SRLCF: LEARNING SCHWARTZ VALUE REPRESENTATIONS FROM STRUCTURED INTERVIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning value representations from interviews requires distinguishing respondents' expressed values from values introduced by interviewers and integrating evidence across turns. We introduce INTERVIEWVALUES-100K, a corpus of 100,000 single- and multi-turn English interview units with machine-generated annotations for 20 Schwartz-based value categories, retained source records, and answer-side evidence. We propose SRLCF, a dual-encoder framework that combines semantic matching between dialogues and value definitions with imbalance-aware classification through label-specific gates. Gated residuals incorporate contextual answer-marker states, while fine-value and higher-order auxiliary supervision shape a dialogue-level value representation. Across three seeds, SRLCF achieves Macro-F1 of 0.510 ± 0.005 on the development benchmark, compared with 0.483 ± 0.003 for DeBERTa + ASL. In a retrospective fixed-model evaluation on 1,000 human-annotated MediaSum dialogues containing at least one expressed value, Macro-F1 is 0.494 versus 0.475 for a fixed-recipe baseline rerun, with the paired 95% confidence interval for the difference spanning zero. Representation analyses show that fine-value supervision improves frozen linear-probe performance without an observed gain in fused classification. Value-centroid distances also correlate with Schwartz's circular distances (Spearman $\rho = 0.502$), without a circular training constraint. These findings distinguish the benefits of semantic–supervised fusion for value detection from those of fine-value supervision for representation learning.

1 INTRODUCTION

Interview questions often name a value before the respondent takes a position on it. An interviewer may invoke security, fairness, or tradition; an answer may endorse that premise, qualify it, or reject it. Value detection therefore needs to distinguish the context supplied by a question from the evidence supplied by an answer. Follow-up exchanges add another difficulty: the evidence for one issue may span several turns, and several values may be expressed together.

Most established value benchmarks use sentences or arguments as their units of analysis (Kiesel et al., 2022; 2023). Direct multi-label classifiers perform well on these tasks, but their fixed output heads do not explicitly compare an utterance with a value's meaning. Label-description methods provide this comparison (Halder et al., 2020; Tunstall et al., 2022), while facing sparse positive supervision for rare values. We investigate whether label-semantic matching and direct supervised classification can be combined in interviews, and whether a linear classifier trained on frozen representations can recover the 20 value labels.

Schwartz's refined theory organizes 19 values along a motivational continuum (Schwartz, 2012; Schwartz et al., 2012). Adjacent values tend to share motivations; distant values can involve competing goals. This structure need not make their expression mutually exclusive: an answer can discuss a conflict between security and self-direction. We use four higher-order motivations as auxiliary supervision and the 19-value ring as a post-training analysis target. The model produces one sigmoid output per value, allowing labels to co-occur.

We introduce INTERVIEWVALUES-100K, comprising 100,000 single- and multi-turn English interview units. Records are traceable to retained source identifiers, dialogue turns, and answer-side evidence spans; this does not imply human-verified labels, redistribution rights, or a stable URL for every unit. Our model, SRLCF (Schwartz-supervised Residual Label-Conditioned Fusion),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

combines MPNet semantic matching with a DeBERTa-v3 direct classifier. Both encoders receive the same serialized dialogue, truncated to 384 tokens per tokenizer when necessary. Label-specific gates combine their logits. Their hidden states also produce a 768-dimensional dialogue-level value representation with fine-value and higher-order auxiliary supervision.

This work makes three contributions:

- A source-linked interview corpus and an answer-evidenced expressed-value detection task, including evidence-negative dialogues and grouped train/validation/test splits.
- A dual-encoder model combining value-definition matching with imbalance-aware classification, with a separately evaluated dialogue-level value representation.
- Development comparisons with seven neural baselines, matched ablations, frozen probes, and Schwartz-distance diagnostics, followed by fixed-model evaluation on 1,000 human-annotated interviews containing at least one expressed value.

2 Related Work

Human value detection. Webis-ArgValues and ValueEval operationalize Schwartz values for argument classification (Kiesel et al., 2022; 2023; Mirzakhmedova et al., 2024). Axies examines context-specific values (Liscio et al., 2021), and EAVIT uses LLM assistance for value detection (Zhu et al., 2025a). Recent comparisons identify recall bottlenecks in hierarchical presence gates (Yeste & Rosso, 2026b) and show that circular decoders can improve structural coherence without improving F1 (Yeste & Rosso, 2026a). We extend value detection to issue-coherent interview exchanges drawn from MediaSum, ConvoKit resources, and NewsInterview (Zhu et al., 2021; Chang et al., 2020; Spangher et al., 2025), with evidence attributed to the respondent.

Learning with label semantics. TARS formulates classification as text–label matching (Halder et al., 2020); NLI methods use label hypotheses (Yin et al., 2019). SetFit and label-description training motivate shared text/label representations (Tunstall et al., 2022; Gao et al., 2023). LEAM uses label-aware attention (Wang et al., 2018), and UCLAF adds overlap-sensitive contrastive learning (Zhu et al., 2025b), related to supervised multi-label contrastive objectives (Khosla et al., 2020; Zhang & Wu, 2024). We pair semantic matching with a direct DeBERTa classifier trained by asymmetric loss (He et al., 2023; Ridnik et al., 2021). The evaluated interface covers the fixed 20-value vocabulary; generalization to unseen value descriptions is not established.

Value representations. Value FULCRA maps model behavior into a multidimensional value spectrum (Yao et al., 2024). Our experiments concern the measurement of expressed values: they test linear readability and correspondence with Schwartz structure. Conditioning a generator on the representation remains a downstream application to evaluate.

3 Task and Dataset

3.1 Task and value taxonomy

Let $D = (t_1, \ldots, t_T)$ contain the ordered, role-preserving turns about one interview issue, and let v denote a value name and definition. We estimate

$$p_{D,v} = p(y_{D,v} = 1 \mid D, v), \qquad y_{D,v} \in \{0, 1\}. \tag{1}$$

Evaluation pairs every dialogue with all 20 values. Any subset can have label 1, including the empty set. A positive label denotes an *expressed value* evidenced in an answer: the value may be promoted, constrained, rejected, or negotiated. It does not denote endorsement, positive sentiment, or a stable trait. Questions provide context; only respondent answers license labels.

An answer weighing personal advancement against helping a relative can express both motivations while favoring only one. Polarity is retained separately in the machine-labeled corpus and is not predicted by the discriminator; the new human file contains binary labels only. Competing values can co-occur. The answer-evidence rule defines the target but does not ensure that a classifier ignores interviewer cues.

The taxonomy comprises the refined 19 Schwartz values plus the ValueEval extension *Universalism: objectivity*. We exclude the extension from circular analyses. We retain the original ten-value mapping, 54 finer indicators, and four higher-order motivations: Openness to change, Self-enhancement, Conservation, and Self-transcendence. Boundary values can belong to two higher-order groups. Appendix A gives the mapping and support counts.

3.2 Dataset construction and quality

INTERVIEWVALUES-100K combines MediaSum (77,550 units), NPR-2P (12,000), Tennis Interviews (8,000), and NewsInterview (2,450). We reconstruct interviewer/respondent turns, form atomic Q/A exchanges, and merge adjacent exchanges conservatively when they concern the same issue, up to four exchanges per unit. The result contains 60,725 single-turn and 39,275 multi-turn units from 78,539 conversations and 61,384 respondent groups. Of these, 47,226 have at least one positive label and 52,774 are evidence-negative. Full-corpus positive-label counts range from 276 to 11,308, a roughly $41\times$ difference.

Deduplication uses source identifiers, normalized Q/A text, and normalized answers. Conversation, respondent-group, exact-answer, and duplicate links define connected components for splitting. Components from the earlier 22,450-unit release retain their assignment; new components receive a stable hash assignment. The resulting train/validation/test counts are 79,750/10,613/9,637, with no conversation, respondent-group, or exact-answer overlap across splits. These checks address observed duplication and group overlap, not all possible semantic overlap between sources.

The new 77,550 MediaSum units were labeled from full Q/A turns through Alibaba Cloud Model Studio's `qwen3.8-max` API, using two prompt variants and a third strict pass for unresolved units. The first two channels produced identical label sets with complete answer-side evidence for 66.05% of records. Automated checks passed for allowed labels, exact evidence occurrence, reconstruction, provenance, and split overlap. No expert human override was applied to these units. Appendix A.3 reports the retained API settings and their reproducibility gaps; the model-input truncation audit appears in Appendix A.4.

The primary release preserves full turns, interviewer/respondent identities when available, source IDs, polarity, and evidence. A linked answer-only projection serves public sentence/argument baselines. It retains the original record ID and labels, which were assigned before projection; the canonical dialogue remains available through that ID. Processing, annotation instructions, and the complete schema appear in Appendix A.

3.3 Human-annotated evaluation set

We evaluate a final set of 1,000 human-annotated MediaSum dialogues from distinct conversations unused in training. Source screening excludes known historical conversation and respondent links, exact duplicates, and detected near-duplicates. Two annotators independently assign the 20 binary labels from full turns; a third person resolves disagreements. The annotation organizers exclude invalid or inadequately annotated records. We use the entire final file without changing its labels or text. This human process is separate from automated training-corpus annotation.

Every retained dialogue has at least one adjudicated expressed-value label. The set contains 843 single-turn and 157 multi-turn dialogues, with label support ranging from 8 to 273. It measures which values are expressed among value-containing dialogues, not whether a dialogue expresses any value at all. The final population was confirmed after earlier human-evaluation results were inspected, making this a retrospective fixed-model analysis. Appendix C gives the annotation protocol, quality-screening scope, and evaluation safeguards.

4 Method

4.1 Answer-marker-aware dual encoding

We prefix each turn with `[QUESTION]`, `[ANSWER]`, or `[TURN]` for other roles, then join turns with newlines. The same text enters `all-mpnet-base-v2` (Song et al., 2020) and DeBERTa-v3-base (He et al., 2023); each tokenizer retains at most 384 tokens, including special tokens. Let

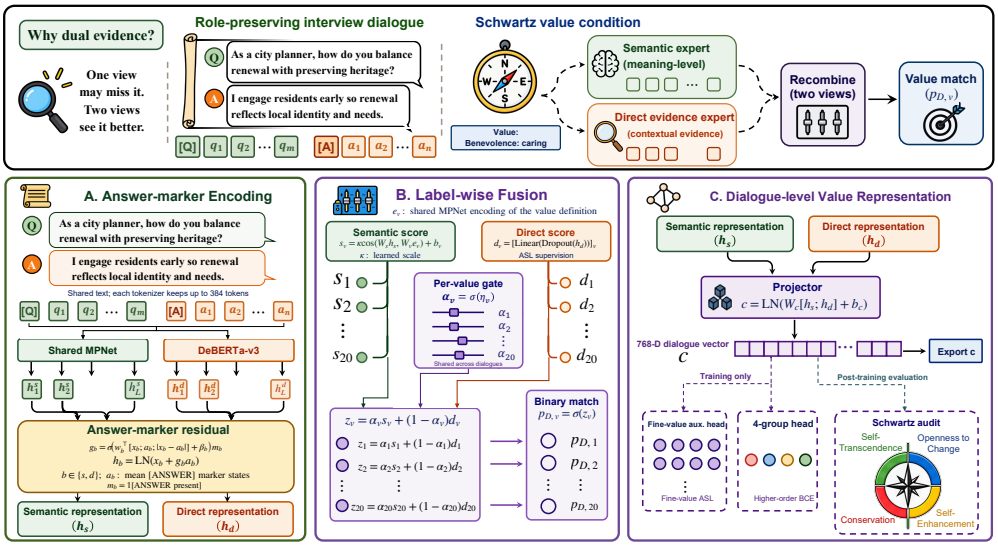


Figure 1: **SRLCF overview.** (A) Shared dialogue text is right-truncated to 384 tokens independently by each encoder's tokenizer; `[Q]`/`[A]` abbreviate role markers. (B) Definition scores s_v and direct scores d_v form logits z_v through dialogue-independent gates α_v; $p_{D,v} = \sigma(z_v)$. (C) The dialogue-level representation c has training-only auxiliary heads. The Schwartz audit and fresh frozen probe (not shown) are post-training evaluations.

x_s be the mean-pooled MPNet state and x_d the first-token DeBERTa state. In branch $b \in \{s, d\}$, a_b averages retained `[ANSWER]` states; if none remain, the implementation sets $a_b = x_b$. With $m_b = \mathbb{1}[\texttt{[ANSWER]} \text{ present}]$, the residual is

$$g_b = \sigma\big(w_b^\top [x_b; a_b; |x_b - a_b|] + \beta_b\big)\, m_b, \tag{2}$$

$$h_b = \mathrm{LN}(x_b + g_b a_b). \tag{3}$$

Answer markers attend to both questions and answers. The residual adds their contextual states; it neither pools only answer tokens nor enforces invariance to interviewer premises.

4.2 Label-conditioned fusion

The semantic branch mean-pools MPNet states for the value description (at most 128 tokens) into e_v. Projections W_s, W_v are shared across all labels; the subscript on W_v denotes the value-description projection, not a separate matrix per label. With $\kappa = \min(\exp(\tau), 100)$, its score is

$$s_v = \kappa \frac{W_s h_s}{\|W_s h_s\|_2}^\top \frac{W_v e_v}{\|W_v e_v\|_2} + b_v. \tag{4}$$

Norm denominators are clamped below at 10^{-12} in code. The cosine form in Figure 1(B) abbreviates this normalized inner product. The direct branch gives $d = \mathrm{Linear}(\mathrm{Dropout}(h_d)) \in \mathbb{R}^{20}$, with $d_v = [d]_v$, and uses asymmetric loss (ASL) to downweight easy negatives (Ridnik et al., 2021). A learned scalar per value combines the logits:

$$\alpha_v = \sigma(\eta_v), \qquad z_v = \alpha_v s_v + (1 - \alpha_v) d_v, \qquad p_{D,v} = \sigma(z_v). \tag{5}$$

The gate is initialized and regularized toward a prior α_0. It is label-specific and shared across dialogues, not dynamically conditioned on each input. The direct heads, semantic biases, and gates use the known 20-value vocabulary.

4.3 Dialogue-level value representation

In parallel with fusion, the two branch states produce one vector per dialogue:

$$c = \mathrm{LN}(W_c[h_s; h_d] + b_c) \in \mathbb{R}^{768}. \tag{6}$$

A fine-value auxiliary head applies 20-label ASL to c, and a four-group BCE head supplies higher-order Schwartz supervision with overlapping boundary memberships. The exported c is a value-supervised dialogue representation; the value-definition embedding e_v is a different object. Neither auxiliary head is needed to export c. A fresh frozen linear classifier tests whether c supports recovery of the 20 labels. The 19-value ring is used only for diagnostics, with no circular training constraint.

4.4 Training objective

The fused classification loss anchors the objective, with weighted branch, representation, and regularization terms:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_z^{\mathrm{ASL}} + \lambda_z \mathcal{L}_z^{\mathrm{BCE}} + \lambda_s \mathcal{L}_s^{\mathrm{BCE}} + \lambda_d \mathcal{L}_d^{\mathrm{ASL}} + \lambda_c \mathcal{L}_c^{\mathrm{ASL}} \\ + \lambda_h \mathcal{L}_{4-\mathrm{value}} + \lambda_a \mathcal{L}_{\mathrm{align}} + \lambda_b \mathcal{L}_{\mathrm{Brier}} + \lambda_g \mathcal{R}_{\mathrm{gate}}. \end{aligned} \tag{7}$$

Here $\mathcal{R}_{\mathrm{gate}} = K^{-1} \sum_{v=1}^{K} (\alpha_v - \alpha_0)^2$, with K value labels. The λ coefficients are fixed hyperparameters, not learned gates. BCE uses training-support weights; alignment averages positive-label semantic log-probabilities within each positive dialogue; Brier loss acts on fused probabilities. Appendix B defines all terms and Table 8 gives the coefficients and prior.

5 Experimental Setup

Comparisons and training. All interview methods use the same frozen split and return an $N \times 20$ probability matrix. We compare TARS, NLI, Dual SetFit, LEAM, UCLAF, Direct DeBERTa + ASL (our implementation of Ridnik et al., 2021), a Schwartz set decoder, and TF–IDF one-vs-rest classification. The first four are task adaptations; UCLAF is an equation-level reimplementation; the Schwartz decoder follows archived official code. SRLCF, DeBERTa + ASL, and Dual SetFit have three seeds (42, 7, 1701); other baselines and most matched ablations have one. Both SRLCF backbones use learning rate 2×10^{-5}, and new heads use 10^{-4}. Training uses length 384, effective batch 24, at most four epochs, and validation Macro-F1 early stopping. Each seed takes about 3.25 hours on one RTX 4090. Appendix B gives settings and adaptations.

Fixed-model human evaluation. Historical experiment checkpoints were deleted after inference. We therefore evaluate the separately retained seed-42 SRLCF deployment checkpoint, frozen before human annotation, and rerun DeBERTa + ASL once with its original seed-42 recipe, training split, and validation selection. The SRLCF thresholds come from its retained validation report; ASL thresholds use only the original validation split. Human labels never enter training, threshold fitting, or model selection. Joint SRLCF branches are diagnostics, not standalone baselines. This single-run follow-up is reported separately from the original three-seed comparison.

Scope of evidence. Seed-42 test diagnostics informed architecture development, so the original test split remains a development benchmark. The human evaluation uses positive-only MediaSum dialogues outside training, not unseen domains or a new blind confirmatory test. It excludes no-value dialogues. Matched removals are configuration-specific; centroid associations are descriptive. Intervals crossing zero are inconclusive, not evidence of equivalence. The experiments assess expression, not endorsement or generation control.

Observed input coverage. On the full 100k corpus, 10.68% of MPNet inputs and 9.78% of DeBERTa inputs are right-truncated, retaining 95.88% and 96.35% of cited answer spans. Canonical dialogues and annotation inputs remain intact. Appendix A.4 reports split-specific rates, denominators, and the greater evidence loss on multi-turn inputs.

Metrics and statistics. Macro-F1 is primary; Micro-F1, Macro-AUPRC, and 15-bin ECE describe detection, ranking, and calibration (Guo et al., 2017). Interview thresholds use a validation grid of 0.02. Development comparisons use 5,000 paired dialogue bootstraps; those intervals omit residual within-conversation dependence. The human evaluation uses 2,000 paired resamples with fixed thresholds; each unit comes from a distinct conversation. All 20 labels remain in every Macro-F1 calculation, with undefined label F1 set to zero. Human-set intervals quantify sample variability

Table 1: Development-benchmark results grouped by replication. Bold marks the best three-seed mean; single-run adaptations are separate. DeBERTa + ASL uses Ridnik et al. (2021) in our implementation; 2021 denotes the loss publication year. Human-evaluation results are reported separately in Table 3.

Method	Macro-F1	Micro-F1	Macro-AUPRC	ECE ↓
Three-seed comparisons (mean ± SD)				
SRLCF (ours)	**0.510±0.005**	**0.578±0.004**	**0.516±0.002**	0.056±0.001
Direct DeBERTa + ASL (2021)	0.483±0.003	0.564±0.004	0.480±0.001	0.174±0.011
Dual SetFit (2022)	0.481±0.003	0.548±0.002	0.470±0.003	**0.038±0.004**
Single-run adapted baselines				
TARS (2020)	0.440	0.512	0.418	0.050
NLI (2019)	0.434	0.510	0.411	0.059
Schwartz decoder (2026)	0.433	0.531	0.399	0.016
LEAM (2018)	0.301	0.413	0.232	0.093
TF-IDF OvR (classic)	0.301	0.397	0.258	0.093
UCLAF (2025)	0.246	0.346	0.189	0.067

conditional on the frozen runs and positive-only population, not training-seed or annotator uncertainty. Pair-level precision and recall describe the transferred operating point.

The auxiliary public-data augmentation study uses an earlier 4k answer-only subset and a separate classifier. Its protocol and results appear in Appendix F.

6 Results

6.1 Development-benchmark performance

SRLCF reaches 0.5100 ± 0.0047 Macro-F1 versus 0.4833 ± 0.0031 for DeBERTa + ASL (Table 1). Micro-F1 rises from 0.5643 to 0.5783 and Macro-AUPRC from 0.4802 to 0.5159. Paired Macro-F1 gains across seeds are 0.0257, 0.0243, and 0.0303; all 95% intervals exclude zero (Appendix B).

ECE decreases from 0.1739 for DeBERTa + ASL to 0.0556 for SRLCF. Dual SetFit has lower ECE (0.0377) among the three-seed models. The decoder's 0.016 ECE uses energy marginals rather than raw backbone probabilities (Appendix B). Brier scores are 0.0263 for SRLCF, 0.0265 for Dual SetFit, and 0.0543 for DeBERTa + ASL. Figure 2(c) shows fused-output reliability; pooled calibration can conceal rare-positive errors.

6.2 Fusion and answer-marker ablations

Within jointly trained SRLCF, the semantic and direct branches obtain Macro-F1 of 0.4681 and 0.4695, while fusion reaches 0.5100. Fusion exceeds the stronger branch for 19 of 20 values; Hedonism decreases by 0.0055. Gate weights correlate with each value's semantic-branch advantage (Spearman $\rho = 0.558$). Gains over the separately trained DeBERTa + ASL baseline are larger for low-support labels ($\rho = -0.566$), including Power: resources (+0.073) and Stimulation (+0.070). Figure 2 shows these associations.

The matched answer-marker ablation uses the configuration without fine-value auxiliary supervision. Removing the residual reduces seed-42 overall F1 from 0.5037 to 0.4964 and multi-turn F1 from 0.4919 to 0.4859; single-turn F1 falls from 0.5077 to 0.5032. The complete model's multi-turn mean is 0.5002, against 0.4624 for DeBERTa + ASL and 0.4642 for Dual SetFit. That comparison includes all architectural differences; the matched removal provides the component-specific result.

Table 2 also shows a small learned-gate gain (+0.0011) and a larger learning-rate effect (+0.0072). Removing hierarchy supervision improves the base configuration by 0.0024. The hierarchy auxiliary thus cannot be credited with the main classification gain. Appendix B identifies the exact configuration for each pair.

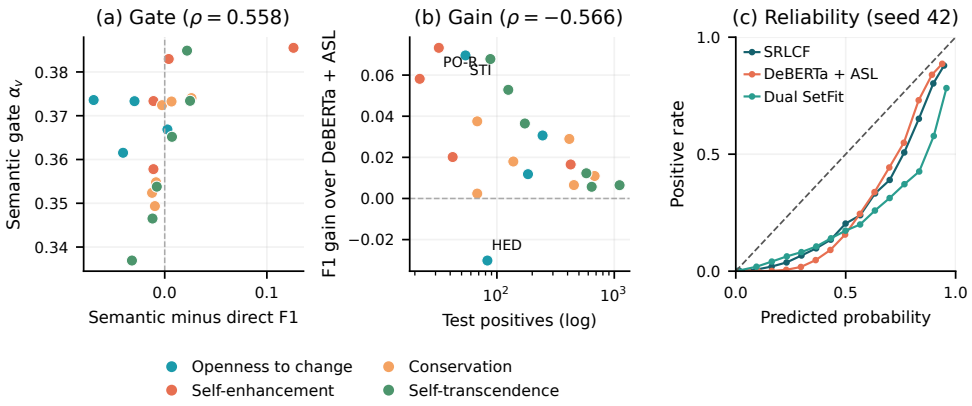


Figure 2: **Fusion and probability diagnostics.** Panels (a–b) use three-seed means across 20 values; colors denote primary higher-order groups. (a) Gate weights versus semantic-minus-direct F1 ($p = 0.0106$). (b) Gains over DeBERTa + ASL versus support ($p = 0.0092$). Both associations are observational. (c) Seed-42 reliability in 15 equal-width bins; the dashed diagonal is ideal.

Table 2: Matched ablations and frozen probes. Upper rows use seed 42, with explicit settings and reference rows; Δ subtracts the reference within each pair. Base configurations differ, so these are not cumulative removals. Lower rows change only fine-value auxiliary supervision under the final settings (three-seed mean$\pm$SD).

Configuration (seed 42)	Setting	Macro-F1	Δ
No fine-value auxiliary	With answer-marker residual	0.5037	+0.0073
	Without answer-marker residual	0.4964	ref.
Base configuration	Learned gate	0.4945	+0.0011
	Fixed gate ($\alpha = .35$)	0.4934	ref.
Base configuration	With hierarchy loss	0.4945	-0.0024
	Without hierarchy loss	0.4969	ref.
Fine-value auxiliary	Backbone LR 2×10^{-5}	0.5055	+0.0072
	Backbone LR 10^{-5}	0.4983	ref.

Fine-value auxiliary (three seeds)	With auxiliary	Without auxiliary	Δ
Fused classifier Macro-F1	0.5100$\pm$0.0047	0.5103$\pm$0.0062	-0.0003
Frozen-probe Macro-F1	.3212$\pm$.0010	.3138$\pm$.0016	+.0074
Frozen-probe Macro-AUPRC	.2033$\pm$.0015	.1960$\pm$.0020	+.0073

Source-level results qualify the aggregate gain (Figure 3(a)). SRLCF exceeds DeBERTa + ASL on all four source means, but Dual SetFit remains stronger on Tennis (0.3903 versus 0.3765). MediaSum dominates the corpus; NewsInterview has only 211 test units. Mean per-seed worst-source F1 is 0.3664 for SRLCF, 0.3235 for DeBERTa + ASL, and 0.3424 for Dual SetFit.

MPNet/DeBERTa truncate 18.67%/17.28% of multi-turn inputs versus 5.52%/4.92% of single-turn inputs. Across the full corpus, 2,004 positive dialogue–label pairs lack a fully retained cited span in either branch; their targets remain unchanged. Appendix A.4 details this input-coverage limitation.

6.3 Dialogue-representation analysis

Fine-value supervision changes fused classifier Macro-F1 from 0.5103 to 0.5100; all three paired confidence intervals include zero. With encoders frozen, a fresh linear probe improves from 0.3138 ± 0.0016 to 0.3212 ± 0.0010, and Macro-AUPRC from 0.1960 to 0.2033 (Table 2). The probe fits 8,490 validation representations, tunes thresholds on the other 2,123, and evaluates on 9,637 test dialogues. This separates label recovery from frozen c from fused classification performance.

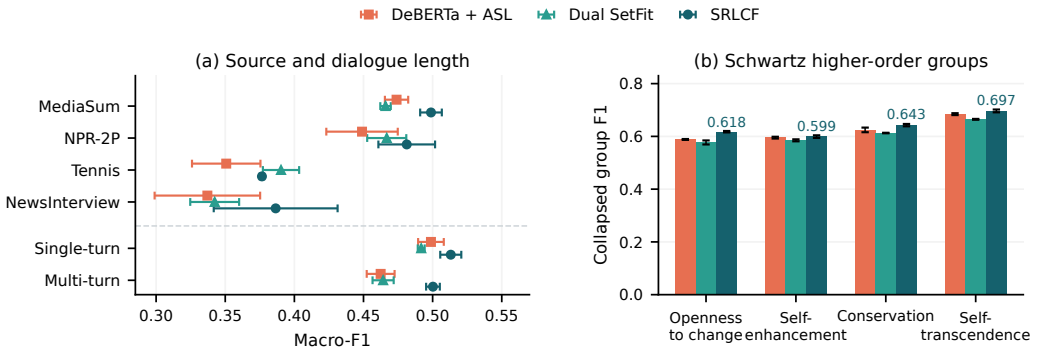


Figure 3: **Source and higher-order recognition.** (a) Three-seed means and SDs by source and turn structure within the shared test corpus; this is not leave-one-source-out transfer. (b) Fine-label predictions collapsed into four overlapping Schwartz groups. Bars and error bars show means and SDs; numbers identify SRLCF scores.

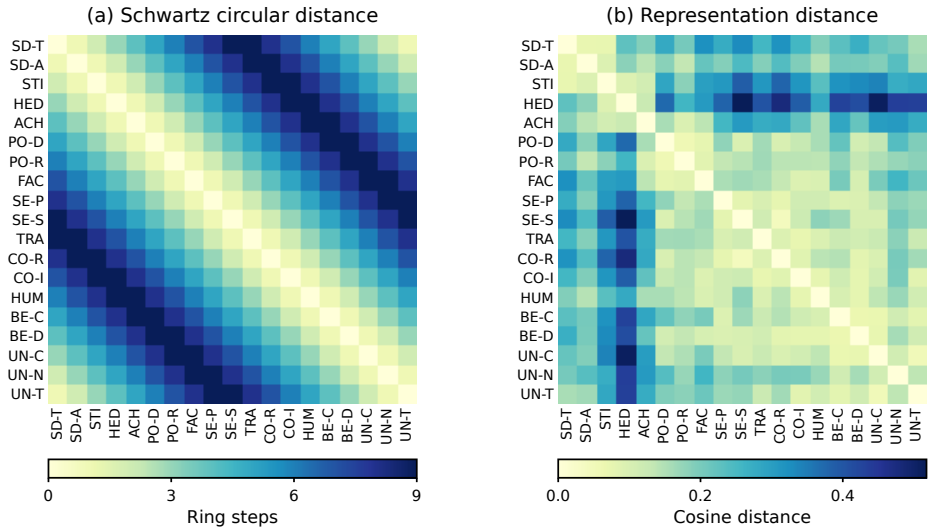


Figure 4: **Post-training dialogue-representation geometry (seed 42).** The 19 classical values follow theoretical circular order. (a) Shortest ring distance. (b) Cosine distance between value-positive dialogue centroids in the original 768-D space, grouping dialogues by label 1. Scales differ; upper-triangle $\rho = 0.502$. Figure 11(b) gives the permutation null.

For the 19 classical values, centroid cosine distances correlate with Schwartz circular distance at $\rho = 0.502$ (Figures 4 and 11(b)). The 5,000 permutations shuffle whole value identities, preserving dependence among the 171 pairs; $p < 0.001$. MPNet value-definition embeddings have a weaker correlation, $\rho = 0.227$. These seed-42 centroids group dialogues by their expression labels, not by endorsement. Co-occurring labels and topic distributions can contribute to the association.

Collapsing fine-label predictions into four overlapping groups gives mean group F1 of 0.639 for SRLCF, 0.623 for DeBERTa + ASL, and 0.610 for Dual SetFit (Figure 3(b)). Group targets take an OR over member labels, so correct group predictions can hide fine-label errors. The circular decoder has lower circular error (0.165 versus 0.353) but lower Macro-F1 (0.433 versus 0.510). These criteria capture different behavior. Appendix E gives the geometric reference, decoder metric, and ring-ordered per-label results.

Table 3: Human-annotated positive-only set ($n = 1{,}000$), one frozen run per system without human-label tuning. Intervals use 2,000 paired resamples.

Model	Macro-F1 [95% CI]	Micro-F1	Macro-AP
DeBERTa + ASL	0.475 [0.435, 0.505]	0.533	0.584
SRLCF (deployment)	0.494 [0.456, 0.524]	0.572	0.598

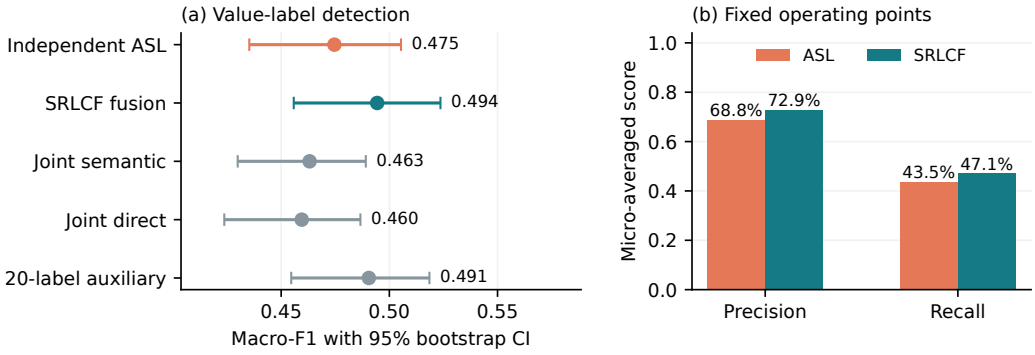


Figure 5: **Positive-only human evaluation.** (a) Macro-F1 and percentile intervals, conditional on each frozen run. Joint outputs share the SRLCF checkpoint. (b) Micro-averaged precision and recall at unchanged thresholds. Both panels use the same 1,000 dialogues; negative labels within these dialogues remain in evaluation.

6.4 Fixed-model evaluation on human labels

On the 1,000 human-annotated dialogues, frozen SRLCF obtains 0.494 Macro-F1 and 0.572 Micro-F1 (Table 3). DeBERTa + ASL obtains 0.475 Macro-F1; the paired difference is $+0.020$, with 95% CI $[-0.012, +0.052]$. SRLCF has higher point estimates on Macro-F1, Micro-F1, and Macro-AP. The Macro-F1 interval crosses zero, however, leaving comparative uncertainty.

SRLCF has precision 0.729 and recall 0.471, compared with 0.688 and 0.435 for ASL (Figure 5(b)). Both fixed operating points miss over half of the annotated positive pairs. Power: dominance ($n = 14$) and Power: resources ($n = 8$) remain weak and sparsely supported (Appendix C).

Fusion exceeds its jointly trained semantic and direct branches by 0.031 and 0.035 Macro-F1; both nominal paired intervals exclude zero. The auxiliary head scores 0.491, close to fusion. The supplementary frozen linear probe reaches 0.468 Macro-F1 without human-label fitting. Fusion remains the fixed primary output.

Multi-turn and multi-value dialogues. On 157 human-annotated multi-turn dialogues, SRLCF reaches 0.512 Macro-F1 versus 0.449 for ASL; on 843 single-turn dialogues, the scores are 0.480 and 0.471. On 472 multi-value dialogues, SRLCF reaches 0.504 versus 0.479 (Appendix C). These descriptive results motivate matched studies of turn integration; differing subgroup support precludes attributing the gaps to the answer-marker mechanism.

7 Conclusion and Limitations

We introduce InterviewValues-100K and SRLCF for detecting expressed values in role-preserving interviews. The corpus retains source records and answer-side evidence; SRLCF combines semantic matching with direct classification, improving development Macro-F1 across three seeds. Fine-value supervision improves frozen-probe label recovery, separating representation quality from fused classification. The human evaluation yields higher point estimates than DeBERTa + ASL. Future work can broaden human validation across sources and value-negative dialogues, strengthen rare-value recognition, and assess these representations for value-conditioned generation.

AI USE STATEMENT

Generative AI was used to annotate the training corpus through Alibaba Cloud Model Studio's `qwen3.8-max` API; to assist with method ideation, code, experiment orchestration, and analysis scripts; and to draft or revise manuscript language and figures. Automated training-data annotation channels used the same model family. The separate human evaluation uses independent A/B annotation with third-person adjudication of disagreements. Annotators received dialogue text with zero-initialized labels and no model predictions. Evaluation uses all 1,000 records in the final human-confirmed file, without further AI relabeling or filtering. AI assistance was used for validation and evaluation code. Reported metrics derive from saved predictions. The authors are responsible for the manuscript and its AI-assisted content.

ETHICS STATEMENT

The corpus contains public interview speech about real people. A label concerns values expressed in one exchange, not stable personality, moral worth, or future behavior. English, media-heavy sources and shared-family LLM training labels can encode cultural and political biases. Source terms and removal requests govern derived-data distribution. Twenty master's students from unrestricted disciplines participated as volunteer annotators. The human annotation file uses pseudonymous speaker identifiers; source mappings remain separate. The evaluation artifact retains adjudicated labels rather than separate A/B judgments; no agreement coefficient is reported. Person-level ranking, eligibility decisions, and covert profiling are excluded uses.

REPRODUCIBILITY STATEMENT

Appendices A–B describe construction, annotation instructions, losses, and original experiments. Appendix C adds the human A/B protocol, quality-screened evaluation population, per-label results, and paired confidence intervals. Model and input hashes, evaluation IDs, exact decisions, thresholds, predictions, and the fixed-recipe baseline configuration are retained. Appendix G distinguishes the retained deployment checkpoint, historical experiments, and baseline rerun. Typesetting uses the included figures and tables without requiring GPU access.

REFERENCES

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. ConvoKit: A toolkit for the analysis of conversations. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes (eds.), *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 57–60, 1st virtual meeting, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigdial-1.8. URL https://aclanthology.org/2020.sigdial-1.8/.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. The benefits of label-description training for zero-shot text classification. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13823–13844, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.853. URL https://aclanthology.org/2023.emnlp-main.853/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. Task-aware representation of sentences for generic text classification. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3202–3213,

Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.285. URL https://aclanthology.org/2020.coling-main.285/.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL https://arxiv.org/abs/2111.09543.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. Identifying the human values behind arguments. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4459–4471, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.306. URL https://aclanthology.org/2022.acl-long.306/.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (eds.), *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2287–2303, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.313. URL https://aclanthology.org/2023.semeval-1.313/.

Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. Axies: Identifying and evaluating context-specific values. In *AAMAS*, pp. 799–808, 2021.

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. The touché23-ValueEval dataset for identifying human values behind arguments. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16121–16134, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1402/.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, 2021. doi: 10.1109/ICCV48922.2021.00015.

Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663, 2012.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

Alexander Spangher, Michael Lu, Sriya Kalyan, Hyundong Justin Cho, Tenghao Huang, Weiyan Shi, and Jonathan May. NewsInterview: a dataset and a playground to evaluate LLMs' grounding gap via informational interviews. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32895–32925, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1580. URL https://aclanthology.org/2025.acl-long.1580/.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts, 2022. URL https://arxiv.org/abs/2209.11055.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2321–2331, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1216. URL https://aclanthology.org/P18-1216/.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8762–8785, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.486. URL https://aclanthology.org/2024.naacl-long.486/.

Víctor Yeste and Paolo Rosso. Beyond independent labels: Schwartz-geometry decoding for human value detection, 2026a. URL https://arxiv.org/abs/2607.05052.

Víctor Yeste and Paolo Rosso. Human values in a single sentence: Moral presence, hierarchies, and transformer ensembles on the schwartz continuum, 2026b. URL https://arxiv.org/abs/2601.14172.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404. URL https://aclanthology.org/D19-1404/.

Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16786–16793, 2024.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. MediaSum: A large-scale media interview dataset for dialogue summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5927–5934, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.474. URL https://aclanthology.org/2021.naacl-main.474/.

Wenhao Zhu, Yuhang Xie, Guojie Song, and Xin Zhang. Eavit: Efficient and accurate human value identification from text data via llms, 2025a. URL https://arxiv.org/abs/2505.12792.

Zhengzhong Zhu, Pei Zhou, Zeting Li, Kejiang Chen, and Jiangping Zhu. Multi-label text classification with label attention aware and correlation aware contrastive learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025b. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/936. URL https://doi.org/10.24963/ijcai.2025/936.

A Dataset and Annotation

A.1 Sources, splitting, and issue boundaries

Figure 6 and Table 4 report the full source distribution. Atomic interviewer/respondent runs form Q/A exchanges. Adjacent exchanges are merged when a boundary model predicts issue continuity above 0.75 or a short explicit follow-up rule applies, with at most four exchanges per unit. The retained boundary evaluation gives ROC-AUC 0.784, merge precision 0.832, and recall 0.307. This conservative threshold favors splitting an extended issue over joining unrelated questions.

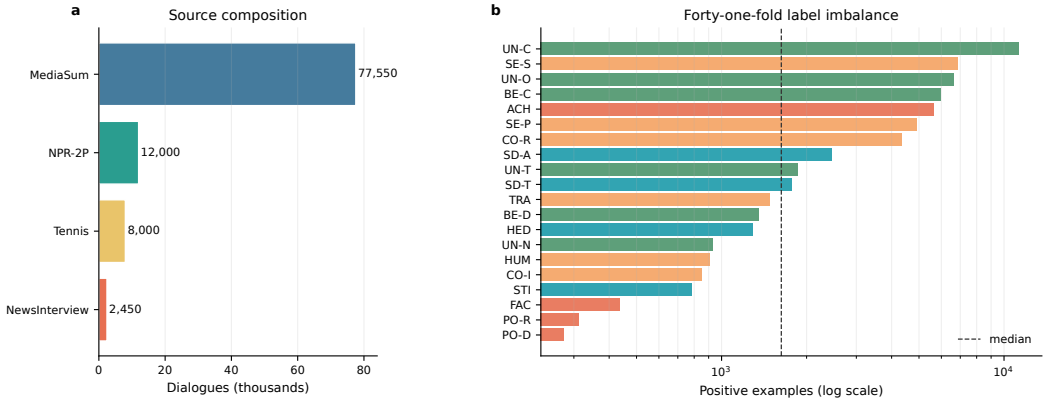


Figure 6: Source composition and full-corpus label support. Support is highly imbalanced across both sources and values.

Table 4: Frozen train/validation/test counts by source.

Split	Dialogues	MediaSum	NPR-2P	Tennis	NewsInterview
Train	79,750	61,833	9,569	6,380	1,968
Validation	10,613	7,972	1,216	1,154	271
Test	9,637	7,745	1,215	466	211
Total	100,000	77,550	12,000	8,000	2,450

The MediaSum build scans 463,596 conversations and 13,919,244 utterances, constructs 794,086 atomic Q/A exchanges, merges them into 754,877 issue units, and retains 743,505 eligible units after duplicate checks. Fixed-seed coverage sampling, source/conversation/respondent caps, and 15% lexical-proxy hard negatives select 77,550 units. Lexical proxies affect sampling, not annotation. Splits use connected components over conversation, respondent group, exact answer, and duplicate links. Seventy-two initially conflicting records were replaced before release validation.

A.2 Annotation instructions and adjudication

The annotation function sends the full role-preserving turns, speaker names, source title, and available question summary, together with the taxonomy definitions. Its system instruction restricts labeling to values expressed, promoted, constrained, rejected, or traded off by the respondent in answer turns. Questions and context are for disambiguation. It prohibits inferring a stable personality profile or treating topic mentions as sufficient evidence.

Channel A requests a conservative evidence-first reading and consideration of competing values. Channel B reverses taxonomy order and requests an independent counter-check rejecting mere topic mentions. The strict evidence rule requires an exact quote and valid answer-turn ID for every retained label. Output consists of the original sample ID, allowed 20-value names, explicit fine-value indicators, polarity, evidence, a concise rationale, and confidence. Empty labels are permitted. The prompt is constructed by `build_messages` in `annotate_dialogue_values.py`; full definitions come from `configs/taxonomy.json`.

Agreement requires identical label sets and complete answer evidence from each channel. It does not require the two channels to select identical quote strings. For the new MediaSum portion, 51,224 units (66.05%) meet this criterion and 26,322 receive a third strict pass. The implementation reruns channel B on the original dialogue; it does not present the first two candidate annotations to a separate judge. This distinction matters because all three calls use the same model family. The new portion loses 7,522 proposed labels lacking valid answer evidence. Four API failures retain only jointly evidenced labels from the first two channels. No expert human override was applied to this portion.

Automated checks cover JSON fields, allowed labels, Q/A reconstruction, evidence occurrence, ID uniqueness, source lineage, projection consistency, and split overlap. The reported checks pass. The generic Draft 2020-12 JSON Schema engine was unavailable, but independent field and consistency validators ran. These checks do not measure semantic agreement. Appendix C adds human labels for previously unused units; a retrospective human audit of the original 100k training and development corpus remains unperformed.

A.3 API IDENTITY AND RETAINED REQUEST POLICY

The exact requested model ID was `qwen3.8-max`, served by Alibaba Cloud Model Studio (DashScope) through its OpenAI-compatible endpoint. The provider lists this ID officially.[1] The execution record dates the new 77,550-unit build and annotation run to 14 August 2026. Earlier records belong to the frozen 22,450-unit release; per-request timestamps were not retained, so the build date should not be read as a timestamp for every annotation in the corpus.

Table 5: Retained annotation configuration and client behavior. Defaults are identified as such; an omitted API parameter is not assigned a reconstructed value.

Item	Recorded setting or behavior
Provider / model ID	Alibaba Cloud Model Studio / `qwen3.8-max`
Endpoint	`https://dashscope.aliyuncs.com/compatible-mode/v1/chat/completions`
Sampling	`temperature=0.1`; `top_p` omitted (historical server default not recorded)
Other request fields	`enable_thinking=false`; JSON-object response format; no explicit seed or output-token limit
Default timeout	20 s connection; 420 s response read
Default attempt limit	Six total attempts per client invocation, including the initial request
Retry triggers	Transport/HTTP failures, invalid JSON, missing fields, or annotation-ID mismatch
Retry delay	$\min(60, 2^j + U[0,1))$ seconds, with zero-based failed-attempt index j
Recovery	Completed IDs skipped on resume; failed batches retried with smaller batches, down to one record
Saved response metadata	Requested alias, request ID, token usage, and normalized annotation; no resolved snapshot or per-request timestamp

Table 5 follows `configs/experiment.json` and `annotate_dialogue_values.py`. The six-attempt limit applies to one invocation, not to all subsequent recovery runs for a record. Retained logs show ten-record batches and later smaller-batch retries. Four persistent third-pass failures use the exact-evidence intersection of channels A and B. Because the client records the requested alias rather than the provider-resolved model version, an exact historical snapshot cannot be recovered from these artifacts. In particular, we do not substitute the currently documented September snapshot for the August calls.

A.4 CLASSIFIER TRUNCATION AND EVIDENCE RETENTION

Annotation uses full stored turns; classifier training subsequently serializes each nonempty turn as its role marker followed by its text, joined by newlines. Speaker names and source IDs are not added to this classifier string. Both branches receive that same string, but tokenize it independently with three

[1] `https://help.aliyun.com/en/model-studio/qwen3-8-max`, checked 23 September 2026. This current listing does not establish which historical snapshot served our requests.

added role tokens. The retained fast tokenizers use right truncation to 384 tokens including special tokens. This keeps the beginning of the dialogue and can remove later answers.

We reran tokenization over all frozen splits using Transformers 4.38.2 and the retained snapshots (MPNet `e8c3b32e`, DeBERTa `8ccc9b6f`). Full revision IDs, split SHA-256 hashes, counts, and the script accompany the audit report. A dialogue is truncated if its full tokenized length exceeds 384. To measure evidence retention, we locate each cited quote inside its named respondent turn, allowing case and whitespace normalization, then map that occurrence to serialized-text character offsets. A quote is retained only if one full occurrence lies within the retained prefix. All 74,531 evidence entries were located before truncation.

Table 6: Right-truncation and answer-evidence retention (%). Within each encoder, columns report truncated dialogues / fully retained cited spans / positive dialogue–label pairs with at least one retained span. Denominators are respectively all dialogues, all evidence entries, and all positive pairs in the indicated group.

Group	Dialogues	Positive pairs	MPNet	DeBERTa
Train	79,750	48,319	10.64 / 95.91 / 96.26	9.74 / 96.39 / 96.70
Validation	10,613	6,455	10.20 / 96.13 / 96.58	9.37 / 96.56 / 97.04
Test	9,637	5,623	11.54 / 95.36 / 95.77	10.55 / 95.76 / 96.09
Full corpus	100,000	60,397	10.68 / 95.88 / 96.25	9.78 / 96.35 / 96.68
Single-turn	60,725	35,289	5.52 / 97.58 / 97.70	4.92 / 97.90 / 98.01
Multi-turn	39,275	25,108	18.67 / 93.66 / 94.21	17.28 / 94.32 / 94.82

Across all 60,397 positive dialogue–label pairs, 58,132 (96.25%) retain at least one cited span in both branches and 58,393 (96.68%) in at least one branch. Thus 2,004 pairs have no fully retained cited span in either branch. Four training dialogues lose their first answer marker in both tokenizers. Multi-turn examples incur greater truncation and evidence loss than single-turn examples. Retaining a quote does not certify sufficient surrounding context or semantic label validity; losing it also does not rule out uncited supporting text in the prefix. Labels are not recomputed after truncation. This audit leaves the evaluated inputs, thresholds, and classifier results unchanged.

A.5 Taxonomy and support

Table 7: Primary labels, full-corpus positives, and higher-order membership. The first group sets figure color only. UN-O is excluded from the classical ring.

ID	Value	Positives	Higher-order membership
SD-T	Self-direction: thought	1,778	Openness
SD-A	Self-direction: action	2,449	Openness
STI	Stimulation	783	Openness
HED	Hedonism	1,291	Openness; Self-enhancement
ACH	Achievement	5,657	Self-enhancement
PO-D	Power: dominance	276	Self-enhancement
PO-R	Power: resources	313	Self-enhancement
FAC	Face	435	Self-enhancement; Conservation
SE-P	Security: personal	4,926	Conservation
SE-S	Security: societal	6,843	Conservation
TRA	Tradition	1,477	Conservation
CO-R	Conformity: rules	4,335	Conservation
CO-I	Conformity: interpersonal	850	Conservation
HUM	Humility	911	Conservation; Self-transcendence
BE-C	Benevolence: caring	5,976	Self-transcendence
BE-D	Benevolence: dependability	1,352	Self-transcendence
UN-C	Universalism: concern	11,308	Self-transcendence
UN-N	Universalism: nature	933	Self-transcendence
UN-T	Universalism: tolerance	1,856	Self-transcendence
UN-O	Universalism: objectivity	6,648	Self-transcendence

A.6 SCHEMA AND BASELINE PROJECTION

The canonical release retains these fields (source URLs may be empty):

```
qa_id, schema_version, conversation_id,
source{dataset, dataset_key, version, license,
       original_conversation_id, program, date,
       url, title, archive},
interviewers, respondents, dialogue_type, annotation_target,
question_turn_ids, answer_turn_ids, turns,
question_text, answer_text, dialogue_text, model_input,
labels_20, fine_values_54, labels_10, labels_4,
polarity, evidence, rationale, confidence, adjudication,
annotation_model, split, respondent_group_id, release_version
```

Only the public-data augmentation experiment in Appendix F replaces the dialogue with `answer_text`, without relabeling. The interview baselines in Table 1 use the canonical `model_input`, including stored role, turn-ID, and speaker strings. SRLCF reconstructs the same turn text with role tokens and newlines, omitting those identity strings. These native serialization choices are retained in the human evaluation. IDs join the answer projection to the primary release; discarded context cannot be recovered from the answer text alone.

B Implementation and Evaluation Protocols

B.1 Loss definitions and numerical conventions

Table 8 instantiates the symbolic objective in Equation 7. The fused ASL coefficient is fixed to one as the reference scale. All loss coefficients and the gate prior remain fixed during training, while the per-value gate parameters are learned with the network.

Table 8: Loss coefficients and prior used in the final model. Ablation-specific changes are identified in the matched-comparison protocol.

Symbol	Term or setting	Value
λ_z	Fused-logit weighted BCE	0.05
λ_s	Semantic-logit weighted BCE	0.05
λ_d	Direct-branch ASL	1.00
λ_c	Fine-value representation ASL	0.25
λ_h	Four-group representation BCE	0.02
λ_a	Semantic alignment	0.005
λ_b	Fused-probability Brier loss	0.02
λ_g	Mean gate-prior penalty	0.01
α_0	Gate initialization and prior center	0.35

Let B be the batch size, $K = 20$, and n_v the training support of label v. For dialogue D_i, write $p_{iv} = p_{D_i,v} = \sigma(z_{iv})$ and $y_{iv} = y_{D_i,v}$. The positive weight in the fused and semantic BCE terms is

$$w_v^+ = \mathrm{clip}_{[1,10]} \sqrt{(N_{\mathrm{train}} - n_v)/n_v}. \tag{8}$$

For zero support the implementation uses weight 1. The weighted BCE averages $-[w_v^+ y_{iv} \log p_{iv} + (1 - y_{iv}) \log(1 - p_{iv})]$ over all BK pairs. Semantic BCE uses $\sigma(s_{iv})$ in the same expression. Higher-order BCE is unweighted and averages over $4B$ outputs.

For ASL, define $q_{iv} = \min(1, 1 - p_{iv} + 0.05)$. The pair loss is

$$\ell_{iv}^{\mathrm{ASL}} = -y_{iv}(1 - p_{iv})^1 \log p_{iv} - (1 - y_{iv})(1 - q_{iv})^4 \log q_{iv}. \tag{9}$$

Probabilities inside logarithms are clipped below at 10^{-8}; the mean over pairs gives the ASL terms for the fused, direct, and representation heads. These terms have separate logits and shared targets.

The alignment term in Equation 7 is defined over dialogues with at least one positive label. For $P_i = \{v : y_{iv} = 1\}$ and $I_+ = \{i : |P_i| > 0\}$,

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{|I_+|} \sum_{i \in I_+} \frac{1}{|P_i|} \sum_{v \in P_i} \log \frac{\exp(s_{iv})}{\sum_{u=1}^{20} \exp(s_{iu})}. \tag{10}$$

It is zero if the batch has no positive dialogue. The softmax is confined to this auxiliary term; inference remains multi-label sigmoid prediction. Brier loss is $(BK)^{-1} \sum_{i,v} (p_{iv} - y_{iv})^2$. The gate prior is a mean over 20 values, as written in Equation 7; an unnormalized squared norm would multiply its strength by 20.

Role-gate weights start at zero and their bias at -1.5. MPNet uses attention-mask mean pooling; DeBERTa uses its first-token state. If no answer marker survives, pooling returns $a_b = x_b$ and $m_b = 0$, hence $g_b = 0$ and $h_b = \mathrm{LN}(x_b)$. The no-role ablation bypasses this residual module, including its LayerNorm, and returns x_b. The semantic scale starts at 10 and is capped at 100. Specifically, W_s and W_v are separate bias-free 768×768 maps; the representation projection W_c has bias b_c. MPNet is shared between dialogue and definition encoding.

B.2 Training settings and baseline adaptations

Validation Macro-F1 selects checkpoints with patience 1. Each label threshold is searched over $0.02, 0.04, \ldots, 0.98$; ties favor greater recall and then the lower threshold. A label absent from validation receives threshold 1. Branch-only evaluations retune thresholds on that branch's validation

Table 9: Final model hyperparameters.

Setting	Value	Setting	Value
Maximum tokens	384	Maximum epochs	4
Train batch/GPU	6	Accumulation	4
Effective batch	24	Evaluation batch	12
Backbone LR	2×10^{-5}	Head LR	10^{-4}
Weight decay	.01	Warmup ratio	.08
Gradient norm	1.0	Dropout	.1
ASL negative/positive powers	4/1	Negative clipping	.05
Representation dimension	768	Threshold step	.02
Output labels	20	Seeds	42, 7, 1701

Table 10: Baseline adaptations to the common interview task.

Method	Input and objective	Reproduction scope
TARS	Dialogue–definition pairs; binary BCE	TARS formulation adapted to DeBERTa
NLI	Dialogue premise and value hypothesis	Unified backbone and fine-tuning protocol
Dual SetFit	Shared MPNet; BCE + contrastive loss	Label-description adaptation of SetFit
LEAM	Label-aware DeBERTa attention	Modernized encoder and common outputs
UCLAF	Label-aware interaction and CACL	Reimplemented equations; code not found
DeBERTa + ASL	DeBERTa with 20 sigmoid outputs	Project classifier with ASL
Schwartz decoder	Direct probabilities and ring energy	Official code; label 20 independent
TF–IDF OvR	Word/character features; logistic heads	Non-neural reference

probabilities. The evaluated branches belong to the jointly trained model, not independently trained standalone models.

The Schwartz decoder adapts the energy protocol associated with the archived implementation (commit bdf5f12). Its retained run selects neighbor reward .10, opposition penalty .20, and cardinality penalty .00 from eight settings, with at most eight candidates and five outputs. Validation Macro-F1 is .4618 before decoding and .4580 afterward, satisfying the 99% floor before geometry-cost selection. The underlying classifier is DeBERTa trained with BCE, not the separately trained ASL baseline; their test-score difference cannot be attributed solely to decoding.

The local implementation returns a maximum-energy set for F1 and continuous marginals for probability metrics: candidate-set weights are proportional to $\exp(E(S))$ at temperature one, and each label score sums weights of sets containing that label. Noncandidate ring labels have score zero; Universalism: objectivity keeps its base probability. Retained test scores include 86,733 strictly interior values among 192,740 pairs, so they are not simply binary set indicators. ECE and AUPRC therefore describe these energy marginals rather than raw backbone probabilities. This output convention differs from evaluating the discrete set alone.

B.3 Matched ablations and run identifiers

Table 2 contains several matched pairs, not one common removal sequence. The base configuration uses direct auxiliary weight .25, no fine-value auxiliary, and backbone LR 10^{-5}. Its gate and hierarchy removals change only the named setting. The role comparison uses direct auxiliary weight 1.0, no fine-value auxiliary, and LR 2×10^{-5}. The LR comparison uses fine-value weight .25 and direct weight 1.0 throughout. The final fine-value comparison changes only its auxiliary weight from 0 to .25 under LR 2×10^{-5}. Table 11 maps these descriptions to the retained method identifiers.

Historical artifact names containing `control`, including `control_head` and `test_control`, refer to the fine-value auxiliary head and dialogue-level representation. We retain these lookup identifiers, while figure labels use the descriptive terms; no generation-control module is implied.

Table 11: Internal identifiers used only for artifact lookup.

Paper description	Retained identifier
Final SRLCF	`srlcf_control_aux25_lr2e5`
Without fine-value auxiliary	`srlcf_direct_aux1_lr2e5`
Matched role removal	`srlcf_direct_aux1_no_role`
Base configuration	`schwartz_role_fusion`
Base, fixed .35 gate	`srlcf_fixed_global35`
Base, hierarchy removed	`srlcf_no_hierarchy`
Fine-value auxiliary, LR 10^{-5}	`srlcf_control_aux25`

B.4 Frozen-probe protocol

For every model seed, the retained validation and test representations are frozen. A seed-controlled permutation divides the 10,613 validation units into 8,490 probe-training units and 2,123 threshold-tuning units. A one-vs-rest averaged SGD logistic regression is fit with `loss=log_loss`, L2 penalty, regularization `alpha=1e-4`, balanced class weights, 2,000 maximum iterations, and tolerance 10^{-4}. Thresholds use the same step-.02 rule on the probe development partition. Final metrics use all 9,637 test units. The probe is implemented in `evaluate_control_probe.py`. It is a fresh classifier, not the trained auxiliary head. Its training data are the validation representations, not the encoder's training split; the reported scores should be read under that protocol.

B.5 Metrics and uncertainty

Macro-F1 averages 20 per-label F1 scores; Micro-F1 pools all pairs. Macro-AUPRC uses unthresholded probabilities. Brier and 15-bin ECE pool dialogue–label pairs; their interpretation is affected by the predominance of negatives. Retained metrics also include MCC, label-ranking average precision, nDCG@3, subset accuracy, Hamming loss, and source/turn groups. Seeds in the main comparison are 42, 7, and 1701; augmentation uses 42, 43, and 44.

Interview bootstrap comparisons use 5,000 paired resamples of dialogue rows, keeping all labels together and fixing validation thresholds. The smallest reportable two-sided p is $2/(5000+1)$. The method-development use of test diagnostics is not corrected by this procedure. Dialogues from one conversation can also remain dependent within a split; resampling individual dialogues does not model that dependence.

Table 12: Paired Macro-F1 comparisons for final SRLCF. All with/without fine-value intervals include zero.

Comparator	Seed	Δ	95% CI	p
DeBERTa + ASL	42	+.0257	[.0097, .0421]	.0024
DeBERTa + ASL	7	+.0243	[.0087, .0403]	.0012
DeBERTa + ASL	1701	+.0303	[.0148, .0454]	.0004
Without fine-value auxiliary	42	+.0018	[-.0087, .0120]	.732
Without fine-value auxiliary	7	-.0017	[-.0119, .0081]	.758
Without fine-value auxiliary	1701	-.0010	[-.0105, .0082]	.839

C Human-annotated Positive-only Evaluation

C.1 Source isolation and evaluation population

The human evaluation contains 1,000 MediaSum dialogues, each from a distinct conversation outside the original training, validation, and development-test splits. The source screen uses the previously structured MediaSum archive and excludes known historical sample IDs, conversations, source links, respondent identifiers/names, conservative surname matches, and exact text duplicates. Binary hashed word-trigram cosine screening rejects similarity at least 0.80 against historical answer texts and within the candidate set. These approximate checks reduce observed overlap; aliases and paraphrases can escape them. Conservative surname exclusions also alter the population. This is a same-archive evaluation, not new-source or temporal generalization.

The final human-confirmed file contains only dialogues with at least one adjudicated positive label. The organizers removed records judged invalid or inadequately annotated before delivering this file; item-level exclusion reasons are not retained in the evaluation artifact. We evaluate every delivered record without additional sampling or relabeling. This quality-screened set estimates value-label identification conditional on a retained dialogue containing an annotated expressed value. It cannot estimate false alarms on entirely value-negative dialogues or prevalence in the source archive, and quality screening may change the difficulty distribution.

C.2 Independent A/B annotation and adjudication

Twenty currently enrolled master's students were randomly recruited as volunteer annotators, without a restriction on academic discipline. For each assigned dialogue, two annotators, denoted A and B, independently inspect the full, role-preserving turns and assign 20 binary expression labels. A third person makes the final decision on disagreements. The consolidated file contains the resulting adjudicated label vector. This human workflow is distinct from the same-model automated channels used for the 100k training release. The recruitment description does not establish specialist expertise in value theory, and no English-proficiency screening record was supplied.

Annotators receive a pseudonymous sample ID, ordered turns with role and speaker aliases, and a dictionary of 20 Chinese-named labels initialized to zero. No model predictions or former value labels are shown. Instructions define answer-evidenced expression: a value can be promoted, constrained, rejected, or negotiated. A question alone does not license a positive label. Annotators inspect every label, set supported labels to one, and leave others at zero without changing the dialogue. A positive label denotes expression, not endorsement or approval.

The final artifact passes unique-ID, complete-taxonomy, binary-label, and unchanged-text checks. A name-based mapping converts Chinese label keys to the fixed English taxonomy order. We use the adjudicated labels without semantic relabeling or assistant-generated exclusion flags. The supplied artifact does not retain the separate A/B decisions or item-level adjudication notes. Consequently, we report the confirmed process but do not reconstruct inter-annotator agreement or disagreement rates from the final labels.

C.3 Final evaluation file and timing

The evaluation input is the complete organizer-confirmed file of 1,000 records, with SHA-256 beginning `db1487504e77`. We retain its exact IDs, order, dialogue text, and final label vectors. No new sampling, balancing, or semantic filtering is performed during evaluation. Stable IDs link each record to its saved model outputs, and name-based label alignment is verified before scoring.

The resulting set has 843 single-turn and 157 multi-turn dialogues. There are 528 dialogues with one label, 290 with two, and 182 with three or more; the mean label cardinality is 1.719. Every label has positive support, ranging from 8 to 273 examples. Table 16 gives all 20 supports.

The final evaluation population was confirmed after an earlier human evaluation had been inspected. This is therefore a retrospective, fixed-model analysis rather than a new blind confirmatory test. We freeze the supplied file before recomputing its statistics; no model, output choice, threshold, or probe setting is changed. Bootstrap intervals quantify sampling variability conditional on the retained population, not quality-screening decisions or prior result inspection.

C.4 Models, inputs, and uncertainty

SRLCF uses the retained deployment checkpoint described in Appendix G, SHA-256 beginning `6aac6ea79603`. Its configuration and four output-specific threshold vectors are unchanged. The semantic, direct, and 20-label auxiliary heads belong to this same jointly trained model; they are not independently trained baselines.

The independent DeBERTa + ASL comparator uses seed 42, the original 79,750 training and 10,613 validation units, and the unchanged benchmark recipe. Early stopping and thresholds depend only on the original machine-labeled validation data. This rerun is not the deleted historical seed-42 checkpoint, and its scores are kept separate from the old three-seed summaries. Native input serialization is retained for each model (Appendix A), without human-evidence-based cropping.

We reuse saved predictions for exactly the 1,000 fixed IDs. Model weights, label order, probability arrays, exact decisions, thresholds, and manifests identify these outputs. For ASL, exact saved decisions avoid threshold-boundary changes caused by float16 probability serialization. Reusing outputs avoids unnecessary retraining and keeps operating points constant.

Macro-F1 averages all 20 binary-label F1 scores; a zero denominator contributes zero. Each of 2,000 paired bootstrap resamples (seed 20260925) samples dialogue rows with replacement and retains all labels. Every row is a distinct conversation. Percentile intervals use the 2.5th and 97.5th percentiles, with thresholds fixed in every draw. Intervals condition on the positive-only population, model runs, and adjudicated labels; they omit training-seed and annotation uncertainty. Paired differences, rather than overlap of separate model intervals, describe comparative uncertainty. Diagnostic branch comparisons have nominal intervals without multiplicity adjustment.

Micro-averaged precision is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and recall is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ over all dialogue–label pairs. A positive-only dialogue still supplies negative labels for values it does not express. Thus, incorrect value assignments remain false positives even though entirely all-zero dialogues are absent.

C.5 Complete results on the 1,000 dialogues

Table 13: Human positive-only evaluation. Each output uses a fixed model run; joint heads share the SRLCF checkpoint. The frozen linear probe is a supplementary representation diagnostic. Precision and recall are micro-averaged.

Model / output	Macro-F1	Micro-F1	Macro-AP	Precision	Recall
Independent DeBERTa + ASL	0.4746	0.5331	0.5835	0.6881	0.4351
SRLCF fusion	0.4944	0.5719	0.5976	0.7288	0.4706
Joint semantic	0.4631	0.5403	0.5526	0.6931	0.4427
Joint direct	0.4595	0.5489	0.5628	0.7014	0.4508
Fine-value auxiliary	0.4905	0.5661	0.5921	0.7399	0.4584
Frozen linear probe	0.4675	0.5407	0.3255	0.4326	0.7208

Table 14: Paired Macro-F1 differences, SRLCF fusion minus comparator, on the same 1,000 dialogues. Intervals are nominal and conditional on fixed runs; branch comparisons are diagnostic rather than independent replications.

Comparator	Δ	95% CI
Independent DeBERTa + ASL	$+0.0198$	$[-0.0117, +0.0523]$
Joint semantic	$+0.0312$	$[+0.0082, +0.0523]$
Joint direct	$+0.0349$	$[+0.0077, +0.0603]$
Fine-value auxiliary	$+0.0038$	$[-0.0149, +0.0214]$
Frozen linear probe	$+0.0268$	$[-0.0078, +0.0562]$

SRLCF and independent ASL reach Macro-F1 0.494 and 0.475, respectively. The paired difference $+0.020$ has interval $[-0.012, +0.052]$: its positive point estimate is not statistically resolved at the nominal 95% level. SRLCF also has higher Micro-F1 and Macro-AP point estimates. Reporting all three metrics separates equal-weighted label coverage, pooled decisions, and threshold-free ranking.

Fusion exceeds its joint semantic and direct outputs by 0.031 and 0.035 Macro-F1, with both nominal paired intervals above zero. The auxiliary head's Macro-F1 is 0.491 versus fusion's 0.494, with a paired interval spanning zero. Together with the development ablations, these diagnostics support complementary branch predictions without attributing the gain to higher-order supervision alone. They compare outputs, not matched retraining ablations, and do not select a different output on the evaluation set.

The ASL rerun completed four epochs in 4,813 seconds, with best original-validation Macro-F1 0.51468. Its environment used Python 3.12.13, PyTorch 2.11.0, Transformers 4.38.2, Tokenizers 0.15.2, NumPy 2.3.5, and scikit-learn 1.9.0. The retained deployment artifact and this rerun have different runtime histories. The temporary best-validation state was removed after prediction export, and the cleanup audit found no remaining weights in the new run directory.

Frozen linear-probe follow-up. We reuse the historical probe recipe without searching new settings. A one-vs-rest averaged SGD logistic classifier fits 8,490 original validation vectors, with thresholds selected on the remaining 2,123. Settings are ℓ_2 regularization 10^{-4}, balanced class weights, at most 2,000 iterations, tolerance 10^{-4}, and seed 42. Recomputing the retained deployment model's old test score reproduces Macro-F1 0.3218606 exactly. On the 1,000 human-annotated dialogues, the probe reaches Macro-F1 0.468 with interval $[0.447, 0.485]$, Micro-F1 0.541, and Macro-AP 0.326. Neither fitting nor threshold selection uses human labels. This measures linear recovery from c, not a gain from auxiliary supervision relative to a matched no-auxiliary checkpoint, which is unavailable for these dialogues. The probe follow-up was added after initial human-output inspection and is supplementary to the fixed primary comparison.

Table 15: Descriptive subgroup Macro-F1 on the human evaluation set. All 20 labels remain in each score, including unsupported labels assigned zero F1. Small groups have unstable support; strata are not independent replications.

Stratum	Dialogues	ASL Macro-F1	SRLCF Macro-F1
Single turn	843	0.471	0.480
Multi turn	157	0.449	0.512
One value	528	0.447	0.484
Multiple values	472	0.479	0.504

Both turn-structure groups and both label-cardinality groups have higher SRLCF Macro-F1 point estimates (Table 15). The larger multi-turn gap is consistent with the intended use of dialogue context, but does not isolate the answer-marker mechanism: architectures differ and the subgroup is small. These comparisons remain descriptive because support and label mixtures differ. We report all four predefined groups.

Aggregate scores hide substantial variation. Achievement has 171 positives, of which SRLCF recovers 49 with three false positives: precision is 0.942 but recall is 0.287 at the unchanged threshold. Face recovers two of 26 positives, while Power: dominance recovers none of 14. Power: resources has only eight positives, making its F1 especially unstable. Future operating-point adaptation requires a separate human development set; these evaluation labels are not used to lower thresholds or change the classifier.

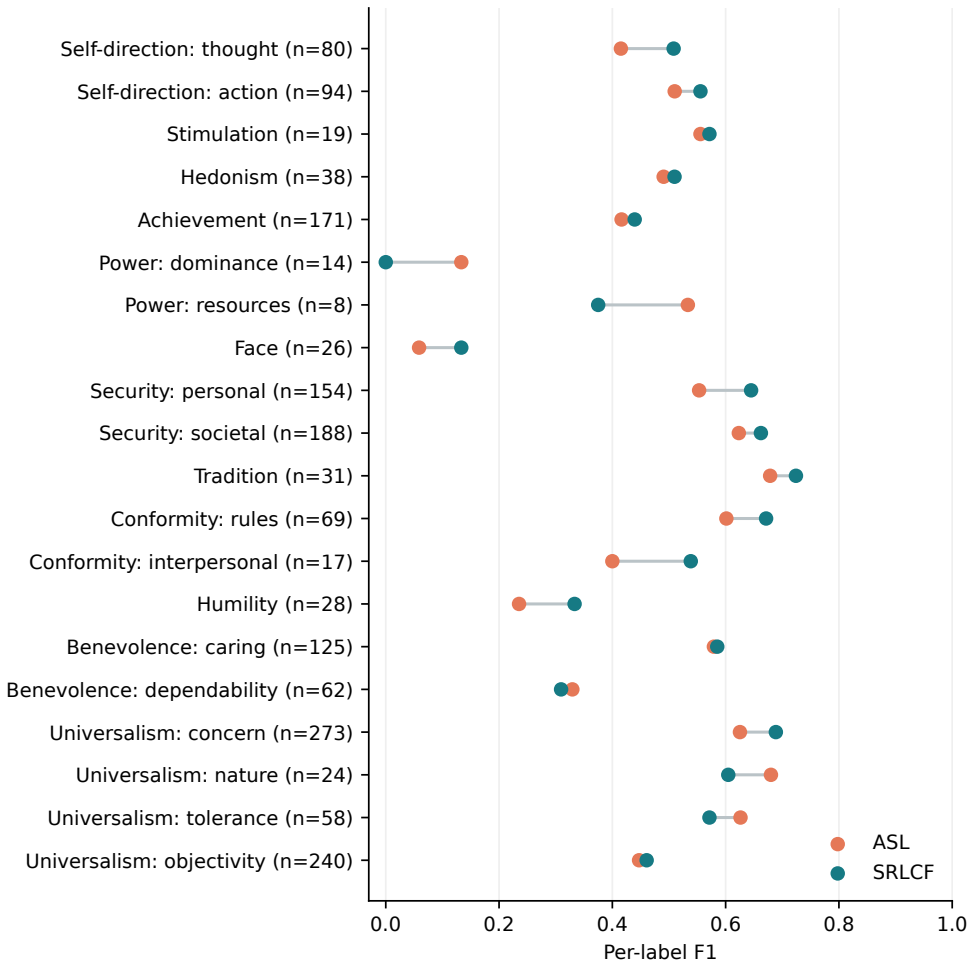


Figure 7: Per-value F1 on the same 1,000 human-annotated dialogues, with positive support in parentheses. Connecting lines show fixed-model point differences, not uncertainty intervals.

Table 16: Per-value support and F1 on the human positive-only set. TP/FP/FN refer to SRLCF fusion and unchanged adjudicated labels. Small-support values require caution.

Value	n_+	ASL F1	SRLCF F1	TP	FP	FN
Self-direction: thought	80	0.415	0.508	31	11	49
Self-direction: action	94	0.510	0.556	40	10	54
Stimulation	19	0.556	0.571	8	1	11
Hedonism	38	0.491	0.510	13	0	25
Achievement	171	0.416	0.439	49	3	122
Power: dominance	14	0.133	0.000	0	1	14
Power: resources	8	0.533	0.375	3	5	5
Face	26	0.059	0.133	2	2	24
Security: personal	154	0.553	0.645	99	54	55
Security: societal	188	0.623	0.662	100	14	88
Tradition	31	0.679	0.724	21	6	10
Conformity: rules	69	0.601	0.671	47	24	22
Conformity: interpersonal	17	0.400	0.538	7	2	10
Humility	28	0.235	0.333	6	2	22
Benevolence: caring	125	0.579	0.585	91	95	34
Benevolence: dependability	62	0.329	0.310	13	9	49
Universalism: concern	273	0.625	0.689	167	45	106
Universalism: nature	24	0.680	0.605	13	6	11
Universalism: tolerance	58	0.626	0.571	26	7	32
Universalism: objectivity	240	0.447	0.461	73	4	167

D Extended Detection Results

Figures 3(a) and 2(c) report source/turn results and calibration in the main text. The four source means for SRLCF are .4988 (MediaSum), .4813 (NPR-2P), .3765 (Tennis), and .3865 (NewsInterview). Small source groups remain difficult. This appendix expands branch and detection comparisons.

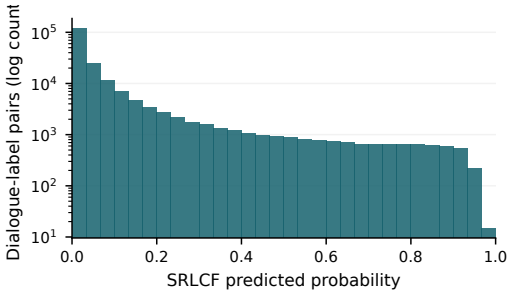


Figure 8: Seed-42 probability distribution underlying Figure 2(c), with log-scaled counts. Low-probability pairs predominate. Main-table ECE averages the available seeds.

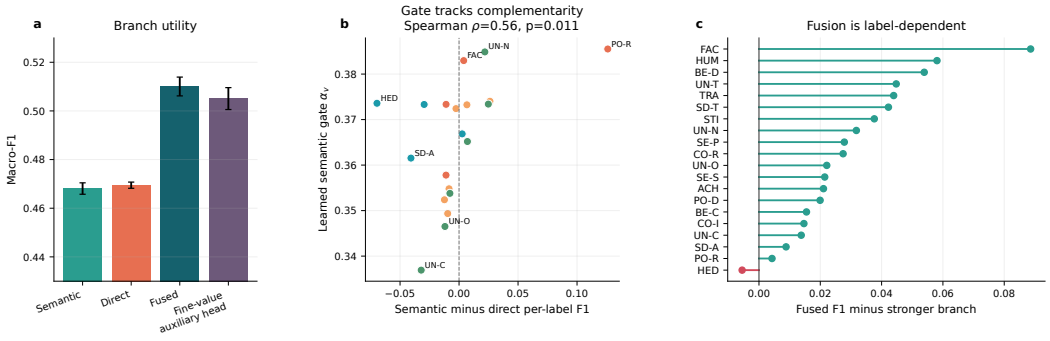


Figure 9: Jointly trained SRLCF branches: scores, gate association, and fusion gain over the stronger branch per value. These outputs are not independently trained baselines.

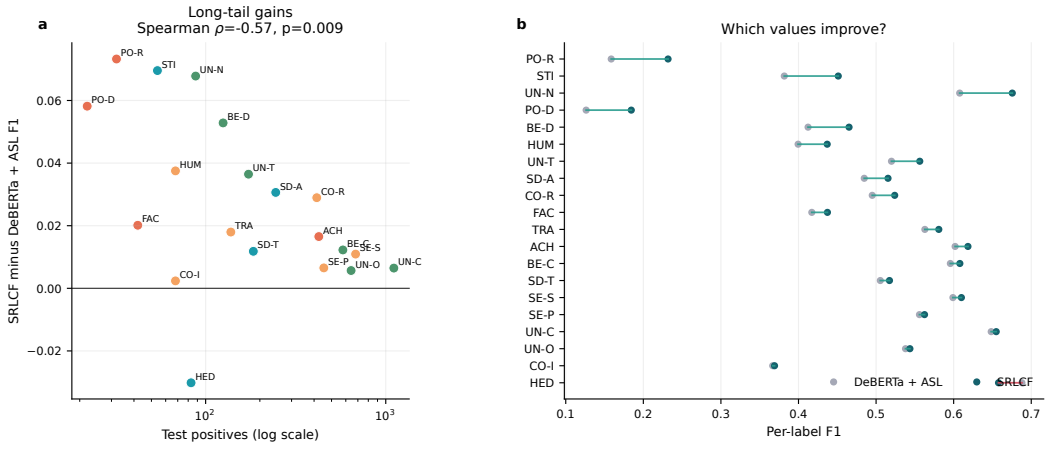
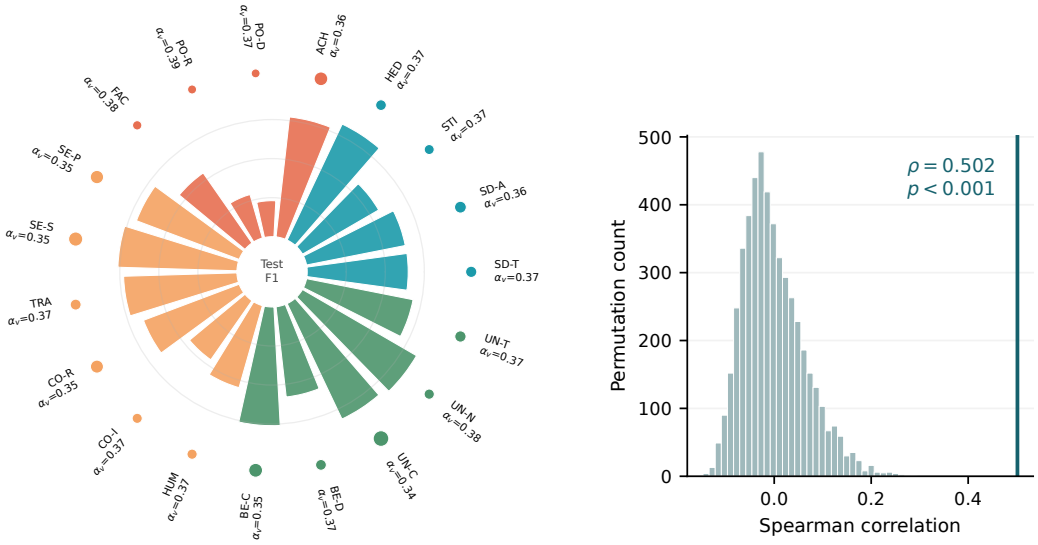


Figure 10: Test-support association and complete per-label comparison with DeBERTa + ASL. Gains are not uniform: Hedonism is lower by about .030.

E Extended Representation Analyses



(a) Per-value audit

(b) Whole-label permutation null

Figure 11: **Development-set Schwartz audits.** (a) Bar length is three-seed mean F1; dot size reflects full-corpus support with square-root scaling; α_v is the mean semantic logit weight. Colors mark primary groups; UN-O is excluded. (b) Seed-42 centroid correlation versus 5,000 label-order permutations. This descriptive analysis precedes the human follow-up.

For theoretical positions i, j among the 19 classical values, circular distance is $r(i, j) = \min(|i - j|, 19 - |i - j|)$. Each centroid averages original test representations over label-positive dialogues for that value. Correlation compares the upper triangle of the cosine-distance matrix with the upper triangle of r. Permutation shuffles value identities 5,000 times while retaining the learned distances, with random seed 42. The 171 label pairs are dependent; identities rather than pairs are permuted. PCA below is illustrative and is not used to calculate the reported correlations. Universalism: objectivity is excluded.

The reported circular-error metric uses normalized distance $d_{uv} = 2r(u, v)/19$ and score matrix p:

$$\mathrm{CE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{u,v \in \mathcal{V}_{19}} y_{iu} d_{uv} p_{iv}}{\max(\sum_{u \in \mathcal{V}_{19}} y_{iu}, 10^{-12})}. \tag{11}$$

An empty gold set contributes zero; a single gold value uses the same formula. There is no normalization by total predicted probability, so this is a distance-weighted activation score, not a bounded set-matching error. Empty thresholded predictions can still have nonzero activation; an identically zero score vector has zero cost. Decoder rows use the energy marginals described above. The metric/decoder code orders the local segment as HUM–BE-D–BE-C–UN-C, whereas the retained descriptive centroid figures use taxonomy order HUM–BE-C–BE-D–UN-C. Each comparison fixes its stated convention; the two analyses are interpreted separately.

Higher-order evaluation collapses fine-label predictions through the taxonomy, allowing overlapping boundary memberships. SRLCF obtains .618 for Openness, .599 for Self-enhancement, .643 for Conservation, and .697 for Self-transcendence. Its separate four-output training head has approximately .650 Macro-F1. That head's accuracy and the fine-label collapse measure different outputs. The one-seed hierarchy removal improves fine-label F1, so group-level accuracy alone does not identify the hierarchy loss as a beneficial classification mechanism.

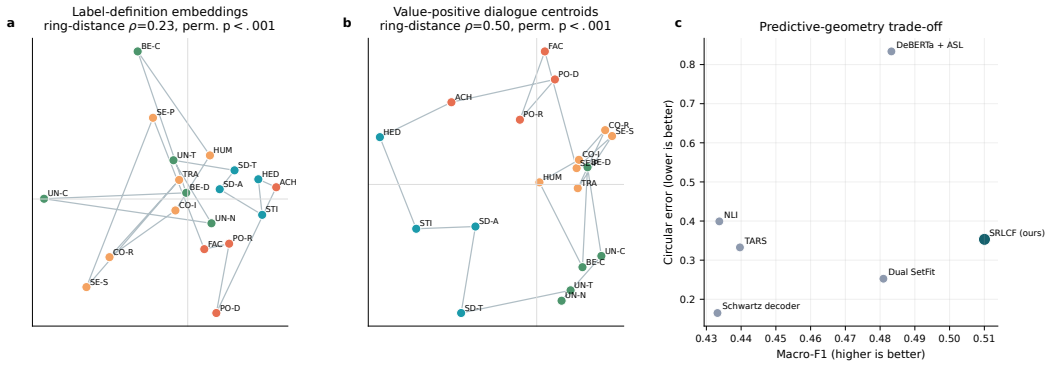


Figure 12: Extended geometry: PCA views of value-definition embeddings and label-positive representation centroids, and the predictive/geometry trade-off. The paths connect theoretical neighbors; they do not demonstrate a recovered geometric circle. Both centroid and embedding analyses use model seed 42.

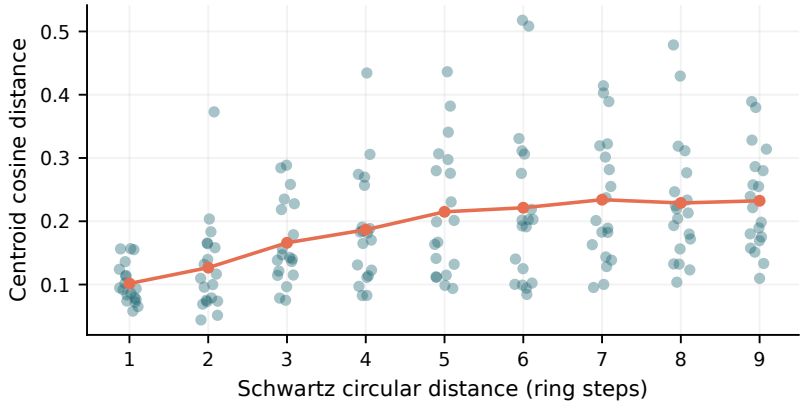


Figure 13: Original-space centroid distances versus theoretical ring separation. Horizontal jitter separates equal ring distances; the line averages cosine distances at each separation. These distances underlie the main heatmaps; Figure 11(b) shows the permutation null.

F PUBLIC-DATA AUGMENTATION

Table 17: Public-data augmentation with 4,000 answer projections from the earlier 22,450-unit pool. Entries are Macro-F1 of three-seed probability ensembles; intervals use 2,000 paired bootstraps with fixed thresholds. Differences use unrounded values. The model is a direct classifier, not SRLCF.

Evaluation set	Frozen route	Public-only	+ Interviews	Δ	95% CI
Webis	webis	0.356	0.372	**+0.016**	[0.002, 0.030]
ValueEval	valueeval	0.424	0.440	**+0.016**	[0.004, 0.028]
Nahj al-Balagha	valueeval	0.183	0.222	**+0.039**	[0.013, 0.065]
Axies-COVID	webis	0.188	0.220	**+0.032**	[-0.021, 0.083]
Axies-Energy	webis	0.140	0.183	**+0.042**	[-0.045, 0.094]

Ensemble test Macro-F1 differences are positive on all five sets (Table 17); intervals exclude zero for Webis, ValueEval, and Nahj. Both Axies intervals cross zero, with 100 examples per domain. Webis Micro-F1 falls from 0.4571 to 0.4397, and Axies-Energy from 0.2381 to 0.2275. This auxiliary result concerns the earlier 4k subset under a direct-classifier protocol, not external validation of SRLCF or the 100k corpus.

F.1 VALIDATION SELECTION AND ITS AUDIT

The selected subset comes from the earlier 22,450-unit interview release, using its documented answer projection. Training uses a direct 20-label DeBERTa classifier. Stage one ranks nested 1k/2k/4k candidates on public validation with seed 42 and retains the top two per source: 2k and 4k for both sources. Stage two repeats those candidates with seeds 43 and 44. It averages validation probabilities across three seeds, then tunes thresholds and computes Macro-F1. Selection maximizes the resulting ensemble improvement, with ties broken by the worst single-seed improvement and then smaller augmentation size. The 1k candidate was not run for three seeds and is not an ensemble candidate in stage two.

The historical figure plotted mean individual-seed F1 but described the curve as the selection criterion. Figure 14 and Table 18 distinguish these quantities. For ValueEval, mean seed F1 favors 2k (.4515 versus .4445), while ensemble F1 favors 4k (.4589 versus .4512). Averaging probabilities and then thresholding is not equivalent to averaging F1 scores. The revision recomputes all four shortlisted ensemble scores directly from retained validation predictions and matches the original selection record, without reading test predictions or changing the chosen sizes.

Table 18: Public-validation selection audit. The ensemble column is the selection criterion.

Source	Added examples	Mean seed F1	Ensemble F1	Ensemble Δ
Webis	0	0.5059	0.5238	–
	2000	0.5171	0.5264	+0.0026
	4000 (selected)	0.5254	0.5311	+0.0074
ValueEval	0	0.4459	0.4448	–
	2000	0.4515	0.4512	+0.0064
	4000 (selected)	0.4445	0.4589	+0.0141

Thresholds in the augmentation pipeline use a .01 grid with minimum precision .4 where feasible; otherwise the full grid is considered. Ties prefer recall and then lower thresholds. This differs from the interview benchmark's unconstrained .02 grid. Thresholds are fitted to the averaged validation probabilities of each condition. The control and augmented conditions therefore have separately selected thresholds, fixed before their final test evaluation.

F.2 ROUTES, TEST METRICS, AND UNCERTAINTY

The retained pre-test routing rule sends Webis and both Axies domains through the Webis-trained model, and ValueEval and Nahj through the ValueEval-trained model. Axies evaluation uses the ten-value mapping. Three-seed probabilities are averaged within each condition. Each of 2,000

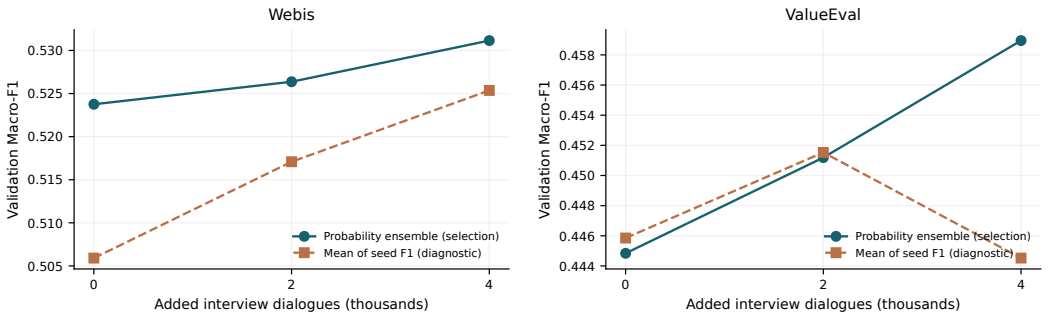


Figure 14: Probability-ensemble validation F1 (used for selection) versus the mean of individual-seed F1 (diagnostic only). Neither curve uses public test scores.

paired bootstrap replicates resamples example rows, retains all their labels, and reapplies the fixed condition-specific thresholds; it does not retune thresholds or select models. Confidence intervals are percentile intervals of the Macro-F1 differences.

Table 19: Additional Micro-F1 results for the fixed 4k augmentation study.

Evaluation set	Public only	With interviews	Difference
Webis	.4571	.4397	-.0174
ValueEval	.5573	.5593	+.0020
Nahj al-Balagha	.2514	.2811	+.0297
Axies-COVID	.2129	.2282	+.0153
Axies-Energy	.2381	.2275	-.0106

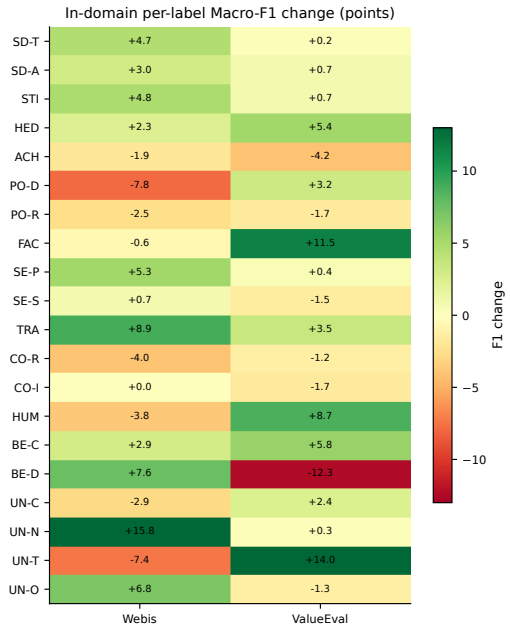


Figure 15: In-domain per-label F1 changes after selected augmentation. Negative cells remain despite positive dataset-level Macro-F1 changes.

F.3 Development chronology

Earlier 1k experiments had mixed public-data results and motivated the later validation-screening protocol. The selected 4k runs and routes were fixed before their final test stage, but this is not a claim that all prior public test information was unknown throughout the research program. A later synthetic symmetric-augmentation study was formalized after test access and remains exploratory; it is excluded from the main table. The 4k experiment evaluates neither SRLCF nor the full 100k corpus. An independent replication using the larger release has not been conducted here.

G Artifacts and Extended Limitations

G.1 Manuscript and experiment artifacts

Compile `main_submission_v6.tex` with the included ICLR style and `sections/v6/`. Statistical figures are vector PDFs; the architecture includes raster icons:

```
latexmk -pdf -outdir=build main_submission_v6.tex
```

The archive includes all typesetting dependencies. The workspace script `scripts/evaluate_confirmed_human1000.py` recomputes the human-set tables and figures from the complete final file and saved predictions. `scripts/evaluate_human_probe.py` reproduces probe fitting on the original validation vectors.

`analysis/revision_v2/` retains the public-data selection audit and geometry. `analysis/human_final1000_20260925/` contains the final-file protocol, exact IDs, hashes, predictions, thresholds, and paired intervals. Model-run and cleanup records remain in `analysis/human_holdout_20260925/`. Table 11 identifies historical variants. The typesetting archive excludes private annotations and source-person mappings.

G.2 Checkpoint handling

Training checkpoints and non-selected weights were removed after saving predictions, metrics, thresholds, and logs. A separately trained seed-42 deployment artifact is retained. It is excluded from all three-seed averages: its test Macro-F1 is .5142 and probe Macro-F1 .3219. Its model file is 1,182,779,359 bytes with SHA-256

```
6aac6ea79603918418e42e12c7349c8d04d2e0eaf2811f2ed4da8e65ece070fb
```

The artifact card records configuration, taxonomy checksum, inference, and smoke test. Human evaluation uses this checkpoint unchanged. The ASL rerun deletes its temporary best-validation state after exporting predictions, thresholds, and metrics. New runs remain separate from historical three-seed means.

G.3 Validity and remaining evidence gaps

Question framing, quotation, irony, social desirability, and strategic speech complicate attribution. Answer evidence does not establish sincerity or stable preference. Expression remains distinct from endorsement; Universalism: objectivity remains outside the classical ring.

Shared-family annotation agreement and exact-span checks do not establish semantic correctness. A retrospective audit of the original 100k labels should stratify by source, rare label, polarity, and question-only cues. The separate human evaluation uses A/B annotation and third-person adjudication, but concerns different dialogues and does not validate the original machine labels. Its final file alone cannot recover pre-adjudication agreement.

The original split supported architecture development. The fixed-model human analysis uses previously unused MediaSum conversations, approximate overlap screens, and a retrospectively defined positive-only population. It leaves no-value rejection untested. One run per system cannot establish seed robustness; intervals omit annotation uncertainty. Most historical component removals also have one seed, and several baselines are task adaptations.