

Open Human Value Mining: Constrained Argument Extraction and Value Concept Induction with LLMs

Shuo Wang^{1,4,*†}, Binghao Fu^{1,3,*}, Ruikang Zhang^{1,*}, Jun Wang^{1,3}, Qi Su^{1,2,✉}

¹Research Center for Digital Humanities, Peking University

²School of Foreign Languages, Peking University

³Department of Information Management, Peking University

⁴NLPIR Lab, Beijing Institute of Technology

shuowang55@bit.edu.cn, binghao_fu@stu.pku.edu.cn, 2300018416@stu.pku.edu.cn,
junwang@pku.edu.cn, sukia@pku.edu.cn

Abstract

Computational value analysis has largely treated values as instances of predefined universal taxonomies, making it difficult to discover culturally specific value dimensions not fully represented in existing schemes. We formulate value analysis as an open-ended discovery problem grounded in the argumentative contexts in which values are expressed. To this end, we propose an LLM-assisted framework that discovers human value dimensions directly from text without relying on predefined taxonomies. It extracts Premise-Conclusion-Stance (P-C-S) argumentative edges under explicit argument tree constraints, expands each edge into structured value-semantic descriptions spanning goal, principle, tradeoff, and perspective dimensions, and induces coherent value clusters via co-association consensus clustering over the embedding representations. We evaluate the framework across ten languages using external label alignment, internal structural stability metrics, and human judgment. Results show that the framework is robust and transferable, with the induced value structures achieving low hallucination and strong alignment with human judgments while revealing culturally distinctive value orientations that both align with and meaningfully complement established psychological frameworks. By shifting computational value analysis from taxonomy-driven classification toward open-ended discovery, our framework offers a new approach to cross-cultural value research, with implications for culturally aware AI alignment. The code is available at: <https://github.com/SWnlpir/OpenHumanValueMining>.

1 Introduction

Values refer to a set of beliefs and guiding principles held by individuals or groups that define

which goals and actions are considered important, appropriate, and worth pursuing. They therefore influence individual and collective judgments and actions across contexts, making values a longstanding central concern in philosophy, psychology, and the social sciences (Allport, 1961; Rokeach, 1973; Hofstede, 1984; Schwartz, 2003, 1992, 2012; Hofstede, 2001). Computational analysis of values in text provides a path toward understanding human thought, morality, and culture. Accordingly, this line of research has attracted attention in areas of natural language processing, computational social science, and AI value alignment (Kiesel et al., 2022; Mirza-khmedova et al., 2024; Ye et al., 2025a). Recent advances in large language models (LLMs), especially in cross-lingual understanding, semantic abstraction and complex reasoning, open new opportunities for exploring value structure from text.

Recently, researchers leverage LLMs to measure values directly from open-ended text. A representative example is *GPV*, which extracts perceptions from open text and maps them to an existing value system via LLM-based generative psychometrics (Ye et al., 2025a). This demonstrates that LLMs can support value analysis beyond direct label prediction. Yet for our goal, it still has important limitations. First, although *GPV* takes open-ended text as input, it ultimately maps generated perceptions onto existing value dimensions, making it less suitable for discovering value dimensions outside predefined frameworks. Second, it is not designed around multilingual or cross-cultural textual structure, which limits its applicability to diverse languages and cultures. Third, its perception-generation stage lacks explicit structural constraints, and the quality of generated perceptions is mainly assessed through human inspection.

To address these issues, we move beyond fixed-taxonomy value measurement and formulate the task as value dimension mining in an open concept

*Equal contribution.

†Work done during internship at Peking University.

✉Corresponding author.

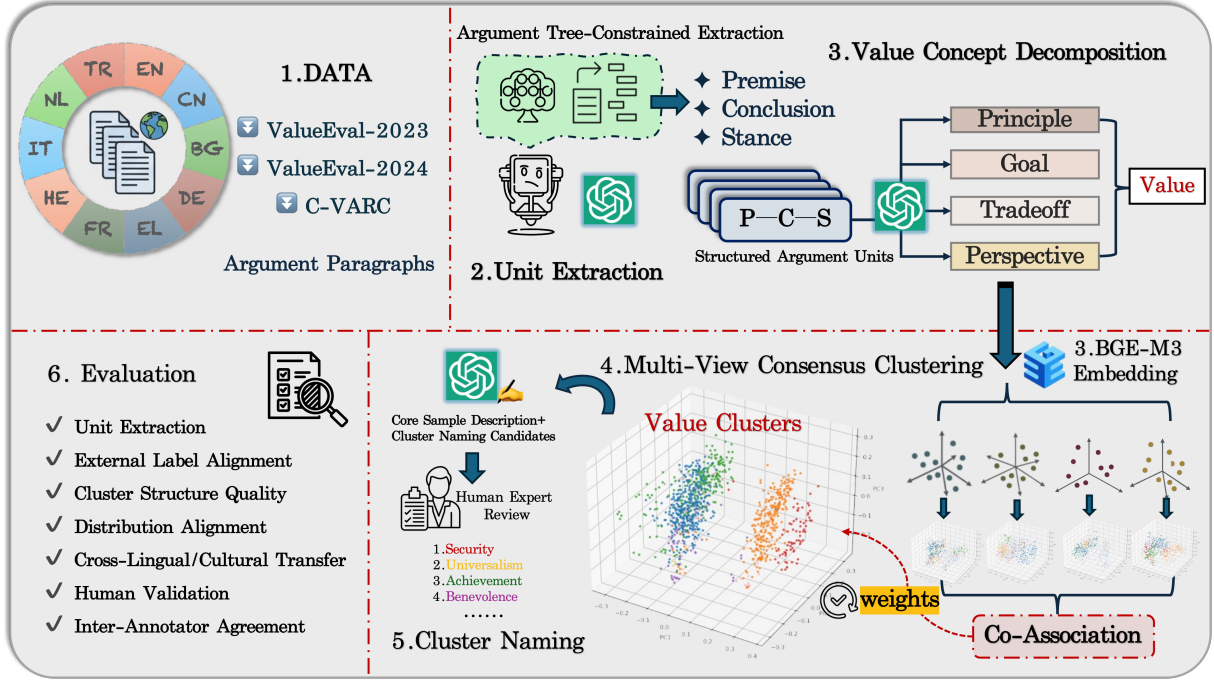


Figure 1: The overall framework for multilingual value mining, including tree-constrained unit extraction, value concept decomposition, consensus clustering, and evaluation.

space. Figure 1 summarizes the full pipeline. We first synthesize paragraph-level inputs from existing Premise-Conclusion-Stance (P-C-S) pairs using argument-structure templates while retaining the original pairs as gold edges for evaluation. Given each paragraph, an LLM then extracts P-C-S argumentative edges under a paragraph-level argument-tree constraint derived from classic argumentation theory (Palau and Moens, 2009; Freeman, 1985). This constraint grounds value expressions in explicit argumentative relations and reduces hallucinated or structurally incoherent edges.

Each extracted P-C-S edge is then expanded into a structured value-semantic representation. Drawing on classic value theories, we decompose value into semantic dimensions such as goal, principle, tradeoff, and perspective (Rokeach, 1973; Schwartz, 1992, 2012; Hofstede, 1984). An LLM then generates short textual descriptions for each dimension; these descriptions are encoded with an embedding model, concatenated, and normalized to form the vector representation used for clustering. We then repeatedly run K-means to obtain base partitions, compute pairwise co-association frequencies, build a sparse consensus graph over local semantic neighbors as the affinity matrix, and derive the final clusters via spectral clustering. The resulting clusters are finally interpreted through LLM-generated descriptions of representative evidence

followed by researcher review and refinement (Ye et al., 2025a,b; Zhu et al., 2025). Because the same procedure can be applied without language-specific redesign, the pipeline transfers across languages while still preserving locally meaningful value concepts.

We evaluate the framework from multiple perspectives. First, we assess the quality of argument unit extraction and the effect of tree-structural constraints on reducing hallucinated relations. Second, we test whether the induced clusters align with an external value-label space, providing evidence that the recovered structure is not arbitrary. Third, we test cross-lingual and cross-cultural transferability by applying the method to Chinese and eight additional languages, examining whether the pipeline recovers value structures that align with each culture’s external value space. Finally, we conduct human validation to assess whether the same- or different- cluster relations implied by the induced clusters align with human judgments of value orientation. Our results show that the proposed pipeline can recover value structure in a stable and interpretable manner. Our main contributions are:

- We propose a novel value mining framework without requiring a fully predefined value taxonomy.
- We introduce an argument-unit extraction

strategy tailored to value analysis, where imposing tree-structural constraints during extraction substantially reduces hallucinated relations and invalid units in LLM outputs.

- We validate the framework on multilingual and culturally diverse data, showing its generalizability across languages and its potential to support future cross-cultural value research and culturally aware AI alignment.

2 Related Work

2.1 Foundations of Human Values

Values are commonly defined as trans-situational guiding principles or desirable goals that shape how individuals and groups evaluate actions, goals, and social norms (Schwartz, 1992, 2012). Among the most influential frameworks in value research are Schwartz’s theory of basic human values, which organizes values into a set of motivationally grounded universal dimensions, and Hofstede’s cultural dimensions theory, which characterizes value differences at the national-cultural level (Schwartz, 1992, 2012; Hofstede, 2001). However, such value systems are primarily *etic* frameworks for universal cross-cultural comparison and therefore cannot fully capture *emic* (i.e., culture-specific) value concepts and modes of expression (Morris et al., 1999; Lu, 2012). In this sense, existing value systems provide important theoretical anchors for value computing, but they fall short of exhausting the full range of value structures expressed in multilingual and multicultural corpora.

2.2 Computational Approaches to Value Analysis

Early work on value analysis mainly followed two directions: lexicon-based methods and traditional NLP approaches. One line of research followed a psycho-lexical perspective, assuming that salient human values leave relatively stable traces in language, and therefore measured value expression through lexical compilation, value dictionary construction, and term matching. Representative studies include *Values in Words*, which predicts personal values from language use, and the *Personal Values Dictionary*, a Schwartz-based lexicon for measuring textual references to basic human values (Boyd et al., 2015; Ponizovskiy et al., 2020). Another line incorporated value analysis into traditional NLP pipelines by using lexical features, semantic features, or pretrained language models

to predict value orientations in text, as exemplified by ValueNet, as well as later argument-based value analysis datasets such as Webis-ArgValues-22 and Touché23-ValueEval (Qiu et al., 2022; Kiesel et al., 2022; Mirzakhmedova et al., 2024). These approaches laid the early foundation for computational value analysis and offered a certain degree of interpretability. However, they remain strongly dependent on predefined theoretical frameworks and lexicon coverage, limiting their ability to handle contextualized meaning, implicit expression, and semantic composition. Moreover, when extended to new languages or cultural contexts, they often require rebuilding lexicons or collecting costly annotations, which restricts their adaptability in cross-lingual and low-resource settings.

With the development of LLMs, value analysis has increasingly relied on their cross-lingual understanding, semantic abstraction, and generative capabilities. One line of work uses LLMs to improve the efficiency and accuracy of existing value identification pipelines; for example, EAVIT first employs a local model to narrow down candidate values and then uses an LLM to make the final judgment in a compressed context (Zhu et al., 2025). Another line moves further toward open-ended value measurement and value-system construction. Examples include GPV, which extracts perceptions from open text for value measurement, and GPLA, which adopts a psycho-lexical perspective to construct a new value system for LLMs (Ye et al., 2025a,b). These studies show that LLMs can not only improve existing value analysis workflows, but also open up new possibilities beyond conventional label prediction. Still, current methods are mainly oriented toward two goals: improving performance on predefined value identification tasks or constructing new measurement systems and value frameworks. In contrast, there is still limited work on directly modeling the recovery of comparable value-concept structures from multilingual corpora in a constrained and interpretable manner. This gap becomes especially important when the goal is not merely to output labels, but to jointly support structural stability, cross-lingual adaptability, and the discovery of distinctive values.

3 Dataset and Task Definition

3.1 Dataset

We evaluate the framework on three datasets with distinct roles. First, *ValueEval23 English* serves as

the main benchmark. It contains 8,865 annotated argument pairs, each consisting of a *Premise*, a *Conclusion*, a *Stance*, and a 20-dimensional human value label. Because these pairs are provided as isolated sentences, we reorganize them into paragraph-level inputs guided by argument-structure templates.

To construct more realistic argumentative contexts, we reorganize the original samples into five common argument structures from prior argumentation analysis: *Convergent*, *Serial*, *Divergent*, *Multi-Argument*, and *Against* (Lawrence and Reed, 2019; Hale and Keyser, 1998). Concrete definitions of these templates are given in Appendix B. For a template, we select sentence-level P-C-S samples that instantiate the desired structure, organize them into a structured plan, and rewrite the plan into a coherent paragraph. This procedure yields 5,000 English instances for the main method validation and ablation.

We also evaluate on the Chinese *C-VARC* dataset to test whether the English-tuned configuration effectively transfers to a different cultural context (Wu et al., 2025). Applying the same paragraph-level construction, we build 5,000 Chinese paragraphs and use the 12 *C-VARC* core value dimensions as external anchors for evaluation. This tests the method’s transferability to Chinese.

We further evaluate cross-lingual generalization on *ValueEval24* (Kiesel et al., 2024), which provides 2,240 texts in eight non-English languages: Bulgarian, German, Greek, French, Hebrew, Italian, Dutch, and Turkish. These texts are annotated with a 38-dimensional value label space, allowing us to test whether the pipeline recovers value structures aligned with each target language’s external annotations. Across all experiments, our study therefore covers a total of 10 languages.

3.2 Unit Extraction

Argumentative structure varies considerably across languages and sources. The same viewpoint may appear in a single sentence in one case, yet rely on multiple layers of support spread across an entire paragraph in another. We therefore begin by extracting minimal, comparable argument units from raw text as the common input for all subsequent experiments.

For a paragraph x , the task of *unit extraction* is to generate a set of argument edges

$$E(x) = \{(p_i, c_i, s_i)\}_{i=1}^m,$$

where p_i and c_i denote the premise and conclusion texts of the i -th argument unit, and s_i denotes its stance (i.e., support or attack). We evaluate extraction at the P-C-S edge level. A predicted edge is considered correct when its premise and conclusion semantically match a gold edge and its stance matches that of the gold edge. To allow for paraphrases and minor boundary shifts, we compute embedding-based edge similarity. The thresholded metrics in Table 1 use a threshold of 0.7, while soft F1 weights each matched pair by its semantic similarity score.

3.3 Value Mining

Given a set of extracted argument units $E = \{(p_i, c_i, s_i)\}_{i=1}^m$, the task of *value mining* is to induce value concepts from these units without presupposing a fixed target taxonomy. Each unit is first expanded into a structured value-semantic representation. These representations are then grouped into clusters, each capturing a coherent value concept shared by multiple argument units.

4 Method

4.1 Argument Tree-Constrained Extraction

Direct LLM extraction of argument units yields low accuracy in our setting because the model tends to over-generate plausible yet unsupported relations (Table 1). This produces structurally inconsistent edges and, in some cases, hallucinated content not grounded in the source paragraph. To address this issue, we introduce an argument-tree-constrained extraction method inspired by classic argumentation theory. Instead of enumerating arbitrary sentence pairs, the method requires extracted relations to form a paragraph-level argument tree, ensuring that local links remain interpretable and the overall structure stays coherent.

Given a paragraph, we perform extraction in four steps (prompts are given in Appendix D). First, the model segments the paragraph into minimal argument units and identifies those that function as conclusions. Second, it selects the root conclusion(s) from among these units. Third, starting from the root conclusion(s), the model recursively expands support and attack relations: for each current target conclusion, it predicts its premise units and their corresponding support/attack relations; if a premise is itself a conclusion unit, it is further expanded in the next round. Finally, all valid links obtained through this recursive process are assembled into a

paragraph-level argument tree and exported as standardized Premise-Conclusion-Stance edges. For refusals, malformed outputs, or predictions that violate the required constraints, we perform a limited number of retries and retain only outputs that satisfy the tree-structure requirement.

The benefit of this tree constraint is not merely a reduction in the number of predicted edges. It also helps the model organize the logic of viewpoints within the paragraph, suppressing invalid links and hallucinated edges while improving the structural quality of the extracted units.

4.2 Value Concept Decomposition

After obtaining tree-constrained argumentative units, the next question is how to represent their value content. A simple solution would be to ask an LLM to directly generate the value label or value description expressed by each argumentative unit. In our task, however, this strategy does not work well in practice. Besides limited precision, LLMs also tend to map arguments to familiar value categories, especially ones resembling Schwartz-style value dimensions.

To address this issue, we decompose value concepts into a set of more specific semantic dimensions. This decomposition views values as normative orientations jointly constituted by multiple associated aspects, such as goals, principles, trade-offs, and social standpoints. Concretely, drawing on authoritative theories of human values (Rokeach, 1973; Hofstede, 1984; Schwartz, 2003), we define seven candidate dimensions for each argumentative edge: *goal*, *principle*, *tradeoff*, *beneficiaries*, *harmed groups*, *norm*, and *perspective* (Appendix F). Conditioned on the Premise-Conclusion-Stance edge and its paragraph context, the model is asked to generate one short description for each dimension. In this way, value analysis is transformed from one-step label prediction into a process of structured semantic expansion.

The main advantage of this decomposition is that it separates components of value reasoning that are often conflated in direct generation. Two argumentative units may pursue a similar goal while relying on different principles; conversely, they may invoke the same principle while differing in trade-off judgments or social perspective. Once these components are represented separately, the comparison between argumentative units no longer depends mainly on surface wording but can instead be carried out along shared dimensions, which makes

subsequent clustering more stable.

In implementation, we start from the full seven-dimensional representation and use ablation to identify the most effective subset, as shown in Table 4. The results show that not all dimensions contribute equally to value-structure recovery. In the final configuration, we retain four core dimensions: *goal*, *principle*, *tradeoff*, and *perspective*. Compared with the remaining dimensions, these four capture the core value logic of arguments more consistently and yield the best clustering performance in the main experiments. The final value-mining pipeline is therefore built on this four-dimensional decomposition rather than on direct value-label generation.

Formally, for each extracted argumentative edge $e_i = (p_i, c_i, s_i)$, we generate the following structured semantic representation:

$$z_i = \{g_i, r_i, t_i, q_i\},$$

where g_i , r_i , t_i , and q_i denote the textual realizations of the *goal*, *principle*, *tradeoff*, and *perspective* dimensions, respectively. We encode each selected dimension text with BGE-M3, concatenate vectors, and apply L2 normalization to obtain the representation used for clustering.

4.3 Consensus Clustering

Open semantic space clustering can be sensitive to initialization, local density variation, and sample perturbation. We therefore employ consensus clustering based on co-association statistics, drawing on prior work in cluster ensembles, evidence accumulation, and resampling-based consensus methods (Strehl and Ghosh, 2003; Fred and Jain, 2005; Monti et al., 2003). In our setting, this strategy is adapted to compute sparse local co-association among open-ended value-semantic edge representations.

We first construct a sparse neighborhood graph on the argument-edge samples. For each sample, we retain only its k nearest neighbors in the semantic space, and compute co-clustering statistics only for these local semantic neighbors. We then generate multiple base partitions from repeated K -means clustering with varied random seeds, different numbers of clusters, and bootstrap subsampling. In bootstrap partitions, samples omitted from a subsample are treated as unobserved and excluded from that round’s co-clustering counts.

For each candidate pair (i, j) , we count how often the two samples are jointly observable across

Model	Pred. edges	Precision	Recall	F1	Soft F1
TreeByGPT-4.1-mini	38,497	0.7487	0.7021	0.7246	0.6701
GPT-4.1-mini	37,675	0.5065	0.9881	0.6697	0.6141
DeepSeek-3.2-Reasoner	37,914	0.5042	0.9897	0.6681	0.6166
DeepSeek-3.2-Chat	39,141	0.4939	0.9897	0.6590	0.6071
GPT-5.2-Chat	43,866	0.4417	0.9918	0.6112	0.5621
GPT-5.2-Medium	49,880	0.3894	0.9942	0.5596	0.5115

Table 1: Argument-unit extraction results at edge-similarity threshold 0.7. Soft F1 weights matched pairs by semantic similarity.

all base partitions and how often they are assigned to the same cluster. The consensus weight is defined as $w_{ij} = \frac{\text{votes}_{ij}}{\text{denom}_{ij}}$, where votes_{ij} denotes the number of times samples i and j are clustered together, and denom_{ij} denotes the number of times they are jointly observable. This weight measures the stability with which a pair remains co-clustered under repeated perturbations. Thus, the consensus graph retains only semantic relations that repeatedly survive perturbation.

Let A denote the resulting sparse symmetric affinity matrix, with $A_{ij} = w_{ij}$ for retained candidate pairs and $A_{ii} = 1$. We derive the final partition using fixed- K spectral clustering (Ng et al., 2001). The method embeds the consensus graph into a spectral space, applies K -means to the embedded samples with 10 initializations and a fixed random seed of 0, and thereby preserves stable co-association neighborhoods while naturally accommodating the sparse affinity graph. For comparability with the external label space, the default K is set to match its dimensionality.

4.4 Cluster Naming

For each cluster, we collect four types of representative argument-edge evidence, as defined in Table 2. Each evidence item contains the premise, conclusion, stance, and the generated value-semantic descriptions. These evidence sets are fed into an LLM to produce a short cluster description and several candidate names, which are then manually reviewed and refined by researchers. This design grounds cluster naming in representative evidence and human interpretation.

5 Experimental Results

5.1 Unit Extraction

We first evaluate the quality of argument-unit extraction. Table 1 reports the results at similarity threshold 0.7. Among all compared methods, the

Evidence type	Selection criterion
Center	Closest to the cluster centroid.
High-confidence	Strongest within-cluster consensus confidence.
Diverse	Covers semantic variation within the cluster.
Boundary	Closest to neighboring clusters.

Table 2: Representative evidence types used for cluster naming.

GPT-4.1-mini constrained by argument tree performs best on the main quality metrics, achieving the highest precision (0.7487), F1 (0.7246), and soft F1 (0.6701). At the same time, it produces 38,497 predicted edges, which remains on a similar scale to several direct-generation baselines.

In contrast, all direct-generation baselines exhibit a markedly different pattern: their recall is extremely high, often close to 0.99, while their precision is substantially lower, roughly ranging from 0.39 to 0.51. This suggests that LLMs tend to over-extract and hallucinate argumentative units. GPT-5.2-Medium, for example, applies stronger reasoning and extracts more units, but its precision is only 0.3894, indicating that most of its extracted units are incorrect. GPT-4.1-mini as well as the two DeepSeek variants show a similar trend: recall is nearly saturated, but precision remains around 0.50. By contrast, the tree-constrained method achieves the best balance between coverage and structural correctness, greatly reducing the chance of generating invalid units.

5.2 Main Results and Ablation on ValueEval23 English

Table 3 reports the consistency between the induced clustering and the human-annotated value space across three levels: global distribution, sample-level external similarity, and internal cluster quality (metrics are defined in Appendix H). At the dataset level, the recovered structure closely pre-

Category	Metric	Value
Distribution	Jensen–Shannon divergence	0.006
	Pearson	0.963
	Spearman	0.956
Alignment	Cosine gap	0.065
	Cosine AUC	0.604
	Random-baseline p	0.010
Structure	Max cluster ratio	0.215
	Silhouette	0.100
	Within-cluster confidence	0.855
	Confidence ratio	0.980

Table 3: Main results on *ValueEval23 English* ($K = 20$), covering distribution consistency, external alignment, and cluster structure.

serves the overall shape of the original label distribution, with a low Jensen–Shannon divergence of 0.0061 and high Pearson (0.9625) and Spearman (0.9564) correlations. At the sample level, the cosine gap reaches 0.0651, yielding a cosine AUC of 0.6045, which is significantly above the random clustering baseline ($p = 0.0099$). Thus, samples assigned to the same induced cluster are, on average, closer to one another in the external human-label space than those assigned to different clusters. The internal structure of the partition is stable and non-degenerate. The largest cluster contains 21.55% of the samples, indicating no collapse into a few dominant groups. Although the silhouette score is relatively low, this is expected because value orientations are often overlapping and compositional. Nevertheless, the consensus-graph confidence scores are high (mean within-cluster co-association = 0.8555, confidence ratio = 0.9799), showing that most samples remain strongly connected to their assigned clusters. We also reran the full consensus-clustering pipeline with five different random seeds on the same 18,250 samples; the resulting partitions are stable, with average ARI and NMI of 0.8030 and 0.9077, respectively. Taken together, these results demonstrate that the final method preserves the global distributional profile of the original annotations and induces a partition that is both externally well grounded and internally stable.

Table 4 reports a compact summary of the value-dimension ablation, with the full results in Appendix Table 14. Two patterns are apparent. First, the combination of *goal + principle + trade-off* already provides a strong semantic core, and adding *perspective* delivers the best overall trade-off. While the top structured variants perform sim-

Setting	Cosine gap	Cosine AUC	Max ratio
Core	0.0615	0.5983	0.2307
Core + N	0.0635	0.6019	0.2134
Core + N + P	0.0631	0.6010	0.2356
Final	0.0651	0.6045	0.2155
All dimensions	0.0367	0.5575	0.3586
Direct baseline	0.0215	0.5330	0.2734

Table 4: Ablation results on *ValueEval23 English*. Core = goal + principle + tradeoff; N = norm; P = perspective.

ilarly, the final setting attains the strongest external alignment without sacrificing cluster balance or consensus confidence. Second, using more dimensions is not uniformly beneficial: the full seven-dimensional representation degrades substantially, suggesting that extra dimensions introduce noise rather than useful structure. Crucially, the direct baseline falls short of every structured variant, achieving a cosine gap of only 0.0215 and a cosine AUC of only 0.5330, both well below those of the final configuration. Together, these findings suggest that structured decomposition yields more stable value representations than direct, one-step value generation.

5.3 Generalization Across Languages and Cultural Contexts

To assess generalization beyond English, we apply the selected configuration to the Chinese C-VARC dataset and to the eight non-English language datasets in *ValueEval24*. We retain the same four-dimensional semantic representation and consensus-clustering pipeline, varying only the number of clusters to match each dataset’s evaluation setting.

Figure 2 visualizes the cross-lingual patterns after normalizing each metric for comparison across languages, and Table 5 reports the overall generalization results. The multilingual results on *ValueEval24* provide strong evidence for cross-lingual generalization. Averaged across the eight non-English languages, the method achieves a cosine gap of 0.1375 ± 0.0175 and a cosine AUC of 0.6598 ± 0.0246 . Moreover, the results for all eight languages significantly outperform random baselines, with cosine AUC values above 0.62 in every case. This indicates that the proposed pipeline can recover a stable, shared value structure across languages, even though the strength of this signal varies by language.

On the Chinese C-VARC dataset, the selected configuration also aligns positively with the exter-

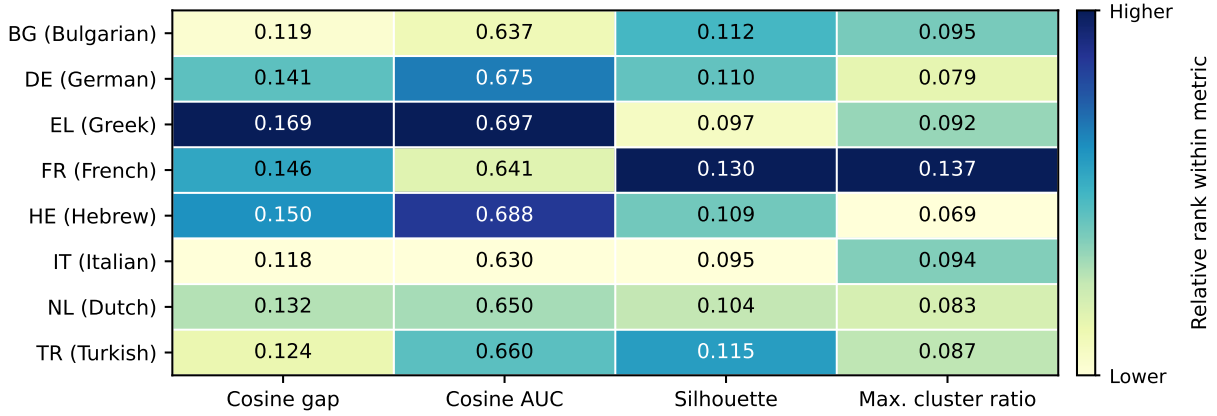


Figure 2: Heatmap of multilingual generalization results on the eight non-English languages in ValueEval24.

Setting	Cosine gap	Cosine AUC	Silhouette	Max cluster ratio
Chinese C-VARC	0.0283	0.5408	0.0299	0.2716
ValueEval24 8-languages (mean)	0.1375 ± 0.0175	0.6598 ± 0.0246	0.1090 ± 0.0110	0.0919 ± 0.0200

Table 5: Generalization results on Chinese C-VARC and the eight non-English languages in ValueEval24.

nal value space, attaining a cosine gap of 0.0283 and a cosine AUC of 0.5408. Although this signal is weaker than in the English main experiment, the induced partition remains structurally sound, with a maximum cluster ratio of 0.2716.

5.4 Sensitivity to the Number of Clusters

We assess sensitivity to the final cluster count by varying only K while holding the value-semantic representations, base partitions, and consensus graph fixed. We test $K \in \{12, 16, 20, 24, 28\}$ for English, $K \in \{8, 10, 12, 14, 16\}$ for Chinese, and $K \in \{d-8, d-4, d, d+4, d+8\}$ for each ValueEval24 language, where d denotes the dimensionality of the corresponding external label space. The main findings remain stable across these ranges. English external alignment peaks near the default $K = 20$; Chinese alignment improves with finer partitions and shows no cluster collapse; and the eight-language mean cosine AUC rises from 0.645 at $d-8$ to 0.676 at $d+8$, while the maximum-cluster ratio falls from 0.112 to 0.080. Thus, the default choices reproduce the main results and lie within the stable range. Full results are reported in Appendix M.

5.5 Case Study

Table 6 presents three qualitative cases that relate induced clusters to established value theories; representative evidence summaries are provided in Appendix N. In the Chinese data, *Patriotic Com-*

mitment combines national identification, civic duty, service-oriented dedication, and community orientation, thereby spanning several theoretical value categories. *Social Order* closely aligns with Schwartz’s *Conformity* and extends it by linking rule-following to harmony, civility, public safety, and a positive social ethos. In the Turkish data, *Religious Tradition Adherence* aligns with Schwartz’s *Tradition* and instantiates it through religious commitment, cultural heritage, historical continuity, and normative adherence.

5.6 Human Validation

We sample 300 English argument-edge pairs from the final clustering and have three annotators independently judge whether each pair expresses the same value orientation. We compare majority-vote human judgments with cluster-implied same/different labels, excluding majority-uncertain cases when computing agreement. The average pairwise agreement is 71.9%, with a mean Cohen’s κ of 0.5419 and a Krippendorff’s α of 0.5409. After excluding uncertain cases, the majority-vote judgments agree with the cluster-implied same/different labels in 82.1% of cases.

6 Conclusion

In this work, we present an open-ended value dimension mining framework that moves computational value analysis beyond fixed-label prediction toward recovering interpretable value struc-

Language / Cluster Name	Key Orientations	Interpretation
CN / Patriotic Commitment	Patriotism, Dedication, Civility, Harmony	Organizes national identification, civic duty, and service-oriented commitment into a coherent community-oriented value cluster.
CN / Social Order	Kindness, Harmony, Rule of Law, Fairness	Organizes respect, civility, legal compliance, and non-violent public conduct into a norm-centered value cluster for maintaining social order.
TR / Religious Tradition Adherence	Religious belief, historical continuity, traditional legitimacy, sacred use	Organizes religious belief, historical continuity, and the normative priority of traditional sacred use into a culturally grounded value cluster.

Table 6: Qualitative cluster cases illustrating culturally meaningful values recovered from Chinese and Turkish data.

Metric	Value
Average pairwise agreement	71.9%
Average Cohen’s κ	0.5419
Krippendorff’s α	0.5409
Majority-vote vs. cluster assignment	82.1%

Table 7: Human validation on the same 300 sampled English argument-edge pairs independently labeled by three annotators.

tures in an open concept space. Our method integrates tree-constrained argument-unit extraction, structured value-semantic decomposition, and co-association consensus clustering. Experimental evaluation across ten languages shows that the induced clusters align with human-annotated value spaces, remain structurally stable, and correspond to human judgments of value consistency. By enabling culture-specific value discovery without language-specific redesign, this work provides a robust, transferable, and empirically grounded foundation for cross-cultural value research and culturally aware AI alignment.

Limitations

Our work has several limitations. First, LLM-based value concept expansion and cluster description currently lack direct evaluation benchmarks; we therefore assess them through complementary evidence from external alignment, structural stability, and human judgment. Second, the tree-constrained extraction stage uses GPT-4.1-mini for efficiency. While it already performs well, comparisons with larger models would clarify how scale affects extraction quality. Third, our human validation is currently limited to English argument-edge pairs and annotators from a relatively narrow cultural background. Extending validation to more diverse

language communities and cultural contexts would strengthen the evidence, complementing the existing cross-lingual alignment and case-study findings. Fourth, as with any LLM-assisted cultural analytics pipeline, the method may inherit biases from models, embedding spaces, annotators, and reference datasets. Future work should therefore pair the computational framework with local expertise and bias-aware qualitative analysis.

Acknowledgment

This work was supported in part by the Fundamental Research Funds for the Central Universities (2025CX11021) and the National Key Research and Development Program of China (Grant No. 2024YFC3308101).

References

- Gordon W. Allport. 1961. *Pattern and Growth in Personality*. Holt, Rinehart and Winston, New York.
- Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2015. [Values in words: Using language to evaluate and understand personal values](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):31–40.
- Ana L. N. Fred and Anil K. Jain. 2005. [Combining multiple clusterings using evidence accumulation](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850.
- James B. Freeman. 1985. [Dialectical situations and argument analysis](#). *Informal Logic*, 7(2):151–162.
- Ken Hale and Samuel Jay Keyser. 1998. The basic elements of argument structure. *MIT Working Papers in Linguistics*, 32:73–118.
- Geert Hofstede. 1984. *Culture’s Consequences: International Differences in Work-Related Values*, volume 5. Sage.

- Geert Hofstede. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, 2 edition. Sage Publications, Thousand Oaks, CA.
- Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. [Identifying the human values behind arguments](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.
- Johannes Kiesel, Çağrı Çöltekin, Maximilian Heinrich, Maik Fröbe, Milad Alshomary, Bertrand De Longueville, Tomaž Erjavec, Nicolas Handke, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Theresa Reitismünstermann, Mario Scharfbillig, Nicolas Stefanovitch, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2024. [Overview of Touché 2024: Argumentation Systems](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, volume 14959 of *Lecture Notes in Computer Science*, pages 308–332, Berlin Heidelberg New York. Springer.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Lung-Tan Lu. 2012. [Etic or emic? measuring culture in international business research](#). *International Business Research*, 5(5):109–115.
- Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei Javaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. [The touché23-ValueEval dataset for identifying human values behind arguments](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.
- Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. [Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data](#). *Machine Learning*, 52(1–2):91–118.
- Michael W. Morris, Kwok Leung, Daniel Ames, and Brian Lickel. 1999. [Views from inside and outside: Integrating emic and etic insights about culture and justice judgment](#). *Academy of Management Review*, 24(4):781–796.
- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. [On spectral clustering: Analysis and an algorithm](#). In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.
- Raquel Mochales Palau and Marie-Francine Moens. 2009. [Argumentation mining: The detection, classification and structure of arguments in text](#). In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. Association for Computing Machinery.
- Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. [Development and validation of the personal values dictionary: A theory-driven tool for investigating references to basic human values in text](#). *European Journal of Personality*, 34(5):885–902.
- Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. [ValueNet: A new dataset for human value driven dialogue system](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.
- Milton Rokeach. 1973. *The Nature of Human Values*. Free Press, New York.
- 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. 2003. [A proposal for measuring value orientations across nations](#). In *Questionnaire Development Package of the European Social Survey*, pages 259–319. European Social Survey.
- Shalom H. Schwartz. 2012. [An overview of the schwartz theory of basic values](#). *Online Readings in Psychology and Culture*, 2(1).
- Alexander Strehl and Joydeep Ghosh. 2003. [Cluster ensembles — a knowledge reuse framework for combining multiple partitions](#). *J. Mach. Learn. Res.*, 3(null):583–617.
- Ping Wu, Guobin Shen, Dongcheng Zhao, Yuwei Wang, Yiting Dong, Yu Shi, Enmeng Lu, Feifei Zhao, and Yi Zeng. 2025. [C-VARC: A large-scale chinese value rule corpus for value alignment of large language models](#). *arXiv preprint arXiv:2506.01495*.
- Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. 2025a. [Measuring human and AI values based on generative psychometrics with large language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26400–26408.
- Haoran Ye, TianZe Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025b. [Generative psycho-lexical approach for constructing value systems in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11968–11991, Vienna, Austria. Association for Computational Linguistics.

Wenhao Zhu, Yuhang Xie, Guojie Song, and Xin Zhang. 2025. [EAVIT: Efficient and accurate human value identification from text data via LLMs](#). *arXiv preprint arXiv:2505.12792*.

A Model Settings

In the unit extraction stage, we use GPT-4.1-mini-2025-04-14, mainly due to the cost considerations associated with Argument-Tree construction. For all other stages, we use GPT-5.2-2025-12-11. The temperature is set to 0 throughout the process to ensure reproducibility. We use BGE-M3 as the embedding model for encoding value-semantic dimension texts and for computing semantic similarity in edge-level extraction evaluation.

B Paragraph Construction Templates

Table 8 defines the five structural templates used to construct paragraph-level inputs from isolated P-C-S pairs. Each template specifies a target argument pattern, and the original pairs selected for that pattern are retained as gold edges for extraction evaluation.

C Example of Paragraph Construction

Table 9 shows a compact example of the template-guided paragraph construction procedure. The example uses a divergent structure, where a shared concern about children’s agency and exploitation supports two related conclusions, while additional premises provide conclusion-specific support. The LLM is asked to realize the structured plan as a coherent paragraph, but the original P-C-S pairs are retained as gold edges for extraction evaluation.

Realized paragraph

A key concern in the entertainment industry is that child actors are for the most part thrown into a role they do not wish to be in, while their parents pocket the gains. This shared issue highlights the lack of agency and potential exploitation faced by children, which supports two important conclusions: We should ban the use of child actors and We should ban kids’ reality shows. The reasoning applies to both situations, as children in reality shows may also be pushed into roles for the benefit of adults. Furthermore, almost all reality shows involves the process of eliminating participants one by one at each level. Facing defeat in front of everyone is not that easy for

children. This may put an emotional toll on children, strengthening the call that we should ban kids’ reality shows. In addition, child actors are robbed of their childhood, and child actors miss out on socialization with kids their own age and fall behind in school, both of which further support the conclusion that we should ban the use of child actors. Thus, the exploitation and emotional harm experienced by children in both acting and reality shows point toward the need for significant reform.

This example illustrates that paragraph construction creates connected discourse for testing paragraph-level extraction, while supervision remains defined by the retained gold P-C-S edges.

D Argument Tree Prompts

The protocol uses four JSON-only steps: segment minimal units, select root conclusions, expand premise links with support/attack stance, and assemble P-C-S edges. Unit IDs must come from the segmented unit list, and extracted text preserves the original language.

Step 1: Minimal Unit Segmentation

Role: Argument structure annotator.
Task: Split paragraph into minimal argumentative units; mark conclusions.
Input: {paragraph_text}
Output:
{
 "units": [{"unit_id": "u1", "text": "..."}, ...],
 "conclusion_unit_ids": ["u2", "u5"]
}
Rules: Preserve original language; use verbatim substrings.

Step 2: Root Conclusion Selection

Task: Select the root conclusion(s) from conclusion_unit_ids.
Input:
{ "paragraph": {paragraph_text}, "units": {units_json},
 "conclusion_unit_ids": {conclusion_unit_ids_json} }
Output: { "root_unit_ids": ["u2"] }
Rules: Root IDs must be a subset of conclusion_unit_ids.

Step 3: Premise Expansion

Task: For each target conclusion, identify premise units and support/attack stance.
Input:
{ "paragraph": {paragraph_text}, "units": {units_json},
 "target_conclusion_unit_ids": {target_unit_ids_json} }
Output:
{ "expansions": [{"target_unit_id": "u2",
 "premises": [{"unit_id": "u3", "relation": "support",
 "stance": "in favor of"}]}]
Rules: relation=support/attack; stance=in favor of/against.

Step 4: Tree Assembly and Edge Export

Procedure: Starting from root conclusion(s), recursively apply Step 3.
Output:
{ "units": [...],
 "pairs": [{"src_unit_id": "...", "tgt_unit_id": "...",

Template	Argument pattern	Selection logic
<i>Convergent</i>	Multiple independent premises support the same root conclusion.	Select one conclusion with several supporting premises.
<i>Serial</i>	A premise supports an intermediate conclusion, which in turn supports a root conclusion.	Select semantically related conclusions so that one can function as an intermediate step toward the other.
<i>Divergent</i>	One shared premise supports multiple related conclusions.	Select a premise-conclusion pair and add semantically similar conclusions that can be supported by the same premise.
<i>Multi-Argument</i>	Multiple independent main conclusions appear in the same paragraph.	Select two topic-related but distinct root conclusions, each with its own supporting premises.
<i>Against</i>	Supporting and attacking premises are both present.	Prefer the same conclusion with both support and attack relations; otherwise select semantically related conclusions with opposite stance patterns.

Table 8: Structural templates used for paragraph construction from isolated P-C-S pairs.

Role in plan	Original premise	Gold conclusion(s)
Shared premise	Child actors are for the most part thrown into a role they do not wish to be in, while their parents pocket the gains.	We should ban the use of child actors; We should ban kids' reality shows.
Reality-show-specific support	Almost all reality shows involves the process of eliminating participants one by one at each level. Facing defeat in front of everyone is not that easy for children. This may put an emotional toll on children.	We should ban kids' reality shows.
Child-actor-specific support	Child actors are robbed of their childhood.	We should ban the use of child actors.
Child-actor-specific support	Child actors miss out on socialization with kids their own age and fall behind in school.	We should ban the use of child actors.

Table 9: A paragraph construction example based on a divergent template. The original P-C-S pairs are organized into a structured plan before being realized as paragraph-level input.

```
"relation": "support/attack", "stance": "in favor of/
against"},
"roots": [{"u2"}]
```

E Example of Tree-Constrained P-C-S Extraction

Using the paragraph in Appendix C, Table 10 shows how the tree-constrained procedure converts connected discourse into standardized P-C-S edges. The procedure first identifies minimal argument units, marks the two ban statements as root conclusions, and then links shared and conclusion-specific premises to the corresponding conclusions.

F Definitions of Value-Semantic Dimensions

The seven value-semantic dimensions are inspired by recurring components in classical value theories and adapted for argument-level text analysis.

Rokeach's theory distinguishes terminal values, which concern desirable end-states, from instrumental values, which concern preferable modes of conduct. This distinction motivates the separation between the *goal* facet, which captures the valued outcome, and the *principle* and *norm* facets, which capture the normative reason and behavioral expectation implied by the argument (Rokeach, 1973). Schwartz's theory defines values as motivational goals and emphasizes that values are organized by relations of conflict and compatibility. This motivates the *tradeoff* facet, which records the competing value, cost, or risk involved in accepting an argument's conclusion (Schwartz, 1992, 2012). Hofstede's cultural-value framework, together with emic/etic accounts of cultural interpretation, highlights that value judgments are shaped by culturally situated standpoints and social relations. This motivates the *perspective* facet and further supports

ID	Extracted unit	Role
u1	Children are pushed into unwanted roles for adult benefit.	Premise
u2	Public elimination may cause emotional harm to children in reality shows.	Premise
u3	Child actors are robbed of childhood and fall behind in school.	Premise
u4	We should ban the use of child actors.	Root conclusion
u5	We should ban kids' reality shows.	Root conclusion

Table 10: Extracted argument units in the end-to-end P-C-S example.

Dimension	Definition
<i>goal</i>	The desired outcome or state of affairs supported by the argument.
<i>principle</i>	The normative reason that makes the conclusion desirable, legitimate, or necessary.
<i>tradeoff</i>	The competing value, cost, risk, or sacrifice involved in accepting the conclusion.
<i>beneficiaries</i>	The people, groups, institutions, or entities expected to benefit.
<i>harmed groups</i>	The people, groups, institutions, or entities expected to be harmed, burdened, or excluded.
<i>norm</i>	The behavioral expectation, duty, or social rule implied by the argument.
<i>perspective</i>	The evaluative standpoint from which the issue is judged.

Table 11: Definitions of the seven candidate value-semantic dimensions.

making the affected parties explicit through the *beneficiaries* and *harmed groups* facets (Hofstede, 1984; Morris et al., 1999; Lu, 2012). This design yields an interpretable representation of how a value is expressed, justified, and situated in a P-C-S argument edge. Table 11 defines the seven candidate dimensions used for structured value-semantic expansion. Each dimension is generated as a short textual description conditioned on the extracted P-C-S edge and its paragraph context.

G Example of Structured Value-Semantic Representation

The final standardized edges are:

- (*u1, u4, in favor of*): lack of agency and exploitation support banning child actors.
- (*u1, u5, in favor of*): the same concern also supports banning children’s reality shows.
- (*u2, u5, in favor of*): public elimination and emotional harm support banning children’s reality shows.
- (*u3, u4, in favor of*): loss of childhood, socialization, and education supports banning child actors.

This example shows how the argument-tree constraint keeps the extraction focused on explicit support relations and exports comparable P-C-S edges for downstream value-semantic decomposition.

Table 12 gives a concrete example of the structured value-semantic representation generated for

one extracted P-C-S edge. The dimension texts are later encoded with BGE-M3, concatenated, and normalized before consensus clustering.

H Evaluation Metrics

See Table 13. Let y_i be the external label vector of sample i , and let $\mathcal{S}_{in}$ and $\mathcal{S}_{out}$ denote within-cluster and between-cluster sample pairs. The metrics are grouped by their evaluation goal.

I Full Ablation Results

Table 14 reports the full value-dimension ablation results, including the internal geometric and consensus-graph diagnostics omitted from the compact main-text table.

J English Cluster Visualization

Figure 3 provides a three-dimensional PCA projection of the English value-cluster representations. The projection is used as an illustrative visualization of the induced geometry, while the quantitative evaluation in the main text relies on external alignment, distributional consistency, and consensus confidence.

K English Value Cluster Similarity

Figure 4 visualizes the pairwise cosine similarity among the centroids of the 20 induced value clusters in the English experiment. Rather than showing individual samples, this figure highlights the higher-level semantic relations between clusters. The reordered matrix reveals that some clus-

Field	Content
Premise	Children in reality shows may be pushed into roles for adult benefit.
Conclusion	This lack of agency and exploitation supports banning child actors and children’s reality shows.
Stance	In favor of.
Goal	Prevent exploitation and harm to children.
Principle	Children should have agency and should not be used primarily for adult benefit.
Tradeoff	Banning such shows may limit child-related entertainment options.
Perspective	Protect children’s rights and welfare.
Induced cluster	Child Caring.

Table 12: Example of the structured value-semantic representation generated for one extracted P-C-S edge.

Category	Metric	Definition
Distribution consistency	JSD ↓	Jensen–Shannon divergence between the induced cluster distribution and the external label distribution.
Distribution consistency	Pearson / Spearman ↑	Correlation between induced and external distribution profiles.
External alignment	Cosine gap ↑	$E_{(i,j) \in S_{in}} \cos(y_i, y_j) - E_{(i,j) \in S_{out}} \cos(y_i, y_j)$.
External alignment	Cosine AUC ↑	Ranking-based separation between within-cluster and between-cluster cosine similarities in the external label space.
External alignment	Cosine p -value ↓	Random-baseline significance test for the cosine-based alignment score.
Cluster structure	Max cluster ratio ↓	Proportion of samples assigned to the largest induced cluster.
Cluster structure	Silhouette ↑	Cosine silhouette score computed in the value-semantic embedding space.
Cluster structure	Mean within-cluster co-association ↑	Average co-association strength between each sample and its same-cluster local neighbors.
Cluster structure	Confidence ratio ↑	Share of each sample’s local co-association mass retained within its final cluster.

Table 13: Definitions of evaluation metrics used in clustering experiments. Arrows indicate the preferred direction when comparing methods.

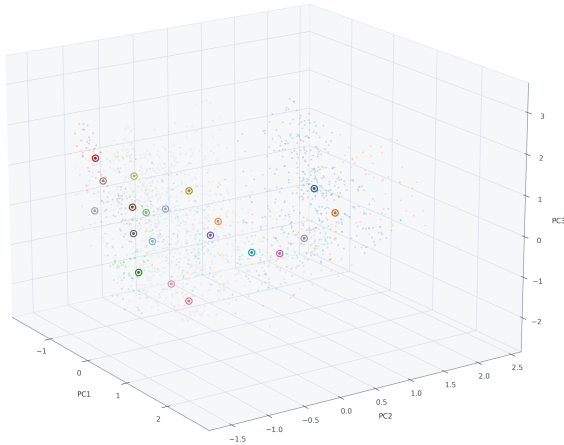


Figure 3: Three-dimensional PCA projection of the induced English value clusters for illustrative visualization.

ters form relatively close semantic neighborhoods, while others remain more distinct, suggesting that the induced value space is structured rather than arbitrary.

L Full Multilingual Results

Table 15 provides the full multilingual results of the final method on the eight non-English languages in ValueEval24. Although the strength of the signal varies across languages, the method consistently produces positive alignment with the external value space and non-degenerate cluster structures. This supports the claim that the framework generalizes beyond English and can recover stable value structure across multiple languages.

Method	Views	Cosine gap	Cosine AUC	Silhouette	Max cluster ratio	Mean within-cluster co-association
Core	goal + principle + tradeoff	0.0615	0.5983	0.1057	0.2307	0.8568
Core + norm	goal + principle + tradeoff + norm	0.0635	0.6019	0.1034	0.2134	0.8571
Core + norm + perspective	goal + principle + tradeoff + norm + perspective	0.0631	0.6010	0.1049	0.2356	0.8642
Final	goal + principle + tradeoff + perspective	0.0651	0.6045	0.0999	0.2155	0.8555
All dimensions	all seven dimensions	0.0367	0.5575	0.0642	0.3586	0.4766
Direct baseline	direct value	0.0215	0.5330	0.0581	0.2734	0.8127

Table 14: Full ablation and direct baseline results on *ValueEval23 English*. Boldface marks the best value in each metric column; lower is better for Max cluster ratio, and higher is better for the other metrics.

Language	n	K	Max ratio	Cosine gap	Cosine AUC	JSD	Pearson	Silhouette
Bulgarian (BG)	870	38	0.0954	0.1193	0.6371	0.0130	0.9733	0.1120
German (DE)	1052	38	0.0789	0.1405	0.6752	0.0152	0.9446	0.1099
Greek (EL)	1079	38	0.0918	0.1692	0.6970	0.0137	0.9629	0.0973
French (FR)	703	37	0.1366	0.1462	0.6410	0.0109	0.9630	0.1298
Hebrew (HE)	955	38	0.0691	0.1497	0.6877	0.0276	0.9123	0.1088
Italian (IT)	979	36	0.0940	0.1178	0.6299	0.0106	0.9682	0.0948
Dutch (NL)	1320	38	0.0826	0.1324	0.6505	0.0101	0.9613	0.1037
Turkish (TR)	1244	38	0.0868	0.1245	0.6600	0.0174	0.9375	0.1152
Mean $\pm$ Std.	—	—	0.0919 $\pm$ 0.0200	0.1375 $\pm$ 0.0175	0.6598 $\pm$ 0.0246	0.0148 $\pm$ 0.0057	0.9529 $\pm$ 0.0203	0.1090 $\pm$ 0.0110

Table 15: Full multilingual results of the final method on the eight non-English languages in ValueEval24. n denotes the number of edge-level metadata instances, K the final cluster number, and Max Ratio the proportion of the largest cluster.

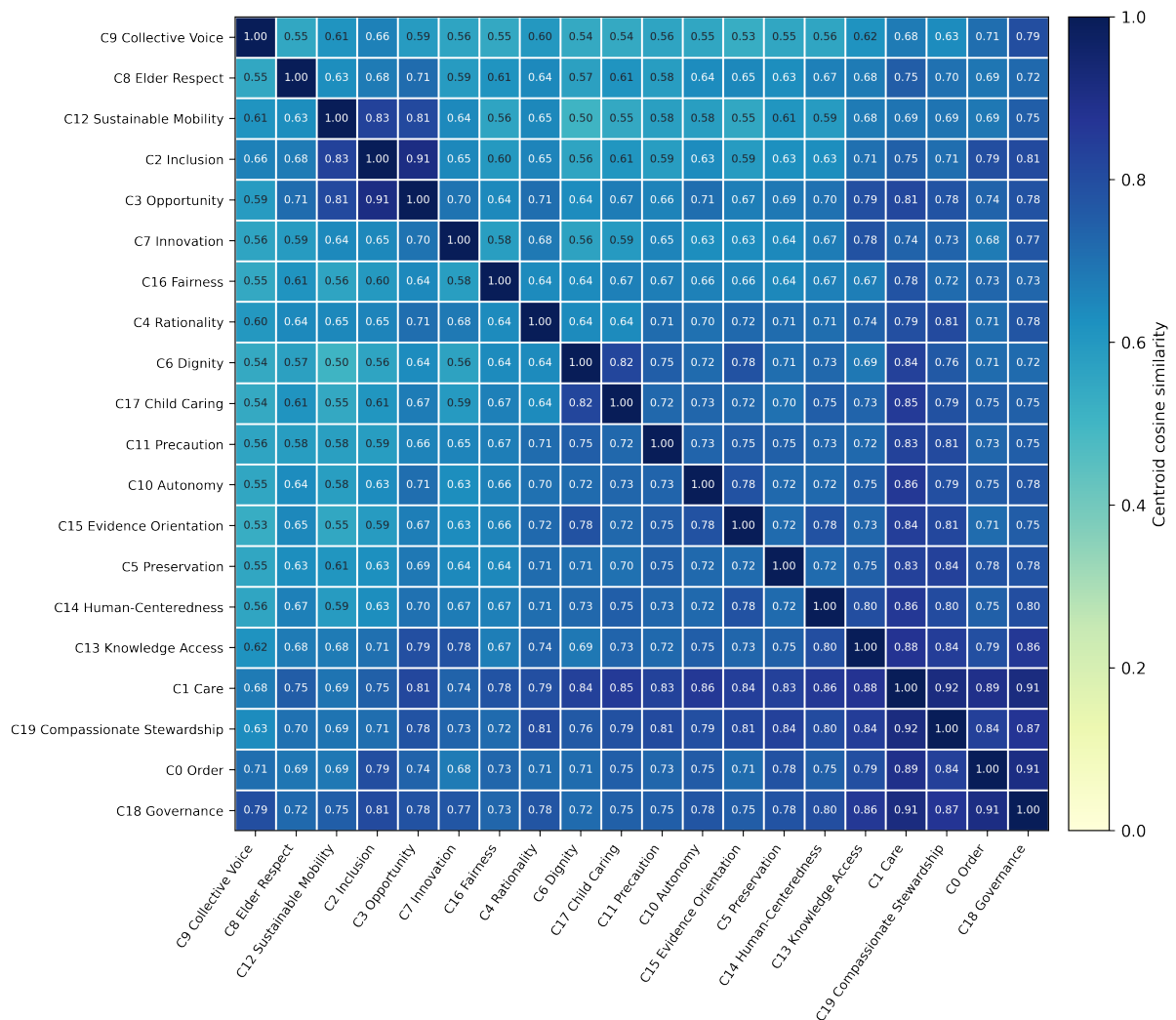


Figure 4: Heatmap of pairwise cosine similarity among the centroids of the 20 induced value clusters. Clusters are labeled with manually curated final names and reordered by hierarchical clustering.

M Sensitivity to the Number of Clusters

We vary only the final cluster count while keeping the value-semantic representations, base partitions, and consensus graph fixed. Tables 16–19 report the complete results. Defaults follow the dimensionality of the corresponding external label space.

Metric	12	16	20	24	28
Cosine gap	0.027	0.031	0.065	0.062	0.047
Cosine AUC	0.544	0.550	0.604	0.597	0.575
Max ratio	0.436	0.372	0.215	0.187	0.236
Silhouette	0.067	0.080	0.100	0.108	0.106
Confidence	0.847	0.851	0.855	0.860	0.859

Table 16: Cluster-count sensitivity on ValueEval23 English ($K = 20$ default).

Metric	8	10	12	14	16
Cosine gap	0.019	0.020	0.028	0.030	0.039
Cosine AUC	0.528	0.529	0.541	0.543	0.556
Max ratio	0.268	0.272	0.272	0.216	0.216
Silhouette	0.035	0.029	0.030	0.025	0.029
Confidence	0.608	0.610	0.615	0.624	0.630

Table 17: Cluster-count sensitivity on C-VARC Chinese ($K = 12$ default).

Metric	$d - 8$	$d - 4$	d	$d + 4$	$d + 8$
Cosine gap	0.121	0.132	0.137	0.148	0.156
Cosine AUC	0.645	0.656	0.660	0.668	0.676
Max ratio	0.112	0.098	0.092	0.086	0.080
Silhouette	0.095	0.103	0.109	0.116	0.123
Confidence	0.553	0.573	0.590	0.603	0.616

Table 18: Mean sensitivity over the eight ValueEval24 languages.

N Representative Evidence for Case Studies

Table 20 provides representative evidence summaries for the qualitative cases discussed in the case study section. The examples illustrate how the induced clusters are grounded in argument-edge evidence, without generalizing about an entire language or culture.

O Human Annotator Information

Three annotators participated in the human validation study. All annotators have postgraduate-level training and academic backgrounds related to the humanities. They independently labeled the same 300 English argument-edge pairs sampled from the final clustering results. Each pair consists of two extracted P-C-S edges. The sampled pairs include both within-cluster pairs and cross-cluster pairs, so that the annotation can evaluate whether cluster membership corresponds to human-perceived value consistency.

Annotators were asked to assign one of three labels to each pair: *consistent*, *inconsistent*, or *uncertain*. We compute majority-vote labels over the three annotations. For the final cluster-human agreement score, pairs whose majority label is *uncertain* are excluded, and the remaining majority labels are compared with the same/different relation implied by the induced clusters. Annotators were compensated according to local research-assistant payment standards, with an average wage of approximately 40 USD per day.

Label Instruction for Annotators

Consistent: The two argument units express the same or highly similar value orientation.
Inconsistent: The two argument units express different or conflicting value orientation.
Uncertain: The value relation is too ambiguous to judge reliably.

Language	Offset	K	Cosine gap	Cosine AUC	Max ratio	Silhouette	Confidence
BG	$d - 8$	30	0.088	0.610	0.168	0.099	0.527
	$d - 4$	34	0.105	0.626	0.108	0.105	0.544
	d	38	0.119	0.637	0.095	0.112	0.563
	$d + 4$	42	0.116	0.634	0.124	0.116	0.579
	$d + 8$	46	0.144	0.668	0.083	0.128	0.596
DE	$d - 8$	30	0.122	0.668	0.125	0.099	0.565
	$d - 4$	34	0.134	0.679	0.102	0.105	0.593
	d	38	0.141	0.675	0.079	0.110	0.611
	$d + 4$	42	0.141	0.676	0.085	0.115	0.624
	$d + 8$	46	0.149	0.687	0.080	0.125	0.632
EL	$d - 8$	30	0.143	0.671	0.105	0.090	0.549
	$d - 4$	34	0.156	0.687	0.093	0.095	0.564
	d	38	0.169	0.697	0.092	0.097	0.580
	$d + 4$	42	0.172	0.701	0.084	0.108	0.589
	$d + 8$	46	0.178	0.704	0.095	0.112	0.603
FR	$d - 8$	29	0.155	0.643	0.107	0.118	0.547
	$d - 4$	33	0.168	0.657	0.090	0.126	0.574
	d	37	0.146	0.641	0.137	0.130	0.594
	$d + 4$	41	0.184	0.669	0.094	0.145	0.612
	$d + 8$	45	0.197	0.684	0.088	0.151	0.621
HE	$d - 8$	30	0.133	0.675	0.104	0.088	0.576
	$d - 4$	34	0.144	0.681	0.087	0.100	0.598
	d	38	0.150	0.688	0.069	0.109	0.615
	$d + 4$	42	0.163	0.698	0.068	0.118	0.628
	$d + 8$	46	0.166	0.699	0.059	0.127	0.643
IT	$d - 8$	28	0.113	0.631	0.082	0.085	0.522
	$d - 4$	32	0.108	0.619	0.107	0.092	0.542
	d	36	0.118	0.630	0.094	0.095	0.556
	$d + 4$	40	0.136	0.649	0.070	0.099	0.572
	$d + 8$	44	0.136	0.643	0.068	0.108	0.582
NL	$d - 8$	30	0.112	0.635	0.117	0.084	0.573
	$d - 4$	34	0.125	0.646	0.095	0.095	0.589
	d	38	0.132	0.650	0.083	0.104	0.605
	$d + 4$	42	0.146	0.666	0.077	0.109	0.613
	$d + 8$	46	0.150	0.667	0.075	0.115	0.626
TR	$d - 8$	30	0.102	0.629	0.089	0.096	0.566
	$d - 4$	34	0.117	0.654	0.101	0.104	0.581
	d	38	0.124	0.660	0.087	0.115	0.596
	$d + 4$	42	0.124	0.654	0.087	0.116	0.608
	$d + 8$	46	0.127	0.652	0.091	0.118	0.623

Table 19: Full cluster-count sensitivity results for the eight ValueEval24 languages. Offset is relative to the external label-space dimensionality d .

Cluster	Representative evidence summary	Value configuration
CN / Patriotic Commitment	Arguments in this cluster connect love of country with civic responsibility, public service, and willingness to contribute to collective well-being.	National identification + civic duty + service-oriented dedication.
CN / Social Order	Arguments emphasize that kindness, harmony, civility, and law-abiding conduct protect public safety and help sustain a positive social ethos.	Rule-following + interpersonal morality + collective harmony.
TR / Religious Tradition Adherence	Arguments connect religious practice with inherited cultural legitimacy, historical continuity, and the normative importance of preserving sacred use.	Religious commitment + cultural heritage + traditional legitimacy.

Table 20: Representative evidence summaries for the qualitative case studies.