistent with the artifacts' intended scientific use.

AI assistance. The authors used AI assistants (large-language-model coding tools and writing aids) during this project for two purposes: drafting and editing supporting code (fine-tuning, evaluation, error analysis, and figure generation), and drafting, paraphrasing, and copy-editing portions of the paper prose. All scientific claims, experimental design, statistical analyses, and final wording were reviewed and approved by the human authors, who take full responsibility for the content. AI assistants are not credited as authors, in accordance with the ACL Generative AI Policy.

References

Anonymous Authors. 2025. Schwartz value annotations on cross-population work-goal responses (indian, asian, ijh, ultra).

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. *Preprint*, arXiv:2406.04214. Proceedings of ACL 2024.

Daniel Schroter, Daryna Dementieva, and Georg Groh. 2023. Adam-smith at SemEval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 532–541, Toronto, Canada. Association for Computational Linguistics.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Academic Press.

Shalom H. Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1).

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values — a survey of alignment goals for big models. *Preprint*, arXiv:2308.12014.

Appendix

A Implementation Details

Model and checkpoint. Adam-Smith (Schroter et al., 2023) is a DeBERTa-large-based classifier (405.2M parameters) with a multi-label sigmoid head over the 20 fine-grained values of Schwartz's refined theory; it was originally trained on the SemEval-2023 Task 4 corpus of stance-bearing arguments. We use the public checkpoint released on HuggingFace as `tum-nlp/Deberta_Human_Value_Detector` with no architectural modifications.

Coarse–fine label projection and multi-positive supervision. We define a fixed surjection from each of the model's 20 fine-grained values to its parent coarse value in the 12-category coarse Schwartz scheme used in our annotations (e.g., both *Self-direction: thought* and *Self-direction: action* map to SD; the four Universalism sub-labels all map to UN). At inference time, predictions are made in the fine label space and projected to the coarse level by selecting the parent of the highest-scoring fine label. During fine-tuning, each coarse-labeled instance is supervised by assigning positive targets to *all* fine-grained labels corresponding to its coarse category, retaining the model's original multi-label binary cross-entropy objective. This multi-positive formulation reflects the hierarchical structure of the label space and avoids introducing arbitrary asymmetries across coarse categories.

Hyperparameters and per-epoch evaluation. All runs use identical hyperparameters: AdamW with learning rate 2×10^{-5}, micro-batch size 2 with gradient accumulation over 4 steps (effective batch size 8), maximum sequence length 256 tokens, and 10 training epochs with no early stopping. These hyperparameter values are standard defaults for transformer fine-tuning; no hyperparameter search was performed (on any test set or otherwise) for this study. After every epoch we evaluate on (i) the run's own training set, (ii) its held-out in-domain test set, and (iii) the held-out test sets of all other populations as out-of-domain probes.

Compute. All experiments run on a single Apple M4 GPU (Mac mini, MPS backend). Total fine-tuning wall-clock across the nine runs is approximately 12 hours (Asian 25 min, Indian 74 min, IJH $3\times\sim$26 min, Ultra 46 min, Combined $3\times\sim$163 min); per-epoch evaluation adds modest overhead.

B Per-Population Fine-Tuning Curves

We here show the per-epoch learning curves of the four per-population fine-tuning runs not shown in the main part of the paper (Asian, Indian, IJH, Ultra); the Combined run is shown in the main part (Figure 3). Each plot reports macro and micro F1 on the run's training set, its in-domain test set, and the held-out test sets of the other three populations as out-of-domain (OOD) probes; epoch 0 is the un-tuned base Adam-Smith model. Strongest OOD transfer is Indian $\rightarrow$ Asian (0.74; Figure 5) and

Asian → Indian (0.61; Figure 4).

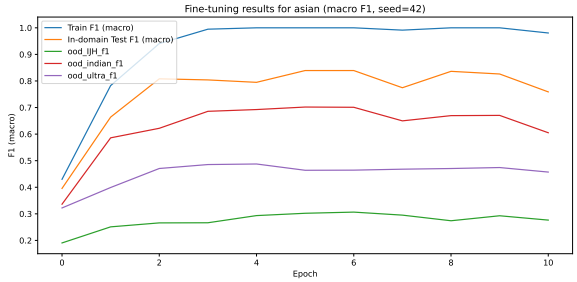

Figure 4: Per-epoch fine-tuning macro/micro F1 for the Asian single-population run (seed 42). In-domain peak 0.84 at epoch 5–6; final 0.76.

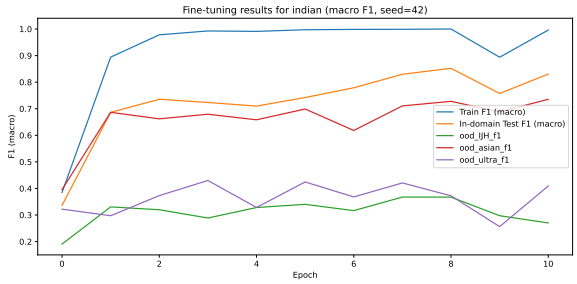

Figure 5: Per-epoch fine-tuning macro/micro F1 for the Indian single-population run (seed 42). In-domain stable at 0.83; no overfitting.

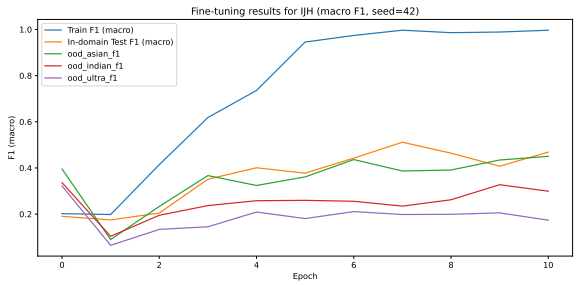

Figure 6: Fine-tuning on IJH (seed 42). Training F1 saturates above 0.9, but in-domain test F1 is capped near 0.35–0.47 throughout, suggesting a performance ceiling driven by dataset characteristics rather than under-training. Epoch 0 is the un-tuned base model (IJH in-domain F1 0.19).

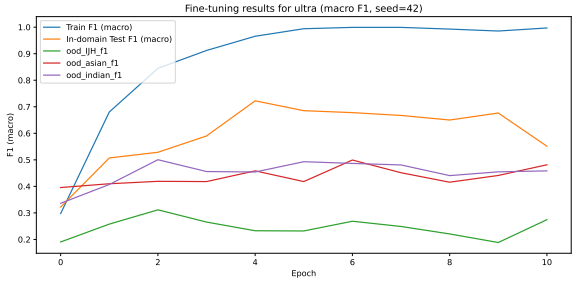

Figure 7: Per-epoch fine-tuning macro/micro F1 for the Ultra single-population run (seed 42). In-domain peak 0.72 at epoch 4; final 0.55, the most pronounced overfitting among the four populations.

C Additional Error-Analysis Figures

For a visual impression, we here show the numbers underlying the error analysis of Section 6 as matrices: the achievement-vocabulary ablation diff (Figure 8) and the Ultra confusion matrix mirroring the IJH AC-attractor (Figure 9).

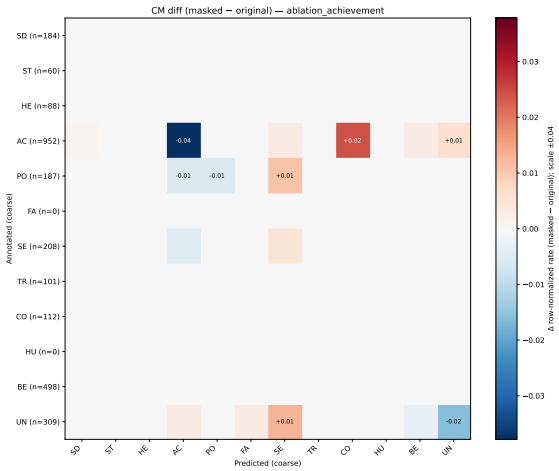

Figure 8: Base Adam-Smith confusion-matrix shift induced by masking the 22-token achievement vocabulary on the full 2,699-instance corpus. Cells are row-normalized rate differences (masked − original); blue indicates predictions that decreased after masking, red indicates predictions that increased. Off-diagonal cells in the AC column lose mass and the rows that previously collapsed into AC redistribute slightly elsewhere, but the magnitude is small ($|\Delta| \leq 1.4$pp on AC-prediction share globally), consistent with the conclusion that vocabulary alone does not account for the AC over-prediction.

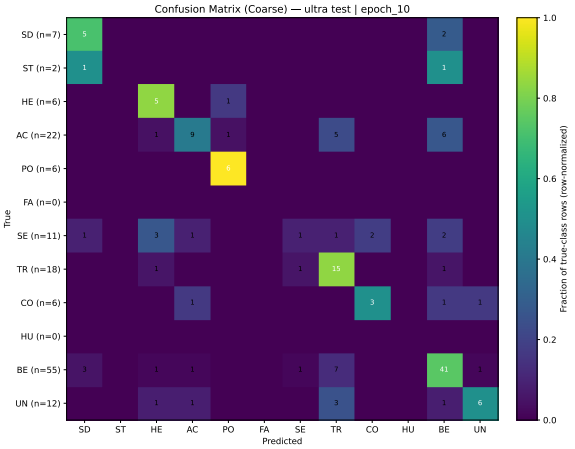

Figure 9: Confusion matrix on the Ultra in-domain test set at fine-tuning epoch 10. Tradition (TR) and Benevolence (BE) jointly absorb most misclassifications, mirroring the IJH Achievement-attractor on the opposite pole of the Schwartz circumplex.

D Base Model: Per-Population Performance

Applied without fine-tuning to the full corpus, Adam-Smith reaches an overall top-1 coarse accuracy of 53.3% and a hit@3 of 81.5%. The gap is informative: the correct value is among the model's top three predictions in over four-fifths of cases, even though the single best prediction is correct only about half the time. Top-1 accuracy varies substantially by population (Table 1): Asian 60.9%, Indian 60.5%, IJH 48.9%, Ultra 40.1%. The Indian and Asian samples, which contain comparatively short, ambition-laden responses, align most closely with the SemEval argument data the model was trained on; the Ultra sample, with its prosocial/communal vocabulary, aligns least.

Population	N	Top-1	Macro F1
Asian	402	60.9	0.40
Indian	1,161	60.5	0.34
IJH	415	48.9	0.19
Ultra	721	40.1	0.32
Overall	2,699	53.3	—

Table 1: Base Adam-Smith model on the full corpus. Top-1 in %, macro F1 on the per-population 20% test set (matching the *base* row of Figure 1).

We additionally provide three views of the un-tuned Adam-Smith model on the full 2,699-instance corpus: (i) annotated and predicted coarse-label distributions (Figures 10–12); (ii) the overall annotated-vs-predicted confusion matrix (Figure 13); and (iii) the same matrix broken down by source population (Figures 14–17). The Achievement column dominates each view, motivating the AC-attractor analysis in Section 6.

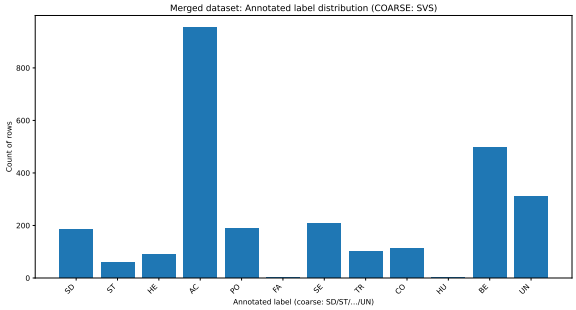

Figure 10: Annotated coarse-label distribution on the full 2,699-instance corpus.

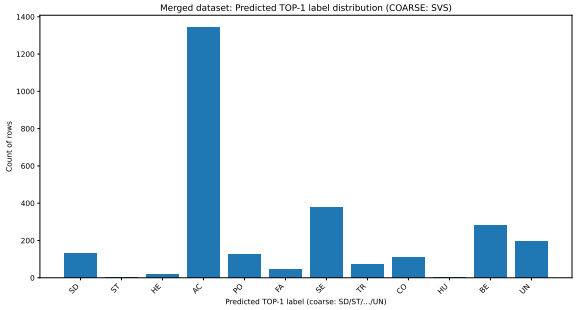

Figure 11: Base Adam-Smith predicted top-1 coarse-label distribution. Disproportionate over-prediction of Achievement relative to the annotated distribution (Figure 10) is the source of the AC-attractor described in Section 6.

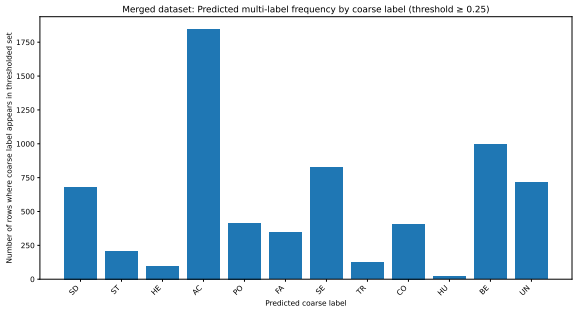

Figure 12: Base Adam-Smith multi-label predicted frequency at threshold 0.25. The over-prediction of Achievement persists under the multi-label view.

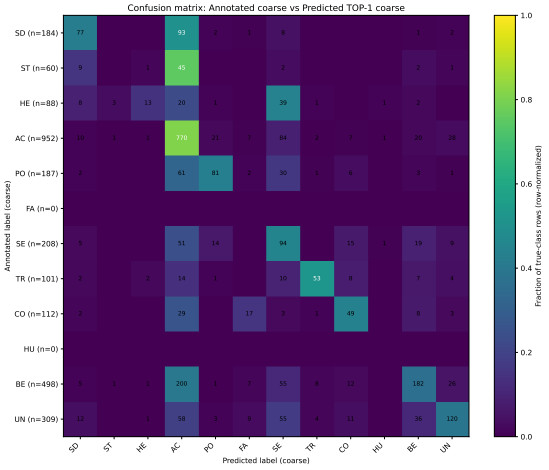

Figure 13: Base Adam-Smith confusion matrix on the full corpus. Rows are annotated coarse values; columns are top-1 predicted coarse values. The Achievement column absorbs misclassifications from many other rows.

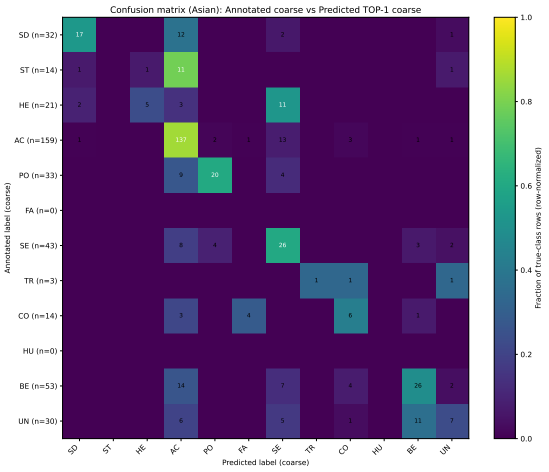

Figure 14: Base Adam-Smith confusion matrix for the Asian sample. Each cell counts instances of the row's annotated category that received the column's top-1 predicted category.

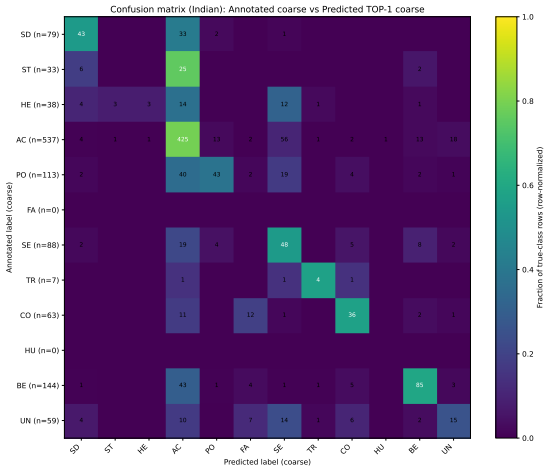

Figure 15: Base Adam-Smith confusion matrix for the Indian sample (the largest population, with the strongest concentration in the Achievement column).

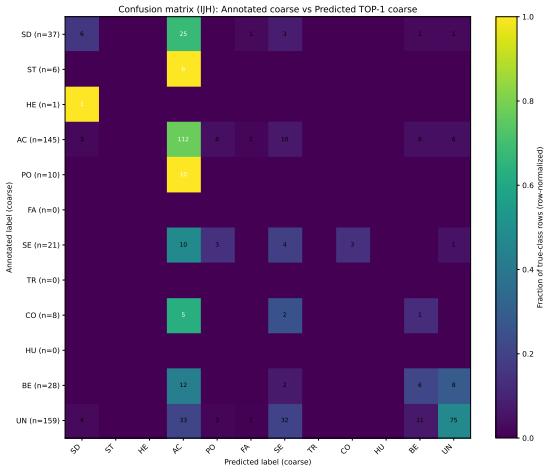

Figure 16: Base Adam-Smith confusion matrix for the IJH sample.

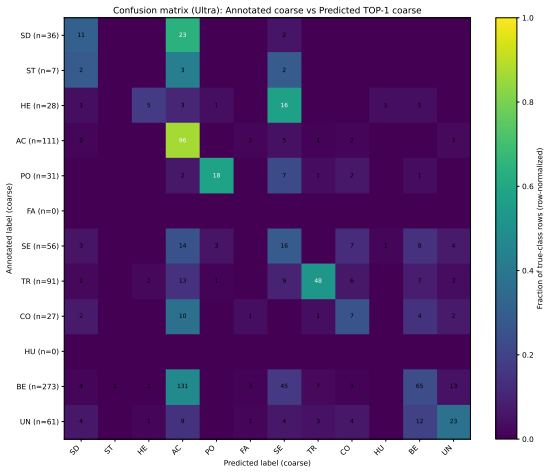

Figure 17: Base Adam-Smith confusion matrix for the Ultra sample (the most diffuse error pattern of the four populations).

E Additional Corpus Descriptives

For completeness, this section collects per-population descriptive statistics referenced in passing from Section 3: a summary table (Table 2), the per-population stacked label distribution (Figure 18), the aggregate coarse-label distribution (Figure 19), and overall response-length distributions in words and characters (Figures 20–21). Together they characterize the two structural properties of the corpus that shape the experiments: short responses (median 3 words overall) and a heavily skewed label distribution (top three values ≈ two thirds of all annotations).

Population	N	Med. wd.	H (bits)	Top label
Indian	1,161	2	2.56	AC (46%)
Ultra	721	4	2.75	BE (38%)
IJH	415	5	2.20	UN (38%)
Asian	402	3	2.74	AC (40%)
Total	2,699	3	—	AC (35%)

Table 2: Per-population corpus statistics. *Med. wd.*: median response length in words. H: Shannon entropy of the label distribution.

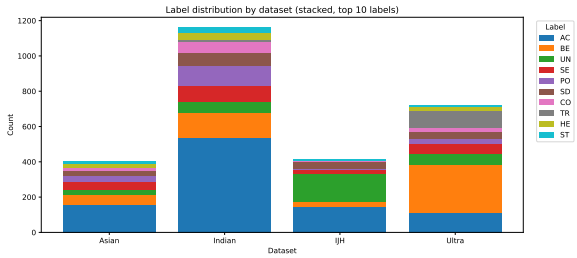

Figure 18: Stacked label counts by population, showing the dominance of a small number of values and the population-specific differences in modal value (AC for Indian/Asian, BE for Ultra, UN for IJH).

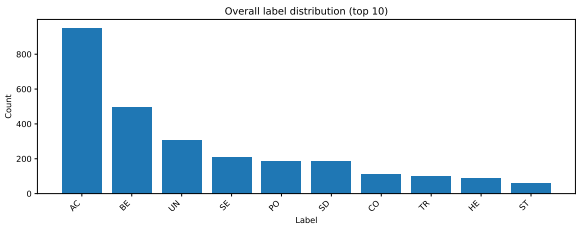

Figure 19: Aggregate coarse-label distribution combining all four samples (top 10 categories).

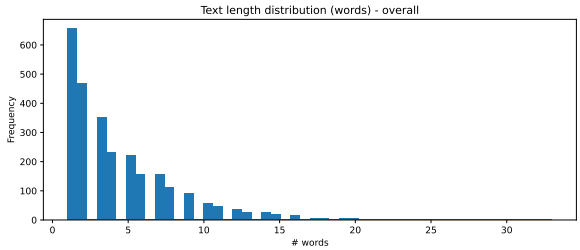

Figure 20: Word-count distribution for the full corpus. The vast majority of responses are under 10 words.

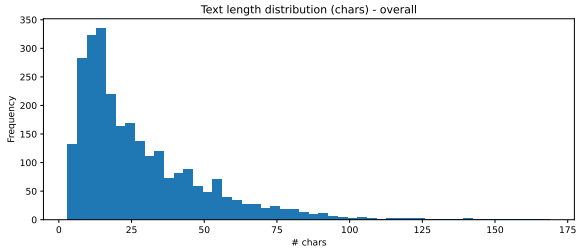

Figure 21: Character-count distribution for the full corpus. The vast majority of responses are under 60 characters.

F Survey Instruments and Procedure

Consent. All participants gave explicit consent to participate by marking a consent checkbox at the start of the questionnaire. Participation was voluntary and could be terminated at any point without penalty.

Schwartz Values Survey (SVS). Each participant first completed an adapted version of the Schwartz Values Survey (Schwartz, 1992), in which respondents rate the importance of predefined value statements as guiding principles in their life. The standard SVS items and rating procedure are described in detail by Schwartz (1992); we used the published instrument with population-appropriate translations (Hebrew for the IJH and Ultra samples; English for the Asian and Indian samples) and minor gender-adaptation of pronouns where required by the source language.

Work-goal elicitation. After completing the SVS, participants were asked to write up to three short sentences describing their work goals, in three separate free-text boxes. The English version of the elicitation prompt read: *"What goals do you aspire to advance at your workplace? Please write 3 goals that you consider most important and meaningful in your work."* For the IJH and Ultra samples, an equivalent Hebrew prompt was used; the original Hebrew text is preserved in the source materials

and is identical in substantive content to the English version above.

These free-text responses (up to three goals per participant) are the units annotated and analyzed in this paper.

Attention checks. The Asian and Indian samples included a single attention check at the start of the survey; participants who failed were given one retry, with a second failure resulting in exclusion. The IJH questionnaire included three attention-check items embedded throughout to ensure response validity.

G Sample Misclassifications: IJH Test Set

Table 3 lists all 27 misclassified instances from the IJH in-domain test set under the seed-42 epoch-10 fine-tuned checkpoint, grouped by direction of confusion. The dominant pattern (19 of 27 errors) is collapse into Achievement (top block); the middle block contains errors between other class pairs (no Achievement involvement); the bottom block contains the rarer reverse direction in which annotated Achievement instances are mistaken for another class. The column d gives the modular distance between true and predicted class on the Schwartz circumplex (1 = adjacent, 6 = diametric, see Section 6).

H Sample Misclassifications: Ultra Test Set

Table 4 lists illustrative misclassifications from the Ultra in-domain test set under the seed-42 epoch-10 fine-tuned checkpoint, organized by direction of confusion. Of 54 total errors, 30 (56%) collapse into the Tradition + Benevolence pair, the mirror analogue of IJH's collapse-into-Achievement (Section 6). 85.2% of Ultra errors cross at least one of Schwartz's higher-order axes (mean circumplex distance 3.3, vs. IJH's 3.6). The column d gives the modular circumplex distance.

I Population-Typical Distinctive Vocabulary

Table 5 lists the top distinctive content tokens of selected (population, value) cells, computed by log-odds-ratio with an informative Dirichlet prior (Monroe et al., 2008) against the same value class in the other populations. We include the four cells most relevant to the attractor analysis of Section 6: Ultra-BE and Ultra-TR (the BE–TR overlap that drives the Ultra communal attractor), and IJH-UN and IJH-AC (the UN–AC overlap that drives the IJH Achievement attractor). Tokens shared between the two rows in each pair are emphasized in **bold**; these are the words the model uses, ambiguously, to discriminate the two classes.

True → Predicted	d	Pattern	Response text
Collapse into Achievement (19 / 27)			
UN → AC	4	impact-as-achievement	"Influence on policy"
UN → AC	4	impact-as-achievement	"Creating a significant change within me"
UN → AC	4	promotion/advancement vocab	"Promoting fairness in public systems"
UN → AC	4	promotion/advancement vocab	"Promoting the social periphery in the north"
UN → AC	4	social-change framing	"Promote significant social change, especially for populations in need."
UN → AC	4	realizing-potential framing	"Providing an opportunity for everyone to realize their personal potential"
UN → AC	4	strategy/innovation vocab	"Digital and data-based social sector"
UN → AC	4	strategic / leading framing	"Promoting a government decision for optimal aging and implementing measures that Eshel will lead in the work plan"
UN → AC	4	values-as-achievement	"Organizational culture will always strive to operate inwardly from the same values that motivate us to act in society."
BE → AC	5	growth/development	"To encourage growth in people"
BE → AC	5	organizational framing	"Trust-based organization"
SE → AC	3	systems-improvement vocab	"Advanced organization"
SE → AC	3	systems-improvement vocab	"Improving technological systems infrastructure and improving processes for employees and managers"
SE → AC	3	abstract / very short	"Tidiness"
SD → AC	3	meaning-as-achievement	"Promoting a sense of self-fulfillment and satisfaction"
SD → AC	3	abstract / very short	"meaning"
SD → AC	3	professional-pride overlap	"A sense of meaning and professional pride"
ST → AC	2	innovation framing	"Expand my field of activity also in professional innovation"
PO → AC	1	PO/AC adjacent on circumplex	"To advance through the ranks"
Other non-Achievement errors (5 / 27)			
BE → PO	6	leadership/team framing	"Help and support my team"
PO → UN	5	prosocial framing of power	"People development"
SE → UN	5	broader framing of stability	"To achieve occupational stability"
HE → UN	3	broader framing of pleasure	"Flexibility in working hours and tasks, improved conditions and pay"
BE → UN	1	BE/UN adjacent on circumplex	"Improving the situation of service recipients"
Reverse: Achievement misclassified as another class (3 / 27)			
AC → UN	4	prosocial framing of work	"Improving government service"
AC → UN	4	broad/societal influence	"National influence"
AC → SD	3	autonomy/innovation framing	"Innovate and influence the products"

Table 3: All 27 misclassified responses from the IJH in-domain test set at the seed-42 epoch-10 fine-tuned checkpoint. Within each block, rows are sorted by descending circumplex distance d: cross-axis ($d \geq 4$) confusions dominate, with adjacent ($d = 1$) confusions appearing only at the bottom of each block where the annotated and predicted categories partially overlap on the Schwartz circumplex. Texts are reproduced as they appear in the test set (translated to English from Hebrew via Google Translate with manual QA; see Limitations).

True → Predicted	d	Pattern	Response text
Collapse into Tradition (16 / 54): values-as-tradition lexicon			
BE → TR	3	"values" lexicon	"Instilling values and virtues"
BE → TR	3	"values" lexicon	"To educate in the right values"
BE → TR	3	"values" lexicon	"Entering the correct values"
BE → TR	3	"values" lexicon	"Education and values before learning"
BE → TR	3	"values" lexicon	"Strengthen their values and resilience"
BE → TR	3	next-generation framing	"Educate the next generation to be good people today"
BE → TR	3	caring-as-tradition	"Love for children"
AC → TR	4	values-based achievement	"Promote value-based social programs"
AC → TR	4	role/status framing	"To enhance the status of the teacher"
AC → TR	4	precision-as-tradition	"Correct singing and precision in playing"
AC → TR	4	precision-as-tradition	"accuracy"
UN → TR	4	community-specific	"That they recognize the needs of the ultra-Orthodox public"
UN → TR	4	community-specific	"Giving more voice to working Haredi women"
UN → TR	4	universalism-as-tradition	"A benevolent world"
SE → TR	1	SE/TR adjacent	"Increase the salaries of ultra-Orthodox teaching staff, including assistants."
Collapse into Benevolence (14 / 54): achievement-as-caring			
AC → BE	5	nurturing framing	"Be effective and promote children"
AC → BE	5	knowledge-as-care	"Professional knowledge"
AC → BE	5	knowledge-as-care	"Significant knowledge"
AC → BE	5	education-as-care	"Study other subjects related to education"
AC → BE	5	education-as-care	"That they will succeed in the matriculation exams"
AC → BE	5	exemplar-as-care	"To do my work in the most honest, fair and optimal manner, thereby setting an example"
SE → BE	4	atmosphere-as-care	"Create an optimal atmosphere"
SD → BE	2	nurturing-environment	"A nurturing and enriching educational environment"
SD → BE	2	abstract / very short	"education"
Reverse: collapse out of communal attractor			
BE → AC	5	imperative as advancement	"educate"
BE → SD	2	autonomy framing	"Teaching skills that will give the student the tools to develop as an independent adult"
BE → SD	2	autonomy framing	"Create a study routine with healthy boundaries, curiosity, encouragement, and joy"
BE → SE	4	diagnostic framing	"Diagnostics, prevention and treatment of abnormalities"
TR → BE	3	care framing	"proximity"
TR → HE	5	detachment framing	"Don't take everything for granted."
Other cross-axis errors			
CO → AC	5	performance framing	"Run the business well — the way the boss wants"
UN → AC	4	advancement framing	"Social promotion"
SE → AC	3	improvement vocab	"Improving learning conditions"
SE → SD	6	autonomy framing	"Inner peace, self-understanding"
SE → HE	4	comfort framing	"Pleasant and calm surroundings"

Table 4: Selected misclassifications from the Ultra in-domain test set (seed-42, epoch 10), prioritized by descending circumplex distance within each block. The top two blocks expose the population-typical *values-as-tradition* and *achievement-as-caring* lexicons that drive Ultra's dual-class attractor (TR + BE). Hebrew-to-English translation via Google Translate with manual QA; see Limitations.

Cell	Top distinctive tokens (descending z-score)
Ultra × BE	students, **children**, give, **education**, success, **love**, student, child, promote, **values**, generation, skills, emotional, social, every
Ultra × TR	torah, jewish, **love**, **children**, israel, joy, fear, god, virtues, **values**, girls, good, educate, **education**, **students**
IJH × UN	**social**, opportunities, **promoting**, change, populations, employment, israel, society, reducing, improving, gaps, justice, field, quality, **significant**
IJH × AC	**promoting**, impact, **social**, processes, influence, ability, creating, **significant**, organization, increasing, professionalism, personal, feel, among, positive

Table 5: Top 15 distinctive content tokens per (population, value) cell, by log-odds-ratio with informative Dirichlet prior (Monroe et al., 2008) against the same value class in the other three populations. Bolded tokens appear in both rows of each pair and drive the cross-class attractor mechanism described in Section 6: in Ultra, BE–TR share *children, education, love, students, values*; in IJH, UN–AC share *social, promoting, significant*. Tokens are lowercased, stripped of punctuation, with stopwords removed.