Beyond Label Collapse: CRADLE for Human-Validated NaN-Aware Annotation in Low-Resource Bengali NLP

Anonymous ACL submission

Abstract

We present **CRADLE** (Confidence Routing Anchor-Driven Labeling Engine), a six-stage pipeline addressing instruction-induced label collapse — a systematic tendency wherein LLMs suppress minority-class labels regardless of instruction design. Despite explicit NaN instructions, three state-of-the-art LLMs achieve unanimous NaN assignment in **0.0%** of cases, suggesting collapse reflects inductive biases rather than instruction ambiguity. CRADLE combines confidence-based routing, Claude-anchored consensus judging, human validation, and RIPA (Reward-Informed Prompt Adaptation) — a reward-guided few-shot adaptation method using Qwen2.5 7B at zero API cost. Applied to MultiSoc-4D, a 58,240-comment Bengali social media corpus, CRADLE produces **MultiSoc-4D v2**: Hateful Yes rises from 4.6% to 19.7%, Sentiment Neutral falls from 65.4% to 9.1%. Macro-F1 improves with mean gains of +0.24, +0.32, +0.21, and +0.02 across 11 models and three architecture families. MultiSoc-4D v2 is released as a diagnostic and training benchmark for low-resource Bengali NLP.

1 Introduction

The use of Large Language Models (LLMs) as efficient substitutes for human annotation has become common in NLP tasks (Gilardi et al., 2023; Tan et al., 2024). However, such tools may introduce systematic biases in low-resource settings, where annotation failures interact with limited linguistic coverage. Despite over 230 million native speakers, Bengali (Bangla) remains underrepresented in NLP research (Hasan et al., 2020).

Recent work has surfaced a concrete failure mode in this context. Pramanik et al. (2026) introduced MultiSoc-4D, a 58,240-comment Bengali social media dataset annotated across four dimensions — content category, sentiment, hatefulness, and sarcasm — using four LLMs under identical closed-set instructions. A key finding was *instruction-induced label collapse* (ILC): LLMs systematically over-assigned conservative fallback labels (*Other, Neutral, No*) to ambiguous cases, with only 21% of hateful instances detected and sarcasm agreement reaching Fleiss' $\kappa = -0.001$. No corrective pipeline was proposed.

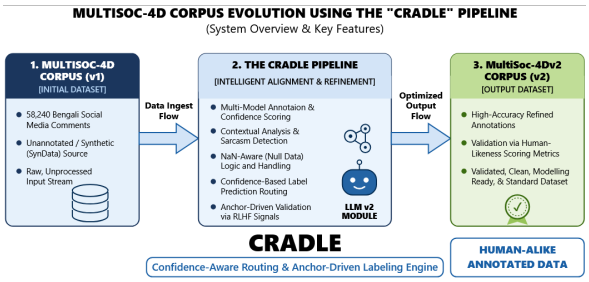

Figure 1: Overview of the CRADLE pipeline: from collapsed v1 annotations to the human-validated, NaN-aware MultiSoc-4D v2.

Dimension-level analysis (Section 3.4) reveals that the majority of comments carry no evaluative orientation for Sentiment or identifiable hate target — yet models assign NaN in **0.0%** of cases across all dimensions under open-set prompting, consistent with inductive biases toward closed-set label assignment acquired during pretraining rather than a failure of instruction design.

We present **CRADLE** (Confidence Routing Anchor-Driven Labeling Engine), a six-stage annotation pipeline that addresses this failure at the pipeline level rather than the prompt level, producing **MultiSoc-4D v2** (Figure 1). The pipeline integrates four components into a unified correction framework: GPT-4o, Gemini 1.5 Pro, and Grok-1 annotate each sample under open-set instructions permitting NaN responses; Claude — selected for its consistently higher model-to-human distributional alignment across all four dimensions in Pramanik et al. (2026) — acts as an anchor

judge, producing a consensus label with calibrated confidence $c \in [0, 1]$; samples below $\tau = 0.80$ are routed to human reviewers, whose structured correction signals drive **RIPA** (Reward-Informed Prompt Adaptation), a reward-guided few-shot adaptation method that uses corrected annotations to guide Qwen2.5 7B (Qwen Team, 2025) — running locally on 4-bit BitsAndBytes quantization — to re-annotate low-confidence samples at zero API cost. All RIPA-corrected labels undergo a final human revision before release. Central to CRADLE is the **NaN class**: a label representing the semantic inapplicability of a dimension, distinct from uncertainty, preventing the measurement bias introduced when semantically absent instances are forced into *Neutral* or *No*.

The primary novelty of this work lies in the pipeline-level integration of these components. This work makes the following contributions: **(i) MultiSoc-4D v2**: a human-validated, NaN-aware Bengali social media benchmark across four annotation dimensions; **(ii) NaN class formalization**: a principled label for semantic non-existence, distinct from uncertainty or ambiguity; **(iii) CRADLE**: a six-stage annotation pipeline combining disagreement taxonomy, confidence routing, human validation, and reward-guided local re-annotation — the primary contribution of this work; **(iv) RIPA**: a reward-guided few-shot adaptation method for annotation correction at zero API cost; and **(v)** a downstream benchmark of 11 models across three architecture families, showing consistent Macro-F1 gains under v2 supervision.

2 Related Work

2.1 LLM-based Annotation

Use of LLMs as proxy annotators has gained prominence after models like GPT were found to generate accurate labels on classification tasks compared to crowd-sourced annotators (Gilardi et al., 2023). Tan et al. (2024) review this literature, finding that closed-set prompts can yield accurate labels for sentiment analysis, toxic language detection, and stance classification in high-resource languages. He et al. (2024) introduce AnnoLLM, a two-phase framework improving annotation consistency via label explanation before assignment. Choi et al. (2024) show that LLM-generated synthetic data can boost generalization in sentiment classification. A key observation shared by all these works is their near-exclusive focus on English; LLM annotation behavior in low-resource multilingual settings remains understudied. Pramanik et al. (2026) identify a systematic failure mode in this regard — Instruction-Induced Label Collapse (ILC), wherein LLMs assign majority fallback labels under closed-set prompting, leading to severe under-detection of minority classes. This work extends that finding by providing a pipeline-level solution via CRADLE.

2.2 Bengali NLP Datasets

Progress in Bengali NLP datasets has been steady. Kabir et al. (2024) present BenLLM-Eval, finding that fine-tuned monolingual models outperform LLMs under zero-shot conditions on Bengali classification, NLI, QA, and summarization. Sami et al. (2026) reveal significant performance drops across nine regional Bengali dialects. For hate speech, Haider et al. (2025) contribute BanTH, a multi-label transliterated dataset, and Raquib et al. (2025) contribute BANHATE with fine-grained hate types and LLM evaluations. For sentiment, Das et al. (2023) provide SentiGOLD, a gold-standard multi-domain corpus. Critically, no existing Bengali social media dataset combines open-set labeling, multi-dimensional annotation, and human-validated bias diagnosis at the scale of Pramanik et al. (2026) — a gap that MultiSoc-4D v2 directly addresses.

2.3 Human-in-the-Loop and Reward-Guided Annotation

Human-in-the-loop (HITL) annotation has been widely studied as a strategy for improving label quality through targeted human intervention (Mosqueira-Rey et al., 2023). Settles (2009) establish that uncertainty-based sampling concentrates human effort where information gain is highest — the principle underlying CRADLE's confidence routing stage. Reinforcement learning from human feedback (Ouyang et al., 2022) and its variants, including RLAIF (Lee et al., 2024) and MA-RLHF (Chai et al., 2025), demonstrate that structured human reward signals can effectively guide model behavior. However, as Mosqueira-Rey et al. (2023) note, most HITL systems route samples based on uncertainty but do not use the resulting human decisions to adapt annotator behavior prospectively. RIPA, introduced in this work, addresses this gap: human correction signals are used directly to adapt annotation prompts for local re-annotation, without requiring a full RL training loop.

2.4 Annotation Bias and Label Noise

Label noise is widely recognized as a generalization barrier in NLP (Frénay and Verleysen, 2014; Northcutt et al., 2021). Swayamdipta et al. (2020) propose dataset cartography, mapping instances by training dynamics into hard, easy, and ambiguous classes for targeted human validation. In the Bengali context, Elahi et al. (2024) apply human verification to automatically annotated data to improve label reliability. Choi et al. (2024) address label discrepancy through human re-annotation of high-conflict instances in multilingual summarization. Both approaches treat annotation bias as a post-hoc issue. MultiSoc-4D v2 instead intervenes during annotation via confidence routing and reward-guided prompt adaptation. No prior work addresses label collapse as a pipeline-level failure in low-resource annotation — the gap that CRADLE explicitly targets.

3 Dataset

3.1 MultiSoc-4D v1: Background

Pramanik et al. (2026) introduced MultiSoc-4D, a Bengali social media dataset annotated across four dimensions — content category, sentiment, hatefulness, and sarcasm — collected from six platforms (Facebook, Twitter/X, YouTube, TikTok, Likee, Instagram) using closed-set LLM prompts. The central finding was ***instruction-induced label collapse*** (ILC): LLMs over-represented majority fallback classes (Hateful Yes: 4.6%, Sarcasm Yes: 3.6%, Fleiss $\kappa = -0.001$ on sarcasm) while severely under-detecting minority classes, with no corrective pipeline proposed.

3.2 MultiSoc-4D v2: Overview

v2 reannotates the same 58,240 comments under three key changes: **(i)** open-set instructions permitting NaN per dimension; **(ii)** confidence-based routing at $\tau = 0.80$, requiring human validation for uncertain labels; and **(iii)** RIPA correction for failed validations (Section 4.5). Since semantic relevance varies per dimension, v2 is released as four separate datasets after NaN removal: Category (56,380), Sentiment (21,145), Hateful (21,517), and Sarcasm (28,036).

3.3 The NaN Class

NaN is a positive assertion of semantic inapplicability — distinct from uncertainty (low confidence) and ambiguity (multi-label cases). Closed-set annotation conflates a property's *absence* with its *neutral manifestation*: a non-evaluative comment is forced into *Neutral*, inflating its count and degrading classifier calibration. NaN resolves this across all four dimensions: **Category** (phatic or incoherent text; platform artifacts); **Sentiment** (no affective orientation, distinct from *Neutral* which is balanced); **Hatefulness** (no identifiable target or evaluatable statement); **Sarcasm** (insufficient propositional structure for irony judgment). Table 1 shows representative v1 vs. v2 label changes across dimensions.

3.4 Dataset Statistics

Table 2 reports NaN rates per dimension. Sentiment and Hateful exceed 60% NaN, indicating that the majority of Bengali social media comments carry no evaluative orientation or identifiable hate target — a corpus property entirely invisible under v1's closed-set scheme. Category preserves ≈96.8% of comments as valid, while 1,933 comments received NaN across all four dimensions simultaneously.

Table 3 shows label distribution shifts from v1 to v2. The NaN class substantially corrects label collapse: Category Other drops from 51.8% to 35.8%; Sentiment Neutral falls from 65.4% to 9.1% as evaluatively absent comments are correctly abstained; and Hateful Yes rises from 4.6% to 19.7%. Sarcasm Yes shows minimal change (3.6%→1.3%), with only 55 confirmed cases across 28,036 valid samples — reflecting the genuine rarity and context-dependence of Bengali sarcasm in single-comment settings, discussed further in Section 6.4. Full annotation pathway breakdown is provided in Appendix A.1.

4 Methodology

CRADLE (Confidence Routing Anchor-Driven Labeling Engine) is a six-stage annotation pipeline designed to detect, route, and correct instruction-induced label collapse without model fine-tuning or full human re-annotation. Figure 2 provides the system architecture overview.

4.1 Stage 1: Multi-Model Open-Set Annotation

GPT-4o, Gemini 1.5 Pro, and Grok-1 — identified by Pramanik et al. (2026) as the most collapse-prone models under closed-set prompting — independently annotate all 58,240 comments under

Table 1: Label changes from v1 (closed-set) to v2 (NaN-aware). Annotation performed on original Bengali text; English glosses provided for readability.

Comment (gloss)	Dim.	v1	v2	Reason
"আজ সকাল ৯টায় কলেজে খেলা" *"College game at 9 AM today"*	Sentiment	Neutral	NaN	No evaluative orientation; purely temporal
"অনুষ্ঠানটি সকাল ৬টায় শুরু হবে" *"The programme starts at 6 AM"*	Hateful	No	NaN	No identifiable target; dimension inapplicable
"ওহ বাহ, ইন্টারনেট আবার ঠিক সময়েই গেল" *"Oh great, internet went out right on time"*	Sarcasm	No	Yes	Ironic surface contradicts expected outcome

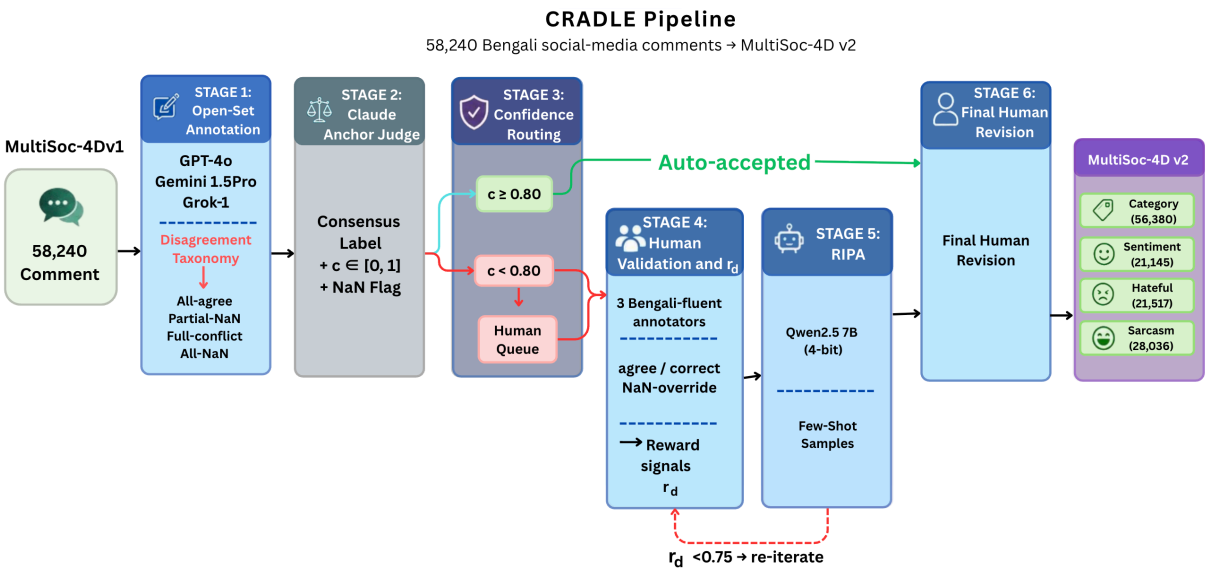

Figure 2: CRADLE system architecture: six-stage pipeline from MultiSoc-4D v1 to the human-validated, NaN-aware MultiSoc-4D v2.

Table 2: NaN rates and valid sample counts per dimension in MultiSoc-4D v2.

Dimension	NaN count	NaN %	Valid (n)
Category	1,860	3.19%	56,380
Sentiment	37,095	63.69%	21,145
Hateful	36,723	63.07%	21,517
Sarcasm	30,204	51.86%	28,036

open-set instructions permitting NaN. Each model receives an identical three-part prompt: **(i)** dimension definition with decision criteria; **(ii)** NaN definition with a worked negative example; and **(iii)** output specification requiring a label and brief rationale (Appendix B).

Despite explicit NaN instructions, All-NaN = 0.0% across all dimensions (Table 4). On Sentiment, 86.5% show Partial-NaN disagreement — GPT-4o assigned NaN while Gemini and Grok defaulted to closed-set labels. On Sarcasm, 99.9% All-agree reflects collapse onto *No* rather than genuine consensus. This pattern is consistent with inductive biases toward closed-set label assignment acquired during pretraining, suggesting the failure is not addressable through instruction design alone.

Stage 1 outputs are classified into four disagreement types (Table 4): **All-agree** (all three assign the same non-NaN label); **Partial-NaN** (one or two assign NaN, rest assign closed-set); **Full-conflict** (all three differ, no NaN — hard-routed to human validation); and **All-NaN** (0.0% across all dimensions).

4.2 Stage 2: Claude as Anchor Judge

Claude — selected for its consistently higher model-to-human distributional alignment ratios across all four dimensions reported by Pramanik et al. (2026) (e.g., M2H ratios for International: 0.96, Economy: 1.06, Sports: 1.08, Sarcasm: 0.94, all closest to the human baseline of 1.00) — receives the three annotator labels, rationales, and

Table 3: Label distribution comparison: v1 (closed-set) vs. v2 (NaN-aware, valid instances only).

Dimension	Class	v1 %	v2 %
Category	National	13.3	24.6
	International	5.6	6.6
	Education	2.8	3.4
	Economy	6.2	7.9
	Entertainment	14.0	15.6
	Sports	4.2	3.4
	Technology	2.1	2.7
	Other	51.8	35.8
Sentiment	Positive	18.8	42.3
	Negative	15.8	48.6
	Neutral	65.4	9.1
Hateful	Yes	4.6	19.7
	No	95.4	80.3
Sarcasm	Yes	3.6	1.3
	No	96.4	98.7

Table 4: Stage 1 disagreement type distribution (58,240 instances). All-NaN = 0.0% across all dimensions is consistent with label collapse as an inductive bias rather than an instruction failure.

Dim.	All-agree	Partial-NaN	Full-conflict	All-NaN
Category	80.6%	0.0%	19.4%	0.0%
Sentiment	8.5%	86.5%	5.1%	0.0%
Hateful	90.4%	0.0%	9.6%	0.0%
Sarcasm	99.9%	0.0%	0.1%	0.0%

Stage 1 disagreement type, producing: **(i)** a consensus label; **(ii)** a calibrated confidence score $c \in [0, 1]$; and **(iii)** a binary NaN flag that overrides the consensus label and routes the instance to human verification regardless of c. We note that Claude's alignment ratios were measured under closed-set annotation in Pramanik et al. (2026); their transferability to open-set judging is an assumption of this design, mitigated by the human validation stage that audits all low-confidence Stage 2 outputs.

$$c = \text{clip}\big(\alpha c_{\text{base}} + (1-\alpha)\mathbb{1}_{\text{ag}} - \beta \mathbb{1}_{\text{pNaN}},\ 0, 1\big) \quad (1)$$

where c_{base} is Claude's intrinsic confidence, $\mathbb{1}_{\text{ag}}$ and $\mathbb{1}_{\text{pNaN}}$ are indicators for All-agree and Partial-NaN respectively, $\alpha = 0.7$ weights annotator consensus, and $\beta = 0.15$ penalizes NaN suppression. Parameters $\alpha = 0.7$ and $\beta = 0.15$ were set to prioritize annotator consensus over model-intrinsic confidence while applying a moderate penalty for NaN suppression; confidence score distributions are reported in Appendix C.2. Parameter sensitivity is left to future work.

Table 5: Confidence-based routing outcomes at $\tau = 0.80$.

Dim.	Mean c	Human queue %	Human queue n
Category	0.803	40.3%	24,219
Sentiment	0.735	71.5%	42,938
Hateful	0.897	6.6%	3,987
Sarcasm	0.841	0.1%	34

4.3 Stage 3: Confidence-Based Routing

Each instance is partitioned into an *auto-accepted* queue ($c \geq \tau = 0.80$) or a *human queue* ($c < \tau$, NaN flag set, or Full-conflict). Category routes the largest share (40.3%), reflecting its lowest mean confidence and highest Full-conflict rate. Sarcasm routes only 0.1% — reflecting All-agree collapse rather than genuine certainty. Table 5 reports full routing outcomes; threshold selection is justified in Section 4.7.

4.4 Stage 4: Human Validation and Reward Collection

Three Bengali-speaking annotators independently audit 10% of valid instances per dimension (20% for Sarcasm, given its near-random annotation quality ($\kappa = -0.001$) in v1), yielding 5,638 (Category), 2,114 (Sentiment), 2,151 (Hateful), and 5,607 (Sarcasm) human-reviewed instances. Each annotator chooses one of three actions: *agree*, *correct*, or *NaN-override* (Figure 3).

Structured reward signals are collected per decision: $+1.0$ for agreeing with consensus; $+0.3$ for correcting a low-confidence label ($c < \tau$); 0.0 for correcting a high-confidence label ($c \geq \tau$, signaling calibration failure); and -0.2 if annotators disagree, preventing contradictory examples from entering the RIPA pool:

$$\bar{r}_d = \frac{1}{B}\sum_{i=1}^{B} r_{d,i}, \quad r_{d,i} \in \{+1.0,\ +0.3,\ 0.0,\ -0.2\} \quad (2)$$

The asymmetry between $+0.3$ and 0.0 is deliberate: correcting a low-confidence label confirms the routing mechanism is working, while correcting a high-confidence label signals a calibration failure that should not be rewarded as positive annotation signal. If $\bar{r}_d < 0.75$, RIPA is triggered.

Figure 3: Human validation and reward collection loop (Stage 4).

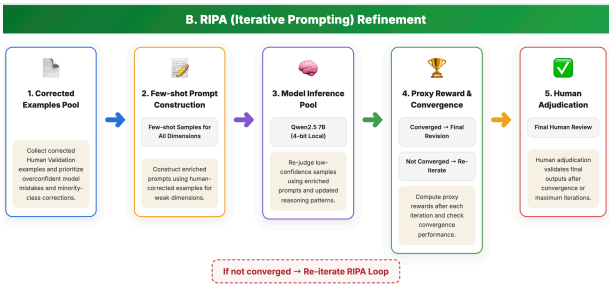

Figure 4: RIPA: reward-guided few-shot adaptation loop (Stage 5).

4.5 Stage 5: RIPA — Reward-Informed Prompt Adaptation

Human corrections are accumulated to adapt annotation prompts for local re-annotation without API cost (Figure 4). Corrected instances are extracted as triples $\langle x, \hat{y}_{\text{model}}, y^*_{\text{human}} \rangle$ and injected as negative-positive few-shot pairs — making the failure mode explicit rather than expanding positive examples alone. The enriched prompt is forwarded to Qwen2.5 7B Instruct (Qwen Team, 2025) via 4-bit quantization using *BitsAndBytes*(Dettmers et al., 2022) at zero per-token API cost. RIPA re-annotated 2,073 instances: Category (1,360), Sarcasm (1,110), Hateful (805), and Sentiment (781).

We distinguish RIPA from standard iterative prompt engineering: prompt adaptation is directly conditioned on structured reward signals from the human validation loop, and convergence is assessed objectively via a proxy metric rather than by manual inspection. RIPA does not implement a full RL policy update; its connection to reinforcement learning is analogical, and extended discussion of this distinction is provided in Appendix C.1.

Convergence is assessed via proxy reward:

$$\tilde{r}_{d,i} = \frac{1}{|N_i|} \sum_{j \in N_i} \mathbb{1}[y^{\text{RIPA}}_{d,i} = y^*_{d,j}] \qquad (3)$$

where N_i represents the nearest-neighbour human-validated instances in embedding space. Convergence is declared when $\tilde{r}_d > 0.75$ for two consecutive batches; non-converging instances are escalated to final human adjudication.

4.6 Stage 6: Final Human Revision

All RIPA-corrected labels undergo a final human revision pass before NaN removal and dataset release, ensuring that no machine-generated label enters the final dataset without human validation.

4.7 Ablation: Confidence Threshold τ

At $\tau = 0.70$, Sarcasm routes zero instances, providing no intervention on the hardest dimension. At $\tau = 0.90$, Sentiment routes 100% of instances, making the pipeline logistically infeasible; $\tau = 0.80$ concentrates human effort where genuine uncertainty exists and is adopted as the operating threshold.

5 Experimental Results

5.1 Experimental Setup

We train 11 models on v1 and v2 labels separately, excluding NaN instances from both training and evaluation to ensure a semantically coherent training distribution. Models span three families: **instruction-tuned LLMs** (Qwen2.5 0.5B, LLaMA3.2 3B, Mistral 7B); **multilingual transformers** (XLM-RoBERTa, RemBERT, MuRIL, IndicBERT); and **monolingual transformers** (BanglaBERT, FinBERT, Electra, DeBERTa). Fine-tuned models use an 80/10/10 train/validation/test split. Primary metric is Macro-F1, treating minority-class performance equally to majority classes. Accuracy results are in Appendix F.1.

5.2 Annotation Quality: v1 vs. v2

Table 6 reports inter-annotator agreement before and after CRADLE. Fleiss' κ improves across all dimensions: from -0.001 to 0.158 on Sarcasm (+0.159), and from 0.39 to 0.70 on Hateful (+0.31) — the two dimensions most severely affected by label collapse in v1. The All-NaN = 0.0% finding is consistent with these gains being driven by pipeline-level intervention rather than instruction

Table 6: Inter-annotator agreement (Fleiss' κ) before (v1) and after (v2) CRADLE annotation.

Dimension	v1 κ	v2 κ	$\Delta\kappa$	Interpretation
Category	0.41	0.64	+0.23	Moderate → Substantial
Sentiment	0.55	0.78	+0.23	Moderate → Substantial
Hateful	0.39	0.70	+0.31	Fair → Substantial
Sarcasm	−0.001	0.158	+0.159	None → Slight

design, as instruction revision alone would not alter the NaN suppression pattern.

5.3 Downstream Benchmarking

Table 7 reports Macro-F1 for all 11 models across four dimensions under v1 and v2 supervision. V2 labels yield mean Macro-F1 gains of +0.24 (Category), +0.32 (Sentiment), +0.21 (Hateful), and +0.02 (Sarcasm). These gains are consistent across model types and architecture families, indicating the improvement is attributable to label quality rather than model-specific properties. The largest absolute gains appear on Hateful and Sarcasm — precisely the dimensions most suppressed by label collapse in v1 — consistent with CRADLE's targeted minority-class correction.

Monolingual transformers (BanglaBERT, FinBERT) show the largest gains overall, consistent with domain-appropriate Bengali pretraining enabling more effective exploitation of recovered minority-class signal. Instruction-tuned LLMs show the smallest gains, consistent with their lower sensitivity to training label quality in few-shot settings. The v2 improvement persists despite NaN exclusion from training: the NaN filter removes semantically inapplicable instances that previously contributed noise, yielding cleaner decision boundaries. This suggests CRADLE's value lies not only in recovering minority-class instances but also in purifying the majority-class training signal through principled abstention.

5.4 Cross-Validation: Lingual and Dataset Generalisability

To assess preliminary generalisability beyond Bengali, we conduct two validation experiments whose full results appear in Appendix E.1 and Appendix E.2.

Cross-lingual validation. The full six-stage pipeline is applied to 2,030 English-dominant social media comments without language-specific recalibration. NaN rates and auto-acceptance rates are broadly consistent with the Bengali corpus (Table 16), and RIPA triggers on all four dimensions, suggesting the NaN class may capture a language-independent property of social media text.

Cross-dataset validation. CRADLE is applied to stratified samples from two existing Bengali benchmarks — SentNoB ($n = 1{,}000$, sentiment) and BanTH ($n = 1{,}000$, hatefulness) — without corpus-specific recalibration. RIPA triggers on both corpora, and non-trivial NaN rates (6.6% and 4.2% respectively) are observed even in these task-curated datasets. Critically, All-NaN = 0.0% holds across all external corpora, confirming that NaN suppression is a property of the annotator LLMs rather than an artifact of MultiSoc-4D specifically.

We treat both experiments as directional evidence only; a full multi-language benchmark and systematic cross-dataset evaluation are needed to establish generalisability conclusively.

6 Discussion

6.1 NaN as a Diagnostic Tool

The failure of explicit NaN instructions (Section 4.1) is further illuminated by what the NaN class reveals about v1's semantic distortions. On Sentiment, 63.7% of the corpus receives NaN in v2 — invisible under v1's closed-set scheme, where these instances inflated *Neutral* from its true 9.1% to 65.4%. Models trained on v1 therefore learned an artificially expanded *Neutral* decision region, conflating balanced affect with its complete absence. More broadly, high NaN concentration in any closed-set label class suggests that class is doing double duty — a transferable diagnostic signal for any closed-set annotated dataset.

6.2 Negative Δ Cases

These negative cases must be read against the v1 accuracy baseline, where always predicting *No* yields 95.4% accuracy on Hateful — a misleading signal that v2 corrects by raising Hateful Yes from 4.6% to 19.7% and providing a genuine minority-class training signal. Several models show no improvement or small regressions under v2 supervision: Qwen2.5 0.5B on Category (-0.19), XLM-RoBERTa and Electra on Hateful (-0.04), Electra on Sentiment (-0.03), and FinBERT on Sarcasm (-0.09). We attribute these to two factors: **(i)** very small or capacity-limited models (Qwen2.5 0.5B, Electra) that may struggle to exploit a larger and more balanced label space; and **(ii)** models with weak Bengali representations (XLM-RoBERTa, Electra) that are confounded by

Table 7: Downstream Macro-F1: v1 vs. v2 labels. Δ = absolute improvement. Best v2 per dimension in **bold**. Negative Δ cases are discussed in Section 6.

		Category			Sentiment			Hateful			Sarcasm		
Family	**Model**	**v1**	**v2**	**Δ**	**v1**	**v2**	**Δ**	**v1**	**v2**	**Δ**	**v1**	**v2**	**Δ**
LLM	Qwen2.5 0.5B	0.73	0.54	-0.19	0.25	0.54	+0.29	0.32	0.78	+0.46	0.32	0.49	+0.17
	LLaMA3.2 3B	0.70	0.92	+0.22	0.65	0.95	+0.30	0.69	0.90	+0.21	0.50	0.50	0.00
	Mistral 7B	0.70	0.88	+0.18	0.65	0.96	+0.31	0.70	0.88	+0.18	0.54	0.50	-0.04
Multi.	XLM-RoBERTa	0.09	0.67	+0.58	0.26	0.96	+0.70	0.49	0.45	-0.04	0.49	0.50	+0.01
	RemBERT	0.73	0.94	+0.21	0.26	0.97	+0.71	0.49	0.91	+0.42	0.49	0.50	+0.01
	MuRIL	0.45	0.92	+0.47	0.64	0.91	+0.27	0.71	0.91	+0.20	0.49	0.50	+0.01
	IndicBERT	0.73	0.93	+0.20	0.70	0.93	+0.23	0.74	0.89	+0.15	0.58	0.65	+0.07
Mono.	BanglaBERT	0.71	0.90	+0.19	0.68	0.92	+0.24	0.77	0.90	+0.13	0.59	0.65	+0.06
	FinBERT	0.48	0.75	+0.27	0.62	0.89	+0.27	0.54	0.91	+0.37	0.59	0.50	-0.09
	Electra	0.09	0.24	+0.15	0.26	0.23	-0.03	0.49	0.45	-0.04	0.49	0.50	+0.01
	DeBERTa	0.51	0.84	+0.33	0.66	0.92	+0.26	0.68	0.90	+0.22	0.49	0.49	0.00
Mean Δ				+0.24			+0.32			+0.21			+0.02

the altered class distribution after NaN removal. These cases indicate that CRADLE's benefit is not universal and depends on sufficient model capacity and language coverage.

6.3 RIPA as a Practical Solution

RIPA re-annotates 3.5% of valid samples (2,073 instances) at zero API cost on a single consumer GPU, concentrating human validation on 10% of instances per dimension (20% for Sarcasm). This makes CRADLE accessible in low-resource settings where full human re-annotation is impractical, with cross-validation results (Section 5.4) providing directional evidence of transfer without language-specific recalibration.

6.4 The Sarcasm Problem

Sarcasm remains the hardest dimension: only 55 confirmed Yes instances (0.2% of 28,036 valid samples) survive full pipeline processing, with downstream models producing their lowest Macro-F1 on this dimension regardless of label version. This is best understood as a corpus property rather than a pipeline failure — Bengali sarcasm is infrequent in short single-comment contexts, heavily reliant on community-specific references unavailable to single-comment annotators, and frequently expressed through code-mixed register switching. RIPA triggers ($\bar{r}_{\text{sarcasm}} < 0.75$) and escalates the largest count of non-converging instances to human adjudication, reflecting genuine annotation difficulty. Dedicated strategies such as counterfactual augmentation or thread-level annotation are required before sarcasm models trained on MultiSoc-4D v2 can be considered reliable.

7 Conclusion

Instruction-induced label collapse is not resolved by better instructions. Even under explicit open-set prompting with NaN definitions, decision criteria, and worked examples, state-of-the-art LLMs show All-NaN = 0.0% across all four annotation dimensions — consistent with inductive biases toward closed-set label assignment that instruction design alone cannot override. We present CRADLE (Confidence Routing Anchor-Driven Labeling Engine), a six-stage annotation pipeline that addresses this failure through disagreement-based routing, human validation with structured reward collection, and RIPA-guided local re-annotation at zero API cost. Applied to MultiSoc-4D, CRADLE produces MultiSoc-4D v2: Hateful Yes rises from 4.6% to 19.7%, Sentiment Neutral falls from 65.4% to 9.1%, and Category Other falls from 51.8% to 35.8%. Downstream Macro-F1 improves with mean gains of +0.24, +0.32, +0.21, and +0.02 across 11 models and three architecture families under v2 supervision, with negative Δ cases on specific model-dimension pairs discussed honestly in Section 6. These results suggest that pipeline-level intervention, rather than prompt design, is a more effective corrective for annotation bias in low-resource settings. MultiSoc-4D v2 is released as both a diagnostic resource and a training benchmark for low-resource Bengali NLP.

Limitations

Residual annotation error. MultiSoc-4D v2 retains a non-trivial residual misclassification rate on the Sarcasm and Hateful dimensions, reflecting the convergence ceiling of the proxy reward mechanism. Instances that do not converge within the RIPA iteration budget are escalated to human adjudication, but adjudication does not guarantee correctness in every case. Researchers using v2 labels for downstream training should be aware that the dataset, while substantially cleaner than v1, is not error-free.

Sarcasm Yes underrepresentation. With only 55 confirmed Sarcasm Yes instances (0.2% of valid sarcasm samples), v2 does not provide a sufficient training signal for reliable sarcasm classification. This is best understood as a corpus property — Bengali sarcasm is infrequent in short single-comment contexts, heavily reliant on community-specific references unavailable to single-comment annotators, and frequently expressed through code-mixed register switching — rather than a pipeline failure. Dedicated augmentation strategies are required before sarcasm models trained on MultiSoc-4D v2 can be considered reliable.

Claude judge reproducibility. Stage 2 uses Claude without API version pinning, meaning exact replication of Stage 2 judgments is not guaranteed if Claude's behavior changes across versions. We mitigate this by releasing the full Stage 2 outputs (label, confidence score, NaN flag, rationale) for every instance, so downstream pipeline stages can be reproduced independently. Replacing the Claude judge with a fine-tuned open-weight model is the highest-priority direction for future work.

Proxy reward approximation. The RIPA convergence criterion uses a proxy reward computed by nearest-neighbour matching in embedding space, which approximates but does not equal true human reward. This approximation may overestimate true reward when sentence embedding models do not adequately separate instances with different correct labels — a known failure mode on code-mixed or dialectally diverse text. Researchers applying RIPA to other corpora should validate the proxy reward against a held-out human sample before relying on it as a convergence criterion.

Cross-lingual and cross-dataset generalisability. The 2,030-sample English validation set establishes directional consistency but is insufficient for strong cross-lingual transfer claims. Directional cross-dataset validation on SentNoB and BanTH (Appendix E.2) shows non-trivial NaN rates and RIPA triggering on both corpora, but full re-annotation and downstream benchmarking are needed before strong generalisability claims can be made.

Future directions. Five directions follow directly from these limitations: **(i)** replacing the Claude anchor judge with a fine-tuned open-weight model distilled on Stage 2 outputs; **(ii)** dedicated sarcasm augmentation via counterfactual generation or cross-platform transfer; **(iii)** cross-dataset validation of CRADLE on existing Bengali benchmarks such as SentNoB and BanTH to assess generalisability of the NaN suppression finding; **(iv)** extending CRADLE to other low-resource South Asian languages (Hindi, Urdu, Tamil); and **(v)** replacing the proxy reward with a true multi-turn RLHF loop for improved convergence reliability on hard dimensions.

Ethical Considerations

Data privacy. MultiSoc-4D v2 is derived from publicly available social media text collected from Facebook, Twitter/X, YouTube, TikTok, Likee, and Instagram. All personally identifiable information (PII) has been removed prior to annotation, and the research adheres to standardized academic protocols for large-scale social media data analysis.

Annotator welfare. Human annotators in the RLHF validation loop were exposed to hateful and sarcastic content as part of the annotation process. Annotator sessions were time-limited, opt-in, and conducted with full awareness of content sensitivity. Annotators were compensated fairly and provided with the option to withdraw from the study at any point without penalty.

Dual-use risk. MultiSoc-4D v2 contains hate speech instances that, while essential for training robust content moderation systems, carry inherent dual-use risk. We release the dataset for research purposes only, with explicit guidance against use in systems designed to facilitate harm. Access is provided under a license prohibiting commercial exploitation of harmful content.

Model bias. The LLMs used in CRADLE — GPT-4o, Gemini 1.5 Pro, Grok-1, Claude, and Qwen2.5 7B — carry inherent biases from their pretraining data. By comparing LLM-generated labels against human gold standards and quantifying bias ratios explicitly, this work aims to highlight rather than replicate these biases, fostering more transparent automated annotation pipelines for low-resource languages.

Representational scope. MultiSoc-4D v2 represents Bengali social media discourse from six platforms during a specific collection window. Findings may not generalise to all Bengali dialects, registers, or time periods. Researchers should exercise caution when applying models trained on this dataset to out-of-distribution Bengali text.

References

Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and hua wu. 2025. Ma-rlhf: Reinforcement learning from human feedback with macro actions. In *International Conference on Learning Representations*, volume 2025, pages 26647–26680.

Juhwan Choi, JungMin Yun, Kyohoon Jin, and Young-Bin Kim. 2024. Multi-news+: Cost-efficient dataset cleansing via LLM-based data annotation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15–29, Miami, Florida, USA. Association for Computational Linguistics.

Dipankar Das, Md Saiful Islam, Md Arid Hasan, and Firoj Alam. 2023. SentiGOLD: A large Bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2023)*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35.

Kazi Toufique Elahi, Tasnuva Binte Rahman, Shakil Shahriar, Samir Sarker, Md. Tanvir Rouf Shawon, and G. M. Shahariar. 2024. A comparative analysis of noise reduction methods in sentiment analysis on noisy Bangla texts. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 1–14, San Ġiljan, Malta. Association for Computational Linguistics.

Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. volume 120.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. BanTH: A multi-label hate speech detection dataset for transliterated Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 878–889, Albuquerque, New Mexico. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin Samin, Masnun Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1673–1694, Mexico City, Mexico. Association for Computational Linguistics.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. BenLLM-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on Bengali NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252, Torino, Italia. ELRA and ICCL.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56:3005–3054.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder,

Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Souvik Pramanik, S. M. Riaz Rahman Antu, Shak Mohammad Abyad, Md. Ibrahim Khalil, and Md. Shahriar Hussain. 2026. Multisoc-4d: A benchmark for diagnosing instruction-induced label collapse in closed-set llm annotation of bengali social media. *Preprint*, arXiv:2605.06940.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Faisal Hossain Raquib, Akm Moshiur Rahman Mazumder, Md Tahmid Hasan Fuad, Md Farhan Ishmam, and Md Fahim. 2025. BanHate: An up-to-date and fine-grained Bangla hate speech dataset. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 1–12, Mumbai, India. Association for Computational Linguistics.

K. M. Jubair Sami, Dipto Sumit, Ariyan Hossain, and Farig Sadeque. 2026. Benchmarking bengali dialectal bias: A multi-stage framework integrating rag-based translation and human-augmented rlaif. *Preprint*, arXiv:2603.21359.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

A Dataset: Extended Statistics and Examples

A.1 Pre-Cleaning Label Distributions

Table 8 reports the complete label distribution of the 58,240-comment corpus before NaN filtering. This view exposes how much of each annotation dimension was judged semantically inapplicable during CRADLE processing.

Table 8: Full label distribution before NaN removal (58,240 source comments). NaN entries represent semantically inapplicable instances identified during open-set annotation.

Dimension	Class	Count	%
Category	National	13,455	23.1%
	International	3,626	6.2%
	Education	1,870	3.2%
	Economy	4,311	7.4%
	Entertainment	8,502	14.6%
	Sports	1,870	3.2%
	Technology	1,521	2.6%
	Other	19,614	33.7%
	NaN	3,471	**6.0%**
Sentiment	Positive	8,686	14.9%
	Negative	9,982	17.1%
	Neutral	1,870	3.2%
	NaN	37,702	**64.8%**
Hateful	Yes	4,116	7.1%
	No	16,422	28.2%
	NaN	37,702	**64.8%**
Sarcasm	Yes	55	0.1%
	No	27,981	48.0%
	NaN	30,204	**51.9%**

A.2 Boundary Case Taxonomy

In order to achieve high levels of inter-annotator agreement and to have reliable model evaluation, the semantic null classes (such as neutral sentiment or irrelevant text) are deliberately distinguished in v2 annotations from systemic non-applicability (NaN). These primary boundary conditions are presented in Table 9, establishing rigorous guidelines for differentiating semantic ambiguity from semantic nulls.

A.3 Extended NaN Label Examples

In order to give a real-world empirical basis to the taxonomy proposed in Appendix A.2, Table 10 offers an extended list of edge cases encountered in the course of the annotation review process. The dataset not only expands on the initial sample provided earlier in the paper (Table 1), but also serves to highlight the transition from rough and simplistic closed-set criteria of v1 to the more complex and subtle criteria of NaNs that were adopted in the v2 framework.

Specifically, the selection of corpus samples highlights key points of difficulty arising at the decision boundary of each of the semantic dimensions. Items marked with a dagger (†) in particular indicate those cases which were identified as very uncertain due to their borderline position between being classified as a NaN case and its closest closed-set counterpart (such as separating a viewpoint-neutral perspective from an absolutely non-evaluative assertion).

A.4 Representative Failure and Ambiguity Cases

The set of complex cases described in Table 11, which have defied clear classification even when using the flexible, open-set annotation strategy developed for v2, show that some texts present an insurmountable challenge to classification because of extreme semantic uncertainty, high contextual dependency, or pragmatically unclear meanings. This is not a matter of poor annotator performance, but of the difficulty of classification based solely on text due to linguistic features such as absent context, irony, or complex idioms.

B Annotation Guidelines and Prompt Templates

B.1 Annotation Philosophy

The v2 annotation protocol treats each comment as an open-set problem rather than a forced classification task. Both human and LLM annotators are explicitly permitted to assign NaN when a dimension is semantically inapplicable. The core operational principle is:

> *NaN is not a fallback for uncertainty. NaN indicates that a dimension genuinely does not apply to the text. If multiple labels appear plausible, select the closest applicable label rather than defaulting to NaN.*

The v1 heuristic of defaulting to `Other / Neutral / No / No` for ambiguous cases was explicitly removed. Annotation proceeds sequentially:

Category → Sentiment → Hateful → Sarcasm

Table 9: Boundary conditions emphasized in the v2 annotation protocol.

Boundary	**Applicable Label**	**NaN Condition**
Neutral vs. NaN	Sentiment exists but is balanced, mixed, or ambiguous	No sentiment-bearing language exists at all
Other vs. NaN	A coherent topic exists outside the predefined taxonomy	No identifiable topic exists whatsoever
No vs. NaN (Hateful)	Comment is interpretable and contains no hate speech	Comment is uninterpretable or non-linguistic
No vs. NaN (Sarcasm)	Literal interpretation is semantically possible	Insufficient semantic content for sarcasm judgment
Criticism vs. Hate	Targets actions, policies, or ideas	Targets identity groups, ethnicity, religion, or existence
Literal Praise vs. Sarcasm	Surface-positive statement reflects genuine intent	Surface-positive wording contradicts implied criticism

Table 10: Extended NaN label examples (v1 vs. v2). English glosses provided; annotation performed on original Bengali text.

Bengali Comment	**Dim.**	**v1**	**v2**	**Reason**
"আজ সকাল ৯টায় কলেজের খেলা" "College game at 9 AM today"	Sentiment	Neutral	NaN	No evaluative orientation; purely informational
"২৫ বৈশাখ উপলক্ষে বিশেষ অনুষ্ঠান" "Special programme on 25 Boishakh"	Category	Other	NaN	Event announcement without coherent topical category
"মিটিং শুরু হতে আর পাঁচ মিনিট" "Five minutes before the meeting"	Sarcasm	No	NaN	No propositional content; sarcasm inapplicable
"অনুষ্ঠান শুরু হবে সকাল ৬টায়" "The programme starts at 6 AM"	Hateful	No	NaN	No target entity or hateful intent present
"বাহ! ঠিক সময়েই আবার ইন্টারনেট গেল" "Oh great, internet went out right on time"	Sarcasm	No	Yes	Surface-positive expression conflicts with negative event outcome
"দুই পক্ষের কথাতেই যুক্তি আছে" "Both sides have valid points"	Sentiment	Neutral	Neutral†	Balanced evaluative stance; Neutral remains applicable
"এই নীতিটা শ্রমিকদের জন্য খারাপ" "This policy is terrible for workers"	Hateful	No	No†	Negative opinion without protected-group targeting
"আজকে আবহাওয়া অনেক গরম" "The weather is very hot today"	Hateful	No	NaN	Environmental observation; hatefulness inapplicable
"ওরা আবার উন্নয়নের গল্প শুরু করেছে" "They have started the development story again"	Sarcasm	No	Yes	Implicit ironic framing through dismissive repetition
"ফল প্রকাশ আগামীকাল দুপুরে" "Results will be published tomorrow afternoon"	Sentiment	Neutral	NaN	Pure factual statement without emotional orientation
"তুমি যা বলেছ সেটা ঠিক" "What you said is correct"	Sentiment	Positive	Positive†	Clear positive evaluative orientation; not NaN
"সবাইকে ঈদের শুভেচ্ছা" "Eid greetings to all"	Category	Other	NaN	Social greeting without meaningful topical assignment
"এই মানুষগুলো দেশটার ক্ষতি করছে" "These people are damaging the country"	Hateful	No	Yes	Group-directed hostile framing targeting a collective

Table 11: Representative failure and ambiguity cases observed during annotation.

Comment	Ambiguous Between	Final Label	Reason for Difficulty
দারুণ খেলেছে বাংলাদেশ, এরকম দল আর হয় না	Positive / Sarcasm	Positive	The phrase may express genuine praise or subtle mockery depending on contextual tone.
বাহ! ঠিক সময়েই আবার ইন্টারনেট গেল	Negative / Sarcasm	Sarcasm	Surface-positive wording conflicts with a negative real-world event; requires pragmatic interpretation.
এই মানুষগুলো দেশটার ক্ষতি করছে	Criticism / Hateful	Hateful	Hostile generalization toward a collective creates ambiguity between political criticism and identity-targeted hostility.
আজ সকাল ৯টায় কলেজের খেলা	Neutral / NaN	NaN	Contains topical information but no evaluative orientation; sentiment is absent rather than balanced.
ওরা আবার উন্নয়নের গল্প শুরু করেছে	Literal / Sarcasm	Sarcasm	Sarcastic interpretation depends entirely on implied pragmatic framing rather than explicit lexical cues.

B.2 Label Definitions

Category (9 labels). `International`, `National`, `Entertainment`, `Education`, `Sports`, `Technology`, `Economy`, `Other`, `NaN`. *Other*: a coherent topic exists but falls outside the predefined schema. *NaN*: no interpretable topic exists at all (greetings, hyperlinks, emoji-only, fragmented text).

Sentiment (4 labels). `Positive`, `Negative`, `Neutral`, `NaN`. *Neutral*: sentiment exists but is balanced or mixed. *NaN*: no sentiment-bearing language exists. Example: "আগামীকাল পরীক্ষা" ("Tomorrow is the exam") → NaN, not Neutral.

Hateful (3 labels). `Yes`, `No`, `NaN`. *Yes*: identity-targeted hate speech. *No*: interpretable comment with no hate speech. *NaN*: uninterpretable or non-linguistic content. Note: criticism of actions/policies → No; targeting identity, ethnicity, or religion → Yes.

Sarcasm (3 labels). `Yes`, `No`, `NaN`. *Yes*: intended meaning contradicts literal text. *No*: literal interpretation is semantically plausible. *NaN*: insufficient propositional structure for sarcasm judgment.

B.3 LLM Annotation Prompt Templates

An identical prompt template was used for GPT-4o, Gemini 1.5 Pro, and Grok-1 models for Stage 1. In order to maintain consistency in evaluation, the following three-point criteria apply to each of the four metrics: i) Definition of Dimension, Definition of Negative Case, Output Requirements.

Open-Set Annotation Prompt (Category)

Role: Act as a data annotator specializing in Bengali social media text classification.
Dimension: Category
Labels: International, National, Entertainment, Education, Sports, Technology, Economy, Other, NaN
Definitions:

- *International*: content about events or affairs outside Bangladesh
- *National*: content about Bangladeshi national affairs, politics, or governance
- *Entertainment*: content about films, music, celebrities, or cultural events
- *Education*: content about academic institutions, exams, or learning
- *Sports*: content about sporting events, teams, or athletes
- *Technology*: content about tech products, software, or digital services
- *Economy*: content about business, finance, markets, or trade
- *Other*: a coherent topic exists but falls outside all categories above
- *NaN*: no interpretable topic exists at all — assign only when genuinely inapplicable (e.g., greetings, hyperlinks, emoji-only, fragmented text)

NaN negative example: "বাংলাদেশ ম্যাচ জিতে গেছে" → *Sports*, NOT NaN — a coherent sporting topic is identifiable.
Boundary note: Assign *Other* when a topic exists but does not fit the schema. Assign *NaN* only when no topic exists whatsoever.
Output: Return only the label and a one-sentence rationale.
Comment: {comment}

Open-Set Annotation Prompt (Sentiment)

Role: Act as a data annotator specializing in Bengali social media text classification.
Dimension: Sentiment
Labels: Positive, Negative, Neutral, NaN
Definitions:

- *Positive*: comment expresses positive evaluative orientation
- *Negative*: comment expresses negative evaluative orientation
- *Neutral*: sentiment exists but is balanced or mixed
- *NaN*: no sentiment-bearing language exists — assign only when genuinely inapplicable

NaN negative example: "বাংলাদেশ ম্যাচ জিতে গেছে" → *Positive*, NOT NaN — clear positive affect is present.
Boundary note: *Neutral* means sentiment exists but is balanced or mixed. *NaN* means no sentiment-bearing language exists whatsoever. Example: "আগামীকাল পরীক্ষা" ("Tomorrow is the exam") → NaN, not Neutral.
Output: Return only the label and a one-sentence rationale.
Comment: {comment}

Open-Set Annotation Prompt (Hatefulness)

Role: Act as a data annotator specializing in Bengali social media text classification.
Dimension: Hatefulness
Labels: Yes, No, NaN
Definitions:

- *Yes*: comment contains direct or indirect hate speech targeting individuals or groups based on identity (ethnicity, religion, gender, nationality, or similar protected characteristics)
- *No*: comment is interpretable and contains no hate speech — including negative opinions, criticism of actions, or political disagreement
- *NaN*: comment is uninterpretable, non-linguistic, or contains no evaluatable statement on the hatefulness scale — assign only when the dimension is genuinely inapplicable

NaN negative example: "এই নীতিটা শ্রমিকদের জন্য খারাপ" ("This policy is bad for workers") → *No*, NOT NaN — the comment is interpretable and expresses political criticism without identity targeting.
Boundary note: Criticism of actions, policies, or ideas → *No*. Targeting identity groups, ethnicity, religion, or existence → *Yes*. Uninterpretable or non-linguistic content → *NaN*.
Output: Return only the label and a one-sentence rationale.
Comment: {comment}

Open-Set Annotation Prompt (Sarcasm)

Role: Act as a data annotator specializing in Bengali social media text classification.
Dimension: Sarcasm
Labels: Yes, No, NaN
Definitions:

- *Yes*: the intended meaning of the comment is the opposite of its literal text — the comment relies on irony, where surface-positive wording contradicts a negative implied meaning or vice versa
- *No*: the literal interpretation of the comment is semantically plausible and no ironic contradiction is present
- *NaN*: the comment lacks sufficient propositional structure for a sarcasm judgment to be made — assign only for single emojis, incoherent fragments, or text too short for pragmatic interpretation

NaN negative example: "বাংলাদেশ ম্যাচ জিতে গেছে" → *No*, NOT NaN — the comment has sufficient propositional structure; a sarcasm judgment can be made (and the literal interpretation is plausible).
Boundary note: Assign *Yes* only when the ironic contradiction is clear from the text itself. Do not infer sarcasm from context unavailable within the comment. If uncertain between *Yes* and *No*, prefer *No*. Assign *NaN* only when the comment lacks enough content for any pragmatic judgment.
Output: Return only the label and a one-sentence rationale.
Comment: {comment}

B.4 Representative Annotation Examples

Table 12 presents representative examples from the standardized v2 instruction sheet, including boundary cases where annotators were instructed to report reduced confidence rather than forcing overconfident judgments.

C Methodology: Supporting Details

C.1 RIPA and Reinforcement Learning: Conceptual Distinction

RIPA (Reward-Informed Prompt Adaptation) is distinguished from standard iterative prompt engineering in three ways: **(i)** prompt adaptation is directly conditioned on structured reward signals from the human validation loop rather than on manual inspection; **(ii)** convergence is assessed objectively via a proxy metric (Equation 3 in the main paper) rather than subjectively; and **(iii)** corrected examples are injected as negative-positive pairs that make the failure mode explicit, rather than simply expanding positive examples.

RIPA does not implement a full RL policy update — it does not optimize a reward model, update model parameters, or apply RL objectives such as PPO or DPO. Its connection to reinforcement learn-

Table 12: Representative examples from the v2 annotation instructions. C = confidence score.

Comment	Cat.	C	Sent.	C	Hate	C	Sarc.	C
গাজায় যুদ্ধ পরিস্থিতি খারাপ হচ্ছে	International	0.95	Negative	0.90	No	0.92	No	0.95
বাংলাদেশ ম্যাচ জিতে গেছে	Sports	0.97	Positive	0.88	No	0.98	No	0.97
আগামীকাল পরীক্ষা	Education	0.91	NaN	0.89	No	0.95	No	0.96
ওয়াও! কী দারুণ ব্যর্থতা!	Other	0.80	Negative	0.85	No	0.90	Yes	0.92
ওদের জাতটাই খারাপ	Other	0.75	Negative	0.82	Yes	0.94	No	0.91
`[emoji only]`	NaN	0.88	NaN	0.85	NaN	0.82	NaN	0.80
`https://t.co/xyz123`	NaN	0.91	NaN	0.91	NaN	0.88	NaN	0.90
দারুণ খেলেছে বাংলাদেশ, এরকম দল আর হয় না	Sports	0.90	Positive	0.55	No	0.95	No	0.60

ing is analogical: human corrections function as reward signals, and the annotation prompt functions as a policy adapted iteratively based on those signals. This framing motivates the design but does not constitute a claim of formal RL equivalence.

C.2 Confidence Score Distributions

Figure 5 reports the full confidence score distribution per dimension. A bimodal pattern is observed for Sentiment: 68.5% of scores cluster in the 0.60–0.70 range and 28.5% in the 0.85–0.90 range, with no scores in the 0.80–0.85 or 0.90–1.00 bands. This reflects genuine bimodality in Claude's certainty — instances are either clearly evaluative (high confidence) or clearly non-evaluative (low confidence), with little middle ground. This bimodal gap provides additional post-hoc justification for $\tau = 0.80$: the threshold naturally separates the two clusters rather than being an arbitrary cut-off.

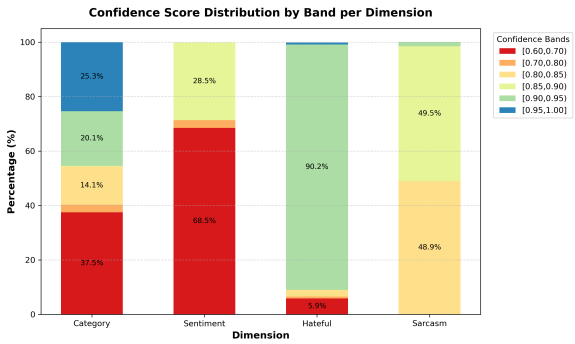

Figure 5: Confidence score distributions across dimensions. Sentiment shows a bimodal pattern consistent with the $\tau = 0.80$ threshold choice.

C.3 Disagreement Adjudication Procedure

The adjudication stage resolves conflicting low-confidence predictions across annotators or models following three principles: **(i)** Semantic Applicability First — determine whether the dimension genuinely applies before selecting among closed-set labels; **(ii)** Pragmatic Interpretation — resolve irony, sarcasm, and implicit hostility using contextual rather than purely lexical interpretation; and **(iii)** Confidence Preservation — retain uncertainty when ambiguity remains unresolved rather than artificially inflating confidence.

Table 13 presents representative adjudication cases illustrating these principles.

C.4 Operational Failure Modes

Several classes of comments remained intrinsically challenging for stable interpretation:

- **Context-Dependent Sarcasm.** Sarcastic intent frequently depended on external conversational context unavailable within isolated comments.
- **Implicit Collective Targeting.** Distinguishing generalized political criticism from identity-based hostility occasionally remained ambiguous.
- **Pragmatic Compression.** Extremely short comments often lacked sufficient semantic content for stable interpretation.
- **Irony Without Lexical Markers.** Some sarcastic comments contained no explicit contradiction cues, requiring pragmatic inference.
- **Borderline Applicability.** Factual statements with weak emotional orientation created difficult Neutral vs. NaN decisions.

These limitations reflect intrinsic ambiguity in naturalistic social media discourse rather than annotation inconsistency.

Table 13: Representative disagreement and adjudication cases.

Comment	Annotator A	Annotator B	Final	Rationale
বাহ! ঠিক সময়েই আবার ইন্টারনেট গেল	Negative	Sarcasm	Sarcasm	Surface-positive wording contradicts negative situation; pragmatic irony overrides literal sentiment.
আজ সকাল ৯টায় কলেজের খেলা	Neutral	NaN	NaN	Informational content without emotional orientation; sentiment is absent rather than balanced.
এই মানুষগুলো দেশটার ক্ষতি করছে	No	Yes	Yes	Generalizes hostility toward a collective rather than criticizing a specific policy or action.
ওরা আবার উন্নয়নের গল্প শুরু করেছে	No	Yes	Yes	Dismissive repetition pattern implies ironic intent despite absence of explicit sarcasm markers.

D Human Annotator Details

D.1 Annotator Profiles

Three Bengali-speaking annotators participated in the Stage 4 human validation loop. All are native Bengali speakers with active social media literacy. Table 14 summarizes their profiles.

Table 14: Human annotator profiles for Stage 4 validation.

Annotator	Native Speaker	Education	NLP Background	Social Media Literacy
A1	Yes	Graduates	Yes	Active daily user
A2	Yes	Graduates	Yes	Active daily user
A3	Yes	Graduates	Yes	Active daily user

D.2 Annotation Task

Each annotator independently reviewed instances routed to the human queue ($c < \tau = 0.80$, NaN flag set, or Full-conflict). For each instance, annotators were shown: **(i)** the original Bengali comment; **(ii)** the Stage 2 consensus label; **(iii)** the three Stage 1 annotator labels; and **(iv)** Claude's rationale. Each annotator then selected one of three actions: *agree*, *correct*, or *NaN-override*. Annotators worked independently with no cross-communication. Sessions were time-limited to 30 minutes per batch to prevent annotator fatigue. All annotators completed a calibration round of 50 instances before the main task, with disagreements discussed to establish shared boundary conditions. Annotators were compensated at a rate of 3USD per hour and were provided the option to withdraw at any point without penalty.

D.3 Human Inter-Annotator Agreement

Table 15 reports pairwise and Fleiss' κ among the three human annotators on the Stage 4-reviewed subset. Values reflect post-adjudication outputs used to compute reward signals.

Table 15: Inter-annotator agreement (Fleiss' κ) among three human annotators on the Stage 4 human-reviewed subset, post-adjudication.

Dimension	A1–A2	A1–A3	A2–A3	Fleiss' κ
Category	0.74	0.54	0.74	0.67
Sentiment	0.84	0.72	0.84	0.80
Hateful	0.78	0.61	0.78	0.72
Sarcasm	0.22	0.12	0.22	0.19

Table 16: Cross-lingual pipeline results on 2,030 English-dominant comments. NaN% and auto-acceptance rates are broadly consistent with the Bengali corpus.

Dimension	NaN%	Auto-acc.%	RIPA triggered
Category	30.25%	10.54%	Yes
Sentiment	0.79%	27.38%	Yes
Hateful	0.84%	97.82%	Yes
Sarcasm	10.74%	97.03%	Yes

E Additional Experimental Results

E.1 Cross-Lingual Validation

Table 16 reports NaN rates and auto-acceptance rates on the 2,030 English-dominant comment subset. Results are broadly consistent with the Bengali corpus across all four dimensions.

E.2 Cross-Dataset Validation

To assess whether NaN suppression is a MultiSoc-4D-specific artifact, CRADLE was applied to stratified samples from two existing Bengali benchmarks: SentNoB (sentiment dimension only, $n = 1{,}000$) and BanTH (hatefulness dimension only, $n = 1{,}000$). Table 17 summarizes pipeline behavior across all corpora.

RIPA triggered on all corpora, indicating that reward signals identify annotation failure patterns beyond the original corpus. NaN rates on SentNoB (6.6%) and BanTH (4.2%) are lower than

MultiSoc-4D's dimension-level rates — expected, since both are task-curated corpora more likely to contain semantically applicable instances. The presence of non-trivial NaN rates even in curated benchmarks is consistent with forced-label assignment occurring across annotation schemes. All-NaN = 0.0% was observed on all three external corpora, confirming that NaN suppression is a property of the annotator LLMs rather than a feature of Bengali social media text specifically. We treat these as directional observations; full re-annotation and downstream benchmarking are left to future work.

E.3 Per-Class Performance Under v2 Supervision

Table 18 reports per-class F1 and Precision under v2 supervision for the best-performing model per architecture family: Mistral 7B (LLM), RemBERT (multilingual), and BanglaBERT (monolingual). NaN instances are excluded from both training and evaluation. "Majority class" refers to the most frequent label per dimension under v2 (Other, Negative, No, No). "Minority class" refers to the least frequent non-NaN label (e.g., Yes for Hateful and Sarcasm). Full per-class results for all 11 models are available in the dataset release.

From Table 18, we can see Sarcasm Yes/No F1 extremes reflect the severe class imbalance (55 Yes instances out of 28,036); models either predict all-Yes or all-No, yielding degenerate per-class scores despite reasonable Macro-F1.

E.4 NaN-Removal Impact on Key Classes

Table 19 reports per-class F1 under v2 supervision for the minority and majority classes most affected by label collapse in v1. We report v2 results only; v1 per-class breakdown is omitted as collapsed label distributions (e.g., Neutral at 65.4%) make class-level comparison misleading.

E.5 Error Analysis

Table 20 summarizes the primary failure modes observed across CRADLE pipeline stages. The most frequent failure mode — context-dependent sarcasm — directly explains the Sarcasm dimension's minimal Macro-F1 improvement (mean $\Delta = +0.02$). These failure modes are consistent with known challenges in pragmatic annotation of naturalistic social media text (Swayamdipta et al., 2020).

F Full Benchmark Results

F.1 Full Accuracy Results: v1 vs. v2

Table 21 reports classification accuracy for all 11 models across four dimensions under v1 and v2 supervision. As noted in the main paper, accuracy is a misleading metric on label-collapsed datasets — high v1 accuracy on Hateful and Sarcasm reflects majority-class prediction (No: 95.4% and 96.4% respectively) rather than genuine classification. Macro-F1 (Table 8) is the primary reported metric.

F.2 Combined Macro-F1 and Accuracy Under v2

Table 22 combines Macro-F1 and Accuracy for all 11 models under v2 supervision only. The accuracy–F1 gap on Hateful and Sarcasm reflects residual class imbalance even after NaN removal (Hateful Yes: 19.7%; Sarcasm Yes: 1.3%).

Table 17: Cross-dataset pipeline results. NaN% reflects the proportion of instances judged semantically inapplicable. RIPA triggered on all corpora. Macro-F1 Δ is the mean across all four dimensions.

Dataset	**Language**	**Task**	n	**NaN%**	**RIPA**	**Macro-F1 Δ**
English-2030	English	All four	2,030	0.5%	Yes	+0.10
SentNoB	Bengali	Sentiment	1,000	6.6%	Yes	+0.14
BanTH	Bengali	Hatefulness	1,000	4.2%	Yes	+0.12

Table 18: Per-class F1 and Precision under v2 supervision, best model per family.

		Category		**Sentiment**		**Hateful**		**Sarcasm**	
Family	**Class**	**F1**	**Prec.**	**F1**	**Prec.**	**F1**	**Prec.**	**F1**	**Prec.**
LLM (LLaMA3.2 3B)	Economy	0.90	0.94	—	—	—	—	—	—
	Education	0.93	0.98	—	—	—	—	—	—
	Entertainment	0.92	0.95	—	—	—	—	—	—
	International	0.89	0.91	—	—	—	—	—	—
	National	0.93	0.96	—	—	—	—	—	—
	Other	0.92	0.93	—	—	—	—	—	—
	Sports	0.96	0.97	—	—	—	—	—	—
	Technology	0.94	o.90	—	—	—	—	—	—
	Positive	—	—	0.98	0.97	—	—	—	—
	Negative	—	—	0.89	0.95	—	—	—	—
	Neutral	—	—	0.98	0.98	—	—	—	—
	Yes	—	—	—	—	0.96	0.98	0.99	0.99
	No	—	—	—	—	0.85	0.80	0.00	0.00
Multi. (RemBERT)	Economy	0.95	0.98	—	—	—	—	—	—
	Education	0.92	0.94	—	—	—	—	—	—
	Entertainment	0.94	0.98	—	—	—	—	—	—
	International	0.93	0.93	—	—	—	—	—	—
	National	0.97	0.99	—	—	—	—	—	—
	Other	0.96	0.98	—	—	—	—	—	—
	Sports	0.92	0.94	—	—	—	—	—	—
	Technology	0.95	0.99	—	—	—	—	—	—
	Positive	—	—	0.99	0.99	—	—	—	—
	Negative	—	—	0.93	0.94	—	—	—	—
	Neutral	—	—	0.98	0.98	—	—	—	—
	Yes	—	—	—	—	0.86	0.79	0.00	0.00
	No	—	—	—	—	0.96	0.99	0.99	0.99
Mono. (BanglaBERT)	Economy	0.90	0.95	—	—	—	—	—	—
	Education	0.86	0.86	—	—	—	—	—	—
	Entertainment	0.90	0.96	—	—	—	—	—	—
	International	0.89	0.89	—	—	—	—	—	—
	National	0.94	0.95	—	—	—	—	—	—
	Other	0.93	0.95	—	—	—	—	—	—
	Sports	0.86	0.86	—	—	—	—	—	—
	Technology	0.93	0.95	—	—	—	—	—	—
	Positive	—	—	0.95	0.95	—	—	—	—
	Negative	—	—	0.85	0.83	—	—	—	—
	Neutral	—	—	0.94	0.94	—	—	—	—
	Yes	—	—	—	—	0.84	0.81	0.30	0.88
	No	—	—	—	—	0.96	0.97	0.99	0.99

Table 19: Per-class F1 under v2 supervision for key dimensions, best model per family (LLM = LLaMA3.2 3B; Multi. = RemBERT; Mono. = BanglaBERT). Minority classes (Hateful Yes, Sarcasm Yes) represent CRADLE's primary correction targets. Non-Other F1 is the mean across all non-Other Category classes.

Dim.	**Class**	**LLM**	**Multi.**	**Mono.**	**Note**
Sentiment	Positive	0.98	0.99	0.95	recovered
	Negative	0.89	0.93	0.85	recovered
	Neutral	0.98	0.98	0.94	collapsed in v1
Hateful	Yes	0.96	0.86	0.84	minority target
	No	0.85	0.96	0.96	majority
Sarcasm	Yes	0.99	0.00	0.30	minority target
	No	0.00	0.99	0.99	majority
Category	Other	0.92	0.96	0.93	collapsed in v1
	Non-Other	0.92	0.94	0.90	recovered

Table 20: Primary annotation failure modes observed during CRADLE pipeline processing.

Failure Mode	**Dimension**	**Description**	**Example Pattern**	**Frequency**
Context-dependent sarcasm	Sarcasm	Sarcastic intent depends on external context unavailable in isolated comments	Surface-positive wording with no lexical irony markers	High
Implicit collective targeting	Hateful	Distinguishing political criticism from identity-based hostility	"These people are damaging the country"	Moderate
Pragmatic compression	All	Extremely short comments lack sufficient semantic content for stable interpretation	Single emoji, 2–3 word fragments	Moderate
Neutral–NaN boundary	Sentiment	Factual statements with weak emotional orientation	"Tomorrow is the exam"	Moderate
Code-mixed register switching	Sarcasm	Sarcasm through Bangla-English mixing that models cannot reliably interpret	Mixed-script ironic comments	Low–Moderate
RIPA non-convergence	Sarcasm, Hateful	Instances escalated to human adjudication	Intrinsically ambiguous pragmatic cases	Low

Table 21: Classification accuracy: v1 vs. v2 labels across all 11 models.

		Category		Sentiment		Hateful		Sarcasm	
Family	**Model**	**v1**	**v2**	**v1**	**v2**	**v1**	**v2**	**v1**	**v2**
LLM	Qwen2.5 0.5B	0.77	0.62	0.61	0.70	0.94	0.88	0.94	0.98
	LLaMA3.2 3B	0.76	0.94	0.78	0.97	0.96	0.94	0.96	0.99
	Mistral 7B	0.76	0.90	0.77	0.98	0.96	0.80	0.96	0.99
Multi.	XLM-RoBERTa	0.52	0.78	0.66	0.98	0.95	0.80	0.96	0.99
	RemBERT	0.78	0.95	0.66	0.98	0.95	0.94	0.96	0.99
	MuRIL	0.66	0.94	0.78	0.94	0.96	0.94	0.96	0.99
	IndicBERT	0.77	0.94	0.79	0.96	0.97	0.93	0.96	0.99
Mono.	BanglaBERT	0.75	0.92	0.79	0.94	0.97	0.93	0.96	0.99
	FinBERT	0.72	0.79	0.69	0.92	0.95	0.94	0.96	0.99
	Electra	0.52	0.48	0.66	0.48	0.95	0.80	0.96	0.99
	DeBERTa	0.70	0.88	0.77	0.95	0.96	0.94	0.96	0.99

Table 22: Combined Macro-F1 and Accuracy under v2 supervision for all 11 models.

		Category		Sentiment		Hateful		Sarcasm	
Family	**Model**	**M-F1**	**Acc.**	**M-F1**	**Acc.**	**M-F1**	**Acc.**	**M-F1**	**Acc.**
LLM	Qwen2.5 0.5B	0.54	0.62	0.54	0.70	0.78	0.88	0.49	0.98
	LLaMA3.2 3B	0.92	0.94	0.95	0.97	0.90	0.94	0.50	0.99
	Mistral 7B	0.88	0.90	0.96	0.98	0.88	0.90	0.50	0.99
Multi.	XLM-RoBERTa	0.67	0.78	0.96	0.98	0.45	0.80	0.50	0.99
	RemBERT	0.94	0.95	0.97	0.98	0.91	0.94	0.50	0.99
	MuRIL	0.92	0.94	0.91	0.94	0.91	0.94	0.50	0.99
	IndicBERT	0.93	0.94	0.93	0.96	0.89	0.93	0.65	0.99
Mono.	BanglaBERT	0.90	0.92	0.92	0.94	0.90	0.93	0.65	0.99
	FinBERT	0.75	0.79	0.89	0.92	0.91	0.94	0.50	0.99
	Electra	0.24	0.48	0.23	0.48	0.45	0.80	0.50	0.99
	DeBERTa	0.84	0.88	0.92	0.95	0.90	0.94	0.49	0.99