VASTU: Language Models Struggle to Recognize Online Community Values

Agam Goyal, Xianyang Zhan, Charlotte Lambert,
Koustuv Saha, Eshwar Chandrasekharan
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign, Urbana, IL, USA
{agamg2, zhan39, cjl8, ksaha2, eshwar}@illinois.edu

Abstract

Online communities develop distinct norms for content they collectively value, yet it remains unclear whether current language models can recognize locally valued contributions in context. We formalize this as **community-conditioned preference prediction** and introduce VASTU (Value-Aware Social Tuning), a benchmark of 75,000 Reddit comments from 15 communities spanning Gaming, Science, Q&A, Advice, and Politics. We evaluate four model families—prompted LLMs, LoRA-adapted SLMs, supervised encoders, and feature-based classifiers—across global, local, and context-conditioned settings. Our central finding is that parametric adaptation consistently outperforms prompting: supervised encoders reach 0.74 AUROC and fine-tuned SLMs 0.64–0.71, while the best prompted result is only 0.62. This gap is not merely quantitative—vanilla prompting yields over 80% false-negative rates, systematically discarding content communities actually value. Conversational context narrows but does not close this divide. Together, these results suggest that local preference recognition requires community-specific training signal, not just better prompting. Our work supports future research on community-aware reward modeling, feed curation, and positive moderation.[1]

1 Introduction

Research on online communities has focused heavily on detecting harmful content such as toxicity, hate speech, and misinformation (Jurgens et al., 2019; Vidgen et al., 2021; Nandi et al., 2025). But content moderation and curation are not just punitive tasks. Communities also need systems that can recognize and surface *constructive* contributions—comments that provide evidence, clarify disagreements, offer support, or exemplify valued behavior (Grimmelmann, 2015; Wang and Diakopoulos, 2021; Lambert et al., 2024; Jundi et al., 2025). However, before such systems can generate interventions, recommend exemplars to newcomers, or guide moderators, they need a detection model: one that can *identify which contributions a particular community values and considers desirable.*

The key challenge is that community-valued content is not the same as generic quality, helpfulness, or non-toxicity: a comment can be constructive in the abstract yet redundant in context, or valued in one community while penalized in another (Weld et al., 2022; Papakyriakopoulos et al., 2023; Goyal et al., 2026b). Prior work on community preferences takes two forms: *characterizing* community norms by decomposing them into interpretable preference dimensions or cultural value frameworks (Park et al., 2024; Li et al., 2025; Zhang et al., 2026), and *aligning* model outputs with community preferences through benchmarking or preference optimization (Sorensen et al., 2024; Kumar et al., 2025; Lin and Wei, 2026; Kim et al., 2026). We study a different capability: can models predict, for a naturally occurring human-written content in its conversational context, whether a community would approve of it? This task requires understanding real online discourse dynamics, where approval depends on community-specific norms, and has direct implications for value-aware feed reranking and positive moderation (Jahanbakhsh et al., 2026).

The consequences of model performance on this task has direct implications for NLP research. For example, a community-conditioned preference detector can function as a reward model. Given a set of candidate responses, it can rank by predicted local approval (Goyal et al., 2026a), construct weak preference pairs for community-conditioned fine-tuning of language models (Kolluri et al., 2025), or retrieve valued exemplars for in-context learning (Chen et al., 2025). This detection problem is thus a prerequisite for building generative systems

[1]Code for experiments and the VASTU benchmark can be found inhttps://github.com/scuba-illinois/VASTU.

that adapt to *pluralistic* community norms without collapsing them into a single generic notion of quality (Feng et al., 2024; Zhang et al., 2025).

We address these gaps by formalizing this as **community-conditioned preference prediction**, where given a community c, a target comment x, and optional conversational context z (parent context), predict whether x is likely to be locally valued by c. We treat upvotes as behavioral signals of local community approval and aggregate preferences. To evaluate models on this task, we introduce VASTU (Value-Aware Social Tuning), a benchmark of 75,000 Reddit comments from 15 communities spanning Gaming, Science, Q&A, Advice, and Politics, with community-specific approval labels and contextual variants.

Using VASTU, we conduct a comprehensive analysis to study three research questions:

RQ1: How well do current NLP models perform on community-conditioned preference prediction? We compare feature-based baselines, supervised encoders, prompted LLMs, and LoRA-adapted SLMs under shared evaluation settings.

RQ2: Which forms of community and conversational context improve detection? We ablate input context (comment-only, community-aware, and parent-aware variants) to test whether models benefit from local grounding beyond comment text.

RQ3: Where do current frontier models fail? We perform a quantitative error profiling and qualitatively develop an error taxonomy to dive deeper into shortcomings of evaluated models.

Summary of Findings: Community-conditioned preference prediction remains far from solved. Our key finding is that adaptation outperforms prompting across all model sizes: locally fine-tuned encoders reach 0.71–0.74 AUROC, fine-tuned SLMs 0.64–0.69, and the best prompted result with full conversational context is only 0.62. Beyond aggregate performance, error profiles reveal a deeper asymmetry: vanilla LLMs discard over 80% of community-valued content, applying generic quality heuristics that miss community-specific signals such as inside references, casual humor, and direct personal experience. Prompt engineering reduces this bias substantially but cannot eliminate it. Conversational context helps across all paradigms but does not substitute for community-specific adaptation—local models consistently outperform global ones, and what predicts value in one community type is often penalized in another.

2 Related Work

Content Values in Online Communities: Understanding what online communities value has been approached through multiple methodological lenses. Survey-based studies have developed taxonomies of community values from user perspectives (Weld et al., 2022) and moderator perspectives on what behaviors they wish to encourage (Lambert et al., 2024), revealing that qualities like prosociality, engagement, and content quality are commonly desired. Complementing these qualitative approaches, Goyal et al. (2026b) used an LLM-based framework to identify values present in highly-upvoted Reddit comments at scale, uncovering values at macro, meso, and micro scales. Research on Reddit's voting mechanisms has shown that upvotes serve as signals of community-approved discourse styles, with Papakyriakopoulos et al. (2023) demonstrating this for political rhetoric in political subreddits. Other computational approaches to measuring valued behavior have largely focused on specific attributes such as partisan animosity (Piccardi et al., 2025) and prosociality (Bao et al., 2021), detected using transformer-based approaches, or identifying causal drivers of community-valued content using feature-based models (Goyal et al., 2026a) for proactive content moderation. Recent work decomposes these community preferences into interpretable latent dimensions (Li et al., 2025) or maps them to cultural value frameworks (Zhang et al., 2026), revealing that community preference variation is both real and structurally complex. *Our work extends this line of work by comparing computational approaches for detecting valued-aligned content, evaluating whether models can learn the implicit value signals expressed through upvotes.*

Detecting and Surfacing Valued Content: Social media feed algorithms primarily optimize for engagement signals (Chan et al., 2026), which can amplify divisive content and misalign with user preferences (Cunningham et al., 2024; Milli et al., 2025). Recent work has explored value-aligned alternatives using platform-agnostic browser extensions that intercept and rerank feeds in real-time (Piccardi et al., 2026). Jia et al. (2024) introduced societal objective functions that translate social science constructs into ranking objectives. Building on this, Jahanbakhsh et al. (2026) implemented feeds ranked by Schwartz's Basic Hu-

man Values, showing that user-designed feeds diverge substantially from engagement-driven defaults. Systems like Bonsai (El Malki et al., 2026) enable natural language-based feed customization, while Alexandria (Kolluri et al., 2026) offers a pluralistic library of values for real-time reranking. CommunityBench (Lin and Wei, 2026) benchmarks community-level alignment of LLMs across diverse groups and tasks, showing that current models systematically struggle with community-specific preferences; our work complements this by focusing on the recognition of valued human-generated content in context rather than AI output alignment. A common thread across this work is the need for models that can reliably detect value-relevant content at scale. *Our work addresses this fundamental challenge by comprehensively comparing modeling approaches for detecting community-valued content.*

3 The VASTU Benchmark

We now introduce the evaluation setup.

Communities: We pick 15 topically diverse Reddit communities, grouped into five domains for balanced coverage and structured analysis: **Gaming** (*r/Games*, *r/2007scape*, *r/DestinyTheGame*); **Science** (*r/science*, *r/askscience*, *r/Futurology*); **Q&A** (*r/AskWomen*, *r/AskHistorians*, *r/IAmA*); **Advice** (*r/personalfinance*, *r/legaladvice*, *r/DIY*); and **Politics** (*r/CanadaPolitics*, *r/PoliticalDiscussion*, *r/socialism*).

Data curation: We extract comments posted in these communities between May 10, 2016 and Feb 4, 2017—a period for which data exists from prior content moderation research (Chandrasekharan et al., 2018, 2019; Lambert et al., 2022). We then remove deleted/removed comments, content from deleted/banned authors, moderator-distinguished posts, likely bots (identified by a deterministic username heuristic), and very short comments ($\leq$ 5 tokens). We operationalize community approval using subreddit-specific *score* (upvotes minus downvotes) quantiles: `HIGH` for comments above the 95th percentile and `LOW` for comments below the 70th percentile following prior work (Goyal et al., 2026b). This intentionally prioritizes separability over full coverage, to ensure the `HIGH` comments are capturing the upper end of content the community values. Prior work has validated using Reddit score to encode community-specific preferences while being a universal signal of approval (Papakyriakopoulos et al., 2023; Park et al., 2024; Lambert et al., 2025; Kumar et al., 2025). For each community, we then sample an equal number of `high` and `low` comments, capped at 2,500 per class (i.e., at most 5,000 comments per subreddit), producing a balanced dataset per community.

Context Variants: Each instance always includes the target comment text. Instances can additionally be conditioned on the subreddit name and description (community-aware), the parent post title and body (post-aware), or the direct parent comment (parent-aware). We evaluate models across these conditions to measure the marginal contribution of each context type.

Features: We compute five types of feature signals: **(i) Embeddings:** Dense representations using SentenceBERT (`all-mpnet-base-v2`) (Reimers and Gurevych, 2019). **(ii) Prosociality:** Following Bao et al. (2021), we measure six attributes: politeness, agreement, supportiveness, donation, laughter, and gratitude. **(iii) Surface-level features:** We compute sentiment via the VADER compound score (Hutto and Gilbert, 2014), Flesch reading ease (Kincaid et al., 1975), and interrogative style (fraction of sentences ending in "?"). **(iv) Toxicity:** We use Detoxify (Hanu and Unitary team, 2020), an open-source toxicity detection model, to estimate comment toxicity. **(v) LIWC:** We apply the LIWC 2015 dictionary (Pennebaker et al., 2015) to quantify psychosocial linguistic processes and social values, complementing features (i)–(iv) with broader coverage of implicit community norms (Weld et al., 2024; Park et al., 2024).

Model families and evaluation protocol: We evaluate four families of models: (i) traditional feature-based models; (ii) a supervised encoder fine-tuned on the task; (iii) prompted frontier LLMs evaluated across context conditions; and (iv) LoRA-adapted SLMs fine-tuned with and without community conditioning. We use community-stratified train/test splits with an 80/20 ratio and report AUROC, F1, and Brier calibration (Brier, 1950). For robustness, we evaluate one representative model per paradigm under a stricter (q=0.99) and a softer (q=0.90) `HIGH` threshold to assess sensitivity to the approval cutoff (Appendix §A.5).

4 RQ1: Performance on Community-Conditioned Preference Prediction

We compare four model families under shared evaluation settings: feature-based classifiers, supervised encoders, prompted LLMs, and LoRA-adapted SLMs.

4.1 Evaluated Models and Setup

Simple Baselines: We include two off-the-shelf baselines: **(1) 1-Toxicity** inverts the Detoxify toxicity score, testing whether non-toxic content proxies for valued content; and **(2) DialogRPT**, trained by Gao et al. (2020) on Reddit data to predict comment scores, which acts as the most natural zero-shot transfer baseline for this task.

Feature-Based Models: We evaluate two feature-based classifiers using the full feature set described in Section 3: L_2-regularized Logistic Regression, with regularization parameter C tuned via cross-validation over $\{0.01, 0.1, 1, 10, 100\}$, and XGBoost with learning rate 0.05, $\gamma = 1$, maximum depth 4, regularization $\lambda = 3$ and $\alpha = 1$, and early stopping after 50 rounds. Both models are evaluated with 5-fold cross-validation. A feature set ablation confirming the incremental contribution of each feature group is provided in Appendix §A.1.

Supervised Encoders: We fine-tune two encoder models on the classification task: *bert-base-uncased* (BERT) (Devlin et al., 2019) and ModernBERT (Warner et al., 2025). Both share the same architecture: transformer-encoded comment representations are combined with linguistic features via learnable fusion weights, then passed through a two-layer classification head (see Appendix §A.6). Encoders are fine-tuned end-to-end for three epochs with a learning rate of 2×10^{-5} and batch size 16.

Language Model Approaches: We evaluate language models in two paradigms—prompting frozen models and fine-tuning open-weight models.

(1) Prompted LLMs: We evaluate GPT-4o-mini and GPT-4o (Hurst et al., 2024) (standard) and GPT-5-mini (OpenAI, 2025) (reasoning model; uses *"medium"* reasoning effort.) under four prompting strategies: **(i) Zero-Shot:** The model receives the comment and subreddit description and predicts whether the comment falls in the top 5% of community upvotes; **(ii) Chain-of-Thought (CoT):** The model reasons step-by-step before predicting (Wei et al., 2022): identifying community values, assessing comment qualities, then synthesizing a prediction; **(iii) Few-Shot (k=3):** Three labeled examples from the same community (training set) precede evaluation. Examples are class-balanced and drawn from different threads to prevent leakage; and **(iv) Value-Augmented (VAP):** The model receives the Schwartz Basic Human Values framework (Schwartz, 1992; Schwartz et al., 2012) as an analytical lens (Epstein et al., 2026), assessing whether the comment's expressed values align with the community's priorities.

All models use temperature=0 to reduce variability. Full prompt templates are in Appendix §A.7.

(2) Fine-Tuned Small Language Models: We fine-tune Gemma-3-4B-IT (Team et al., 2025) and Qwen3-4B-Instruct (Yang et al., 2025) using 4-bit quantization and LoRA (Hu et al., 2022) for one epoch on community-specific training splits. Models are trained to output only the classification label (`HIGH`/`LOW`), enabling direct comparison with prompted LLMs at lower inference cost.

Training Scope: For feature-based classifiers and supervised encoders, we evaluate two training scopes. **Global** models are trained on data pooled from all 15 communities and learn shared patterns of community approval. **Local** models are trained separately per community to capture community-specific norms; we report the mean performance ($\pm$std) across all 15 local models.

Model	Scope	AUROC ↑	F1 ↑	Brier ↓
Feature-based Models				
LogReg	Global	0.65	0.61	0.23
LogReg	Local	0.69 ± 0.04	0.64 ± 0.03	0.22 ± 0.01
XGBoost	Global	0.66	0.61	0.23
XGBoost	Local	0.67 ± 0.04	0.63 ± 0.02	0.23 ± 0.01
Supervised Encoders				
BERT	Global	0.70	0.67	0.22
BERT	Local	0.71 ± 0.05	0.68 ± 0.03	0.23 ± 0.02
ModernBERT	Global	0.73	0.70	0.21
ModernBERT	Local	0.74 ± 0.04	0.72 ± 0.03	0.21 ± 0.02
Baselines				
1-Toxicity	–	0.45 ± 0.05	0.41 ± 0.04	0.81 ± 0.06
DialogRPT	–	0.59 ± 0.04	0.55 ± 0.06	0.17 ± 0.02

Table 1: **Detection performance across model families.** Global models train on pooled data; Local reports the mean ($\pm$std) across community-specific models. Per-topic breakdowns are in §A.2.

4.2 Results: Feature-Based/Encoders Models

Simple baselines are insufficient. The 1-Toxicity baseline performs below chance (0.45 AUROC), demonstrating that desirability and non-toxicity are distinct constructs: content can be non-toxic yet not valued, or seem toxic yet be highly rewarded by a

community. DialogRPT performs better (0.59 AUROC) and achieves strong calibration (0.17 Brier) owing to its score-prediction training objective, but still trails the simplest feature-based model by 0.06 points, indicating that Reddit-wide engagement patterns do not capture community-specific approval.

Local adaptation outperforms global training: Locally trained models outperform their global counterparts across all models (Table 1). The gap is largest for Logistic Regression (+0.04 AUROC) and smaller for fully fine-tuned encoders (+0.01–0.02), suggesting that model capacity can partially compensate for the absence of community-specific supervision, but not eliminate it. ModernBERT consistently outperforms BERT under both training scopes (+0.03 AUROC), achieving the strongest overall result at 0.74 AUROC. Performance is consistently highest for Politics communities and lowest for Science across all model families, highlighting that different communities have norms that models may learn with different levels of difficulty. See Table A1 in Appendix §A.2 for a detailed breakdown.

Calibration is consistent across approaches: We find that Brier scores are stable across evaluated models (0.22–0.24), indicating that while detection performance varies substantially, all models produce reasonably calibrated probability estimates—an important property for downstream moderation workflows where confidence thresholds guide decisions (Goyal et al., 2025).

See Appendix §A.3 for SHAP experiments on feature-based models for interpretability analysis.

4.3 Results: Language Model Approaches

Since LLMs return discrete predictions rather than probabilities, we omit Brier calibration metrics for this subsection.

Prompted LLMs underperform supervised approaches: Across all prompting strategies, LLMs fall well short of fine-tuned encoders and SLMs (Table 2). The best prompted result—GPT-4o with Few-Shot or Value-Augmented prompting (0.60 AUROC)—trails even simple Logistic Regression (0.65 Global) and is 0.14 points below the best fine-tuned encoder (ModernBERT Local: 0.74). This gap is consistent across all community types (see Table A2 in Appendix §A.2), mirroring the supervised model ordering: Politics remains most predictable and Gaming most challenging. These results indicate that recognizing locally valued content requires learning community-specific patterns that cannot be elicited through prompting alone.

Model	Strategy	AUROC ↑	F1 ↑
Prompted LLMs			
GPT-4o-mini	Zero-Shot	0.56	0.38
	Chain-of-Thought	0.55	0.39
	Few-Shot (k=3)	0.56	0.40
	Value-Augmented	0.55	0.45
GPT-4o	Zero-Shot	0.58	0.41
	Chain-of-Thought	0.56	0.40
	Few-Shot (k=3)	0.60	0.42
	Value-Augmented	0.60	0.48
Prompted LRM (Reasoning)			
GPT-5-mini	Zero-Shot	0.53	0.20
	Chain-of-Thought	0.55	0.37
	Few-Shot (k=3)	0.53	0.22
	Value-Augmented	0.57	0.38
Fine-tuned SLMs			
Gemma-3-4B-IT	LoRA + Zero-Shot	0.65	0.42
Qwen3-4B-Instruct	LoRA + Zero-Shot	0.64	0.43

Table 2: **Language model performance across paradigms.** Fine-tuned SLMs substantially outperform all prompted configurations. Reasoning models perform worst overall; Chain-of-Thought provides no consistent benefit, suggesting this task requires learned community norms rather than test-time reasoning.

Fine-tuned SLMs perform substantially better: LoRA fine-tuning significantly narrows the gap: Gemma-3-4B-IT (0.65 AUROC) and Qwen3-4B-Instruct (0.64 AUROC) both outperform all prompted LLMs including GPT-4o, despite being approximately 100× smaller. This aligns with recent findings that fine-tuned SLMs can match or exceed LLMs for content moderation tasks (Zhan et al., 2025). Fine-tuned SLMs approach feature-based model performance but still trail fully fine-tuned encoders by 0.09 points, suggesting that the combination of contextual representations and linguistic features captures signals that pure language modeling misses.

Reasoning capabilities do not help: Neither explicit reasoning prompts nor reasoning-optimized models improve performance. Chain-of-Thought fails to outperform Zero-Shot for any model, and in some cases degrades it (GPT-4o: 0.58 to 0.56 AUROC). GPT-5-mini achieves the worst overall results (0.53 AUROC, 0.20 F1 Zero-Shot), underperforming both standard LLMs and fine-tuned SLMs. This pattern suggests that predicting community-valued content is not a reasoning task amenable to chain-of-thought decomposition: it requires implicit knowledge of community norms that must be learned from examples rather than inferred at inference time.

Value-Augmented prompting improves calibra-

Model	comment-only		+community		+community+parent	
	AUROC↑	F1↑	AUROC↑	F1↑	AUROC↑	F1↑
Prompted LLMs						
GPT-4o-mini	0.56	0.38	0.57	0.37	0.59	0.41
GPT-4o	0.58	0.41	0.61	0.40	0.62	0.44
GPT-5-mini	0.53	0.20	0.54	0.24	0.55	0.21
Fine-tuned SLMs						
Gemma-3-4B	0.65	0.42	0.66	0.43	0.71	0.49
Qwen3-4B	0.64	0.43	0.65	0.41	0.70	0.48

Table 3: **Context ablation across model families.** Community and parent comment context improves prediction accuracy. SLMs continue to show the best overall performance, with AUROC reaching upto 0.71.

tion but not discrimination: Value-Augmented prompting consistently raises F1 across prompted models (e.g., GPT-4o: 0.41 to 0.48) by improving precision-recall trade-offs, likely because the Schwartz values framework helps calibrate decision thresholds. However, AUROC is unchanged, indicating no improvement in the underlying ability to rank valued over non-valued content. Structured analytical prompts improve the threshold decision but cannot substitute for supervised training on community-specific data.

5 RQ2: Impact of Community and Conversational Context on Prediction

We now test whether community metadata or parent-comment context helps prompted models recognize locally valued contributions. We evaluate three input conditions across all prompted models: *comment-only* (no context), *+community* (subreddit name and description prepended), and *+community+parent* (community metadata plus the direct parent comment).

Parent-comment context provides the most consistent AUROC gain: The `+community+parent` condition yields the largest improvement for every model. For prompted LLMs, AUROC gains over comment-only range from +0.02 (GPT-5-mini: 0.53 to 0.55) to +0.04 (GPT-4o: 0.58 to 0.62). Fine-tuned SLMs show even larger gains: Gemma-3-4B improves from 0.65 to 0.71 and Qwen3-4B from 0.64 to 0.70, both +0.06. The consistency of this pattern suggests that conversational grounding from the parent comment provides crucial context that neither model parametric scale nor community descriptions alone can replicate.

Community description alone is inconsistent and sometimes counterproductive: Adding community metadata without conversational grounding produces mixed results. GPT-4o-mini shows a slight F1 drop on `+community` (0.38 to 0.37) despite a marginal AUROC gain, while GPT-5-mini's F1 spikes to 0.24 on `+community` before dropping back to 0.21 on `+community+parent`. Qwen3-4B shows a similar F1 dip on `+community` (0.43 to 0.41). These inconsistencies suggest that community descriptions shift decision thresholds rather than providing reliable alignment signal, producing variable trade-offs between precision and recall across model types.

Fine-tuned SLMs show a more structured context response: For SLMs, the gain from `+community` to `+community+parent` is substantially larger than from comment-only to `+community` (approximately $6\times$ the AUROC improvement). This asymmetry is absent in prompted LLMs, where context gains are distributed more evenly. The pattern suggests that fine-tuning already encodes community-level signals parametrically, leaving parent-comment context as the primary remaining informational lever at inference time.

While context helps, the adaptation gap persists for prompted models: Even the strongest prompted condition—GPT-4o with `+community+parent` (0.62 AUROC)—remains below fine-tuned SLMs at comment-only (0.64–0.65) and well below locally trained encoders (BERT Local: 0.71, ModernBERT Local: 0.74). Fine-tuned SLMs with parent context (0.70–0.71), however, begin approaching the lower range of local encoder performance, suggesting that parametric adaptation combined with conversational grounding is the most promising direction.

6 RQ3: LLM Error Analysis

Recent work on content moderation has shown promising performance of language models on the punitive side of this task, i.e., removing norm-violating content (Kumar et al., 2024; Kolla et al., 2024; Goyal et al., 2025). However, our task captures the positive content curation side of moderation, where we see that language models perform much worse. In order to understand further how models fail, we conduct a targeted error analysis of false positives (FPs) and false negatives (FNs). This is also crucial as systematic errors can encode miscalibrated assumptions about community norms and lead to harmful ranking decisions even when overall performance might appear strong. We therefore examine misclassified instances across communities to identify recurring failure patterns.

This analysis yields both a quantitative error pro-

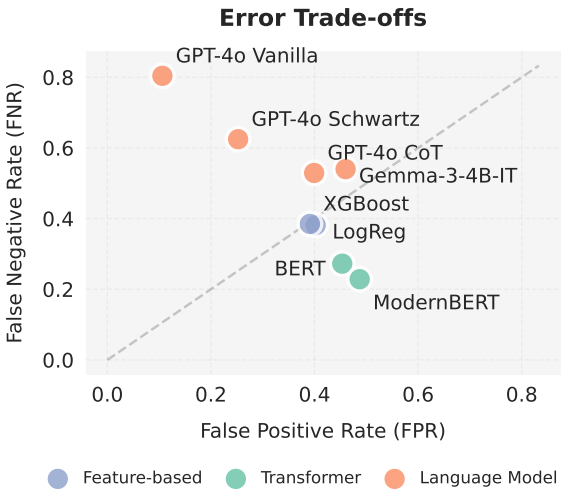


Figure 1: **Error trade-offs across method families.** While fine-tuned encoders and feature-based models tend to minimize False Negatives (prioritizing recall), vanilla LLMs show extreme conservatism in predicting valued content, resulting in high False Negative rates. Prompt engineering strategies and fine-tuning successfully shift LMs toward a more balanced error profile.

file, and a qualitative error taxonomy of distinct failure modes, which we use to diagnose when different approaches are likely to break down, and to surface concrete directions for future work, including development of hybrid systems.

Quantitative Error Profiles: Figure 1 quantifies these systematic biases by visualizing the trade-off between False Positive Rate (FPR) and False Negative Rate (FNR) for the best performing models in each category. We observe a clear divergence in how different model families approach the curation task. Supervised methods, including both feature-based and fine-tuned Transformers, cluster in the central and lower-right regions, favoring lower False Negative rates. This indicates a "high-recall" behavior essential for curation—they successfully surface the majority of community-valued content, at the cost of allowing some lower-quality items into the feed. In contrast, off-the-shelf LLMs, especially with vanilla prompting exhibit extreme conservatism, characterized by a really high FNR. In a curation context like ours, this behavior is highly undesired, as the model acts as a strict gatekeeper, discarding over 80% of the content the community actually values. Notably, prompting strategies that induce reasoning or apply the Schwartz values framework drive a significant shift along the error axes, reducing False Negatives by nearly 30 percentage points. However, even with these interventions, prompted LLMs remain significantly more conservative than fine-tuned models, suggesting that without parametric adaptation, general-purpose models struggle to recognize the idiosyncratic norms of specific subreddits.

Qualitative Error Taxonomy: To better understand the mechanism behind these quantitative gaps, we manually inspected 150 misclassified examples in the `comment-only` condition for GPT-4o with CoT using thematic analysis (Braun and Clarke, 2012) (10 per community), applying an inductive coding workflow. Two authors independently open-coded an initial subset of 50 explanations to identify recurring justifications (e.g., "fits community norms," "lacks context," "helpful/actionable," "hostile tone"). They then met to reconcile disagreements, refine and finalize a compact codebook with inclusion/exclusion criteria, and agree on higher-level themes. The first author then applied this finalized codebook to code the remaining instances. We subsequently derived a taxonomy of six distinct failure modes.

The most prevalent error types explain the high conservatism (FNR) observed in Figure 1. Specifically, **T2: "More effort is better"** and **T6: Profanity penalty** reveal a mismatch between the LLM's generic "helpfulness" prior and actual community dynamics. The model frequently rejects concise, witty, or casually profane comments that are highly valued by communities for their relatability, mistakenly flagging them as "low effort" or "unprofessional." Similarly, **T1: Wrong community norm** highlights the difficulty of zero-shot transferability of these models. For example, in gaming communities, the model often rejected inside jokes or memes as "irrelevant," failing to recognize that shared cultural references are a primary driver of value in that space. Conversely, False Positives often stemmed from **T4: Evidence bar mismatch**, where the model over-valued comments that *sounded* authoritative but were factually incorrect or missed the nuance required by expert communities like *r/AskHistorians*. This taxonomy underscores that while LLMs possess a strong general notion of "quality," they lack the specific grounding required for community-specific curation. The other two failure modes are **T3: Missing thread context** and **T5: Humor / sarcasm**. See Table A4 in Appendix §A.4 for expanded results and comparison to SHAP features for feature-based models.

7 Discussion and Implications

Beyond "One-Size-Fits-All" Models: Our findings challenge the prevailing assumption that "quality" is a universal construct that can be captured by a single, general-purpose model. The consistently high performance of community-specific local models over global baselines demonstrates that *value* is fundamentally contextual: what constitutes a high-quality contribution in *r/AskHistorians* which has rigorous discussions with a neutral tone is qualitatively different from valued content in *r/2007scape* which thrives on humor and community-specific jargon. For platforms, this suggests that the current paradigm of training a single reward model for platform-wide ranking may be inherently limited. Instead, we envision a shift toward decentralized and modular community-specific architectures. By deploying lightweight, community-specific adapters (e.g., LoRA modules), platforms can respect the pluralistic nature of human values (Kolluri et al., 2026; Sorensen et al., 2024). Future work could formalize this using explicit routing mechanism directing content to the specific expert models that understands that community's norms, similar to systems like MoMoE (Goyal et al., 2025). Such a system would allow for scalable personalization without flattening cultural nuances central to online communities.

Understanding the Limits of LLMs for Detecting Community-Valued Content: While LLMs are increasingly being looked at as the future of content curation (Epstein et al., 2026; Jahanbakhsh et al., 2026; Goyal et al., 2026b), our error analysis reveals that current off-the-shelf LLMs act as "moderators" rather than "curators." Vanilla LLMs exhibited prohibitively high False Negative rates (> 0.8), systematically discarding niche or implicit content that communities actively value. This behavior likely stems from safety-tuning (RLHF) that conflates "desirable" with "sanitized," failing to recognize that a low-effort joke or an edgy debate might be valued in specific communities.

We showed that strategies like Value-Augmented Prompting (VAP) can partially mitigate this by providing a structured lens for evaluation, improving calibration. However, we found that prompting alone cannot fully override the model's inherent conservatism and fine-tuning on community-specific data might be required. This points toward practical implications for future curation systems such as BONSAI (El Malki et al., 2026) and PILOT (Choi and Chandrasekharan, 2025): the actual discovery mechanism—the "engine" of feed curation—requires high-recall models that have been parametrically adapted to community preferences, which also aligns with the idea of "candidate generation" in RecSys literature (Covington et al., 2016; Huang et al., 2025). LLMs may instead be best suited for downstream roles such as explaining why content was ranked, providing interpretable rationales, or serving as a final sanity check rather than driving initial retrieval or ranking.

Towards Efficient Models: One of the key practical findings of our work is that massive scale of models is not a prerequisite for effective value alignment. We show that fine-tuned SLMs with just 4B parameters consistently outperformed LLMs, offering a favorable trade-off between accuracy and efficiency. For developers building real-time systems—whether feed rankers, recommendation engines, or moderation tools—this is a crucial insight as this makes it computationally feasible to run much more accurate, community-aligned rankers on edge, avoiding the API cost and rate limits of calling larger, closed-source models. Furthermore, this efficiency also unlocks new possibilities for user agency. If such "value" detectors are small and cheap to run, platforms can move beyond opaque algorithms for feed curation and offer users a catalog of "curation lenses", which could allow a user to swap in a *'science-focused'* ranker on their news feeds or a *'support-focused'* ranker for advice or well-being communities. This would democratize the alignment process, turning curation from a platform-imposed constraint into a user-controlled utility. Future research should explore the user experience of such controls, using VASTU for evaluating models powering them.

8 Conclusion

We show that recognizing community-valued contributions is a harder task than generic quality assessment—a finding that holds consistently across the four model families we evaluate on 15 communities. Value, or desirability, is fundamentally contextual: community-specific models consistently outperform global models, while the same linguistic features predict approval in opposite directions across community types. Fine-tuned SLMs and supervised encoders substantially outperform prompted LLMs, which exhibit a false-negative bias, discarding over 80% of community-

valued content under vanilla prompting. Conversational context narrows, but does not close, this gap, highlighting the importance of parametric adaptation to community-specific norms. These findings point toward modular, community-adapted architectures as a path for feed curation, positive moderation, and reward modeling in community-aware systems. VASTU provides a shared evaluation substrate for future work in this space.

Limitations

Approval labels and non-linguistic signals: Our study uses Reddit *score* as a proxy for community desirability. While this metric is a validated signal of collective approval, it remains an imperfect proxy subject to visibility bias, timing effects, and bandwagon dynamics that may conflate content quality with popularity. Our focus throughout is on evaluating model families based on language and content signals; we recognize that preference prediction also depends on non-linguistic factors such as comment timing, posting rank, and thread position. A more thorough investigation of how these structural signals interact with linguistic content is left for future work, as it requires purpose-built evaluation beyond the scope of this study.

Platform and temporal scope: Our analysis focuses exclusively on Reddit data from 2016–2017 due to open-sourced data availability. As a consequence, our findings reflect the specific demographics, platform norms, and community dynamics of that period and may not fully generalize to other social media environments or contemporary community behavior.

Interpretability and causality: While our SHAP and Chain-of-Thought analyses provide valuable interpretability, they identify correlations and self-reported reasoning rather than established causal drivers of community value. The features and reasoning patterns that predict approval may reflect surface regularities rather than the underlying norms communities actually hold.

Task formulation: We operationalize community value detection as binary classification, which captures preference at the extremes of the score distribution but does not directly evaluate listwise reordering in the information retrieval sense. Future work should connect these detectors to ranking and retrieval evaluations to assess their utility in downstream curation and recommendation settings. Similarly, since our comments are drawn from a relatively short, contiguous window of about 9 months, we did not structure the training and evaluation as a time-period based split where models learn to predict future values from from past data. However, future work could explore that setting with more longitudinal data available.

Experimental and model coverage: Due to budget constraints, we run a single evaluation seed for language models and reduce variability by setting temperature to 0, though some uncertainty may remain. We also leave community-specific LoRA fine-tuning and mixture-of-expert architectures for future exploration. More broadly, the rapid evolution of LLM capabilities means that newer architectures or agentic workflows could yield different performance trade-offs worth investigating.

Acknowledgments

This research was supported by NSF CAREER IIS-2439433.

This work used the Delta system at the National Center for Supercomputing Applications through allocation #240481 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

AI Involvement Disclosure

The research conducted in our work used artificial intelligence tools (e.g., Grammarly, ChatGPT) solely for improving the grammar and readability of certain sections of the manuscript, and for partial development of the codebase for the evaluations conducted through Claude Code, Codex, and Cursor. All scientific content, interpretations, and decisions reflect the original work, judgment and intellectual contributions of the research team.

Ethical Considerations

Our work uses publicly available Reddit data, which we handle in accordance with the platform's terms of service and privacy policies. All usernames and author identifiers are anonymized in our analysis. While our dataset focuses on highly-upvoted content, it may contain offensive language or controversial viewpoints reflective of diverse community norms. The authors do not endorse such content but analyze it to understand what communities reward in practice.

Our methods could potentially be misused by malicious actors to game recommendation systems or artificially inflate content visibility by reverse-engineering community preferences. Additionally, while value-aligned curation aims to improve user experience, overly optimized systems risk creating echo chambers or entrenching existing biases within communities. We emphasize that our models should be deployed with appropriate safeguards, transparency about ranking mechanisms, and ongoing monitoring for unintended consequences. Finally, platforms implementing such systems should provide users with meaningful control over curation algorithms, ensuring it serves user agency rather than replacing it.

References

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the web conference 2021*, pages 1134–1145.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Jackie Chan, Fred Choi, Koustuv Saha, and Eshwar Chandrasekharan. 2026. Examining algorithmic curation on social media: an empirical audit of reddit'sr/popular feed. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 391–406.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 1–30.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of the ACM on Human-Computer Interaction*, volume 2, pages 1–25.

Quan Ze Chen, Kevin Feng, Chan Young Park, and Amy X Zhang. 2025. SPICA: Retrieving scenarios for pluralistic in-context alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 748–765, Vienna, Austria. Association for Computational Linguistics.

Frederick Choi and Eshwar Chandrasekharan. 2025. Designing usable controls for customizable social media feeds. *arXiv preprint arXiv:2509.19615*.

Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Tom Cunningham, Sana Pandey, Leif Sigerson, Jonathan Stray, Jeff Allen, Bonnie Barrilleaux, Ravi Iyer, Smitha Milli, Mohit Kothari, and Behnam Rezaei. 2024. What we know about using non-engagement signals in content ranking. *arXiv preprint arXiv:2402.06831*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Omar El Malki, Marianne Aubin Le Quéré, Andrés Monroy-Hernández, and Manoel Horta Ribeiro. 2026. Bonsai: Intentional and personalized social media feeds. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Ziv Epstein, Farnaz Jahanbakhsh, Tiziano Piccardi, Isabel Gallegos, Dora Zhao, Johan Ugander, and Michael Bernstein. 2026. Whose values? measuring the (subjective) expression of basic human values in social media posts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 738–759.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Agam Goyal, Charlotte Lambert, and Eshwar Chandrasekharan. 2026a. The language of approval: Identifying the drivers of positive feedback online. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Agam Goyal, Charlotte Lambert, Yoshee Jain, and Eshwar Chandrasekharan. 2026b. Uncovering the internet's hidden values: An empirical study of desirable behavior using highly-upvoted content on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 923–938.

Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. 2025. MoMoE: Mixture of moderation experts framework for AI-assisted online governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12645–12660, Suzhou, China. Association for Computational Linguistics.

James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.*, 17:42. Publisher: HeinOnline.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junjie Huang, Jizheng Chen, Jianghao Lin, Jiarui Qin, Ziming Feng, Weinan Zhang, and Yong Yu. 2025. A comprehensive survey on retrieval methods in recommender systems. *ACM Transactions on Information Systems*, 44(1):1–43.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the AAAI ICWSM*, volume 8, pages 216–225.

Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, Zachary Robertson, Ziv Epstein, Sanmi Koyejo, and Michael S Bernstein. 2026. Value alignment of social media ranking algorithms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Chenyan Jia, Michelle S Lam, Minh Chau Mai, Jeffrey T Hancock, and Michael S Bernstein. 2024. Embedding democratic values into social media ais via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–36.

Iman Jundi, Eva Maria Vecchi, Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. It is not only the negative that deserves attention! understanding, generation & evaluation of (positive) moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11360–11395.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Jiwon Kim, Maya Ajit, Sherry Gong, Soorya Ram Shimgekar, Dong Whi Yoo, Eshwar Chandrasekharan, and Koustuv Saha. 2026. Llumi: Improving llm writing assistance for mental health support with online community feedback. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP), Main Conference*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Akaash Kolluri, Renn Su, Farnaz Jahanbakhsh, Dora Zhao, Tiziano Piccardi, and Michael S Bernstein. 2026. Alexandria: A library of pluralistic values for realtime re-ranking of social media feeds. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 1309–1330.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. 2025. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30096–30111, Suzhou, China. Association for Computational Linguistics.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A. Smith, and Hannaneh Hajishirzi. 2025. ComPO: Community preferences for language model personalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8246–8279, Albuquerque, New Mexico. Association for Computational Linguistics.

Charlotte Lambert, Frederick Choi, and Eshwar Chandrasekharan. 2024. "positive reinforcement helps breed positive behavior": Moderator perspectives on encouraging desirable behavior. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 548–559.

Charlotte Lambert, Koustuv Saha, and Eshwar Chandrasekharan. 2025. Does positive reinforcement work?: A quasi-experimental study of the effects of positive feedback on reddit. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. 2025. Prefpalette: Personalized preference modeling with latent attributes. In *Second Conference on Language Modeling*.

Jiayu Lin and Zhongyu Wei. 2026. Community-bench: Benchmarking community-level alignment across diverse groups and tasks. *arXiv preprint arXiv:2601.13669*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. 2025. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062.

Arghodeep Nandi, Megha Sundriyal, Euna Mehnaz Khan, Jikai Sun, Emily K. Vraga, Jaideep Srivastava, and Tanmoy Chakraborty. 2025. The psychology of falsehood: A human-centric survey of misinformation detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8508–8525, Suzhou, China. Association for Computational Linguistics.

OpenAI. 2025. Introducing GPT-5 — openai.com. https://openai.com/index/introducing-gpt-5/.

Orestis Papakyriakopoulos, Severin Engelmann, and Amy Winecoff. 2023. Upvotes? downvotes? no votes? understanding the relationship between reaction mechanisms and political discourse on reddit. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanu Mitra, David Jurgens, and Yulia Tsvetkov. 2024. ValueScope: Unveiling implicit norms and values via return potential model of social interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16659–16695, Miami, Florida, USA. Association for Computational Linguistics.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey Hancock, Jeanne L. Tsai, and Michael S. Bernstein. 2025. Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 390(6776):eadu5584.

Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne Tsai, and Michael Bernstein. 2026. Reranking social media feeds: A practical guide for field experiments. *Trans. Soc. Comput.*, 9(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Yixue Wang and Nicholas Diakopoulos. 2021. Highlighting High-quality Content as a Moderation Strategy: The Role of *New York Times* Picks in Comment Quality and Engagement. In *ACM Transactions on Social Computing*, volume 4, pages 1–24.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 2526–2547.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. What Makes Online Communities 'Better'? Measuring Values, Consensus, and Conflict across Thousands of Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1121–1132.

Galen Weld, Amy X Zhang, and Tim Althoff. 2024. Making online communities 'better': a taxonomy of community values on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1611–1633.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2025. SLM-mod: Small language models surpass LLMs at content moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8774–8790, Albuquerque, New Mexico. Association for Computational Linguistics.

Lily Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, and 1 others. 2026. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *International Conference on Learning Representations*, volume 2026, pages 147076–147215.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. 2025. Exploring chain-of-thought reasoning for steerable pluralistic alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25636–25649, Suzhou, China. Association for Computational Linguistics.

A Appendix

A.1 Feature Set Ablation

Before conducting computationally expensive encoder fine-tuning, we investigated which feature sets contribute most to detection performance. Using feature-based models (Logistic Regression and XGBoost), we evaluate five incremental configurations: (1) embeddings only, (2) + prosociality, (3) + toxicity, (4) + surface attributes, and (5) + LIWC. As shown in Figure A1, performance improves monotonically as features are added, with LIWC categories providing the largest marginal gain (+0.01 AUROC for both models). All experiments in the main paper use the full feature set.

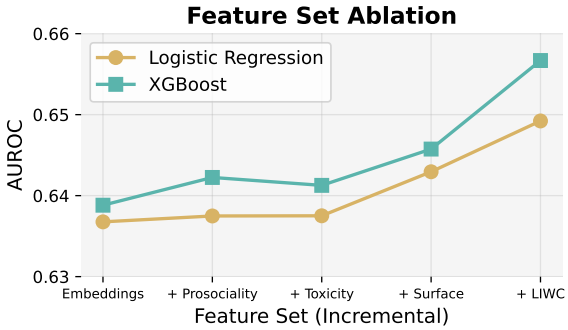


Figure A1: **Feature set ablation for feature-based models.** AUROC improves incrementally as feature sets are added, with LIWC providing the largest marginal gain.

A.2 Per-Topic Performance Breakdowns

Tables A1 and A2 report AUROC broken down by topic domain for feature-based/encoder models and language models respectively.

From Table A1, we observe meaningful variation in detection difficulty across topics. Politics communities show the highest predictability (0.76 AUROC for ModernBERT), possibly because valued content in these communities exhibits more consistent linguistic patterns (e.g., well-reasoned arguments, civility). Science and Gaming communities prove most challenging (0.64–0.72 AUROC), potentially due to the diverse nature of valued contributions across different scientific domains. Q&A and Advice communities fall in between, with moderate predictability (0.68–0.75 AUROC).

Similar patterns hold for the LLM and fine-tuned SLM predictions.

Topic	LogReg	XGBoost	BERT	ModernBERT
Gaming	0.67	0.66	0.69	0.72
Science	0.65	0.64	0.69	0.72
Q&A	0.68	0.68	0.72	0.75
Advice	0.68	0.67	0.71	0.74
Politics	0.69	0.69	0.73	0.76

Table A1: **Topic-level AUROC for feature-based and encoder models (local training).** Politics shows highest predictability; Science shows lowest across all architectures. Highlighted cells indicate per-column best and worst.

Topic	4o-mini	4o	5-mini	Gemma-3	Qwen3
Gaming	0.53	0.56	0.51	0.59	0.60
Science	0.55	0.55	0.51	0.63	0.62
Q&A	0.57	0.59	0.54	0.65	0.65
Advice	0.56	0.58	0.54	0.68	0.65
Politics	0.59	0.62	0.54	0.70	0.68

Table A2: **Topic-level AUROC for language models.** Prompted LLMs use Zero-Shot; fine-tuned SLMs use LoRA. Politics communities show the highest predictability across all models, while Gaming proves most challenging, mirroring the pattern in Table A1.

A.3 SHAP Analysis for Supervised Models

To understand what signals supervised models rely on, we compute SHAP values (Lundberg and Lee, 2017) for feature importance analysis. We focus on XGBoost as it achieves the strongest performance among feature-based models and produces stable importance estimates. While SHAP can be applied to transformer models, their learned internal representations make interpretation less straightforward, and we therefore restrict this analysis to feature-based approaches where input features map directly to interpretable constructs. To aid interpretation, we aggregate our 80+ individual features into 11 semantically coherent groups (e.g., *Prosociality* combines agreement, politeness, and support scores; *Cognitive* combines LIWC categories for cognitive processes, insight, causation, certainty, and tentativeness). Full groupings are provided in Figure A3. For each topic category, we compute mean signed SHAP values across communities and test instances, where positive values indicate the feature group pushes predictions toward "valued" and negative values push toward "not valued."

Feature importance varies substantially across community types: From Figure A2, we observe the heterogeneity in what signals predict valued content across topic categories. The most notable example is *Prosociality*: in Advice communities, prosocial language is the single strongest predictor of valued content (+0.121 mean SHAP), yet the same features show negative or negligible effects in Gaming (-0.056), Science (-0.048), and Politics (-0.029). Similarly, *Self-focus* (first-person singular pronouns) strongly predicts valued content in Advice (+0.098)—where personal experience is central—but is penalized in Science (-0.070) and Q&A (-0.066) communities that prioritize objective, generalizable information. This pattern reinforces our RQ1 finding that community-specific models outperform global ones: a model learning that prosociality universally predicts value would systematically misclassify content in Gaming and Politics communities.

Question-asking and readability show interpretable community-specific patterns: Beyond prosociality, other feature groups exhibit patterns that align with community purposes. *Questions* (interrogative framing) positively predict valued content in Q&A (+0.049) and Science (+0.036)—communities oriented around inquiry—but negatively predict value in Advice (-0.067) and Politics (-0.035), where users seek definitive guidance or argumentation rather than open questions. *Readability* shows broadly positive effects, particularly in Politics (+0.059) and Science (+0.057), suggesting that clear writing is valued in communities dealing with complex topics. Interestingly, *Cognitive* language (reasoning, causation, certainty) is penalized in Politics (-0.037) and Science (-0.023) but weakly positive in Advice (+0.013), perhaps because overly analytical framing reads as hedging in debate-oriented communities but signals thoughtfulness in support contexts. These interpretable patterns suggest that supervised models learn genuine community norms rather than spurious correlations in data.

Robustness of SHAP-derived patterns: To verify that the SHAP feature-group patterns are stable rather than seed- or split-dependent, we quantify stability along three axes: **(1) Cross-fold rank stability:** across the 5 CV folds, the mean pairwise Spearman correlation of feature-group importance rankings is $\bar{\rho} = 0.86$ (min 0.79); **(2) Cross-subreddit rank stability:** within each topic, the mean pairwise Spearman across the three constituent subreddits is $\bar{\rho} = 0.71$, indicating that group-level orderings are largely shared within a topic while retaining community-specific variation; and **(3) Top-k stability:** the top-5 feature groups have a mean cross-fold Jaccard overlap of 0.82,

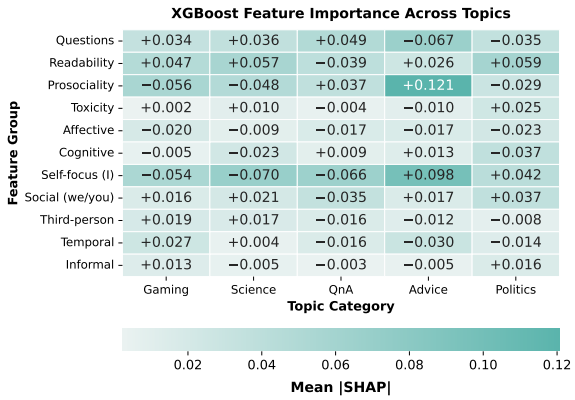


Figure A2: **SHAP feature importance for XGBoost across topics.** Cells show mean signed SHAP values aggregated across communities within each topic. Feature importance varies substantially by communities.

Feature Group	Constituent Features
Prosociality	agreement, politeness, support
Toxicity	toxicity
Affective	vader compound, posemo, negemo
Readability	flesch reading ease
Questions	interrogative fraction, interrog
Self-focus (I)	i
Social (we/you)	we, you, social
Cognitive	cogproc, insight, cause, certain, tentat
Temporal	focuspast, focuspresent, focusfuture, time
Informal	informal, swear, netspeak
Third-person	shehe, they

Table A3: Mapping of individual features to interpretable feature groups used in SHAP analysis.

with a coefficient of variation of of 0.11. Overall, these show that the SHAP analysis is fairly stable with no changes to our interpretation, and readers of our work can get better insights into how much they can rely on the analysis accordingly.

A.4 Extended Results on LLM Error Analysis

Table A4 presents the entire set of six failure modes identified and presented in §6. We additionally present a comparative analysis of the interpretive patterns we observe from the thematic analysis in comparison to the SHAP analysis in §A.3.

Feature-based and LLM signals are complementary but not identical: Comparing SHAP patterns to CoT themes highlights a key difference in *representation*. Feature-based models rely on measurable proxies (e.g., prosociality, interrogative style, self-focus, readability), which produce clear community-conditional effects (Figure A2). In contrast, LLM explanations operate at a more semantic and normative level—explicitly reasoning about *why* a community might value a contribution (e.g., "evidence-based," "on-topic," "actionable," "constructive")—but this abstraction flattens community nuance. In several incorrect cases (analyzed in RQ3), LLMs over-apply these generic heuristics even when those styles are rewarded in that community based on the SHAP analysis. Together, these results suggest that supervised feature-based models capture stable, interpretable correlates of community value, while LLM explanations reflect a higher-level normative theory that is often miscalibrated.

A.5 Threshold Sensitivity Analysis

Our primary experiments define `HIGH` as the top 5% of subreddit-specific scores (q=0.95) and `LOW` as the bottom 30% (q=0.70). To assess whether model rankings and key conclusions depend on this choice, we re-evaluate one representative model from each paradigm at two alternative `HIGH` thresholds—a softer cutoff (q=0.90) and a stricter cutoff (q=0.99)—holding all other conditions fixed.

Table A5 shows that absolute AUROC rises with threshold stringency: at q=0.99, the most highly upvoted comments are more linguistically distinctive, yielding a more separable classification problem. The shift is roughly uniform across paradigms ($\approx$+0.05 from q=0.95 to q=0.99, $\approx$−0.03–0.04 from q=0.95 to q=0.90), so no paradigm benefits disproportionately from a stricter label. Crucially, model rankings are fully preserved: ModernBERT $>$ LogReg $>$ Gemma-3-4B $>$ GPT-4o at every threshold, with Spearman $\rho = 1.00$ between each alternative and the default. These results confirm that the adaptation-versus-prompting gap reported in the main text is not an artifact of the specific approval cutoff chosen to define community-valued content.

A.6 Transformer-Based Model Architecture

Figure A3 shows a visual depiction of the transformer architecure we use for modeling community-valued content in our work with supervised encoders.

Type	What Typically Goes Wrong	Type
T1: Wrong community norm	Misunderstands what subreddits typically rewards (including community jargon), so comment is judged as "not a good fit."	FN
T2: "More effort is better"	Treats short / punchy / obvious comments as low-quality, even when that style is commonly upvoted.	FN
T3: Missing thread context	Judges replies as "unclear" because they rely on the quotes or shared context that is not present in the standalone text.	FN
T4: Evidence bar mismatch	Applies the wrong standard for evidence (too strict or too loose), especially in Science and Q&A communities.	FN / FP
T5: Humor / sarcasm	Misreads jokes, irony, or sarcasm (takes it literally or dismisses it), leading to mismatched value judgments.	FN / FP
T6: Profanity penalty	Over-penalizes profanity even when it is not targeted harassment and is common/acceptable in the community.	FN

Table A4: **LLM Error Taxonomy.** We summarize failure patterns in LLM misclassifications.

Model + Paradigm	q=0.90	q=0.95	q=0.99
GPT-4o Zero-Shot	0.57	0.60	0.64
Gemma-3-4B LoRA	0.62	0.65	0.70
LogReg Local	0.66	0.69	0.74
ModernBERT Local	0.70	0.74	0.79
Spearman ρ vs. q=0.95	1.00	—	1.00

Table A5: **Threshold sensitivity across paradigms.** AUROC at three `HIGH` label thresholds for one representative model per paradigm. Spearman ρ measures rank correlation of model AUROC scores relative to the defaultd.

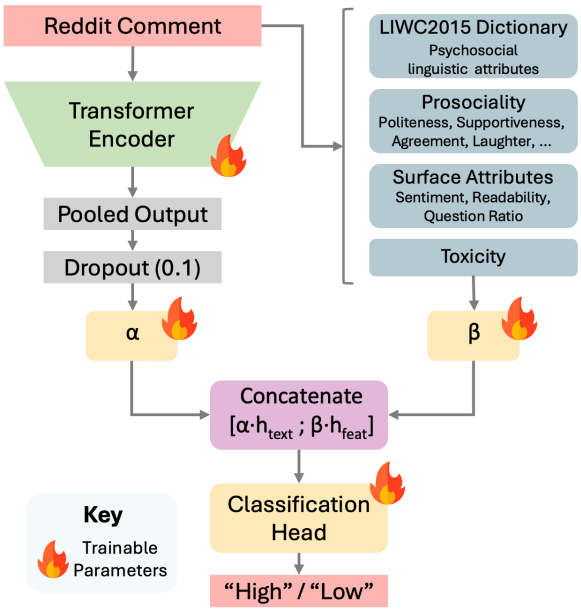


Figure A3: **Architecture of our transformer-based feature fusion model.** The input Reddit comment is processed through a transformer encoder to obtain contextual representations, while linguistic features are extracted in parallel. Learnable weights allow the model to learn the optimal contribution of text embeddings versus linguistic features. The concatenated representation is passed through a classification head to predict whether the comment will be highly-valued by the community.

A.7 Prompting Strategies for Large Language Models

We evaluate four distinct prompting strategies for LLM-based detection of community-valued content. Below we provide the complete prompts used for each condition. All prompts are shown for the `comment+community` context, with addition of the preceding comment for the `comment+community+parent` condition, and removal of the `community` context for the `comment-only` condition.

A.7.1 Condition 1: Zero-Shot Vanilla

The baseline condition with minimal instruction.

System Prompt:

> *You are an expert at understanding online community dynamics and what content resonates with different communities.*

User Prompt Template:

> *You are evaluating whether a comment would be highly upvoted in the Reddit community r/subreddit_name.*
>
> ***Community Description:*** *subreddit_description*
>
> ***Comment to Evaluate:*** *target_comment*
>
> *Would this comment likely receive high approval (top 5% of upvotes) from this community?*
>
> *Respond with exactly one of: HIGH or LOW*

A.7.2 Condition 2: Chain-of-Thought (CoT)

Adds explicit reasoning instruction to elicit step-by-step analysis.

System Prompt:

> *You are an expert at understanding online community dynamics and what content resonates with different communities. You think carefully and systematically about what makes content valuable to specific communities.*

User Prompt Template:

> *You are evaluating whether a comment would be highly upvoted in the Reddit community r/subreddit_name.*

Community Description: *subreddit_description*

Comment to Evaluate: *target_comment*

Think step-by-step:

1. *What is this community about and what do they typically value?*
2. *What qualities does the target comment have that this community might appreciate or dislike?*
3. *Based on this analysis, would this comment likely receive high approval (top 5% of upvotes) from this community?*

After your reasoning, provide your final prediction.

Format your response as:

PREDICTION: [HIGH or LOW]

REASONING: [BRIEF step-by-step analysis]

A.7.3 Condition 3: Few-Shot (k=3)

Provides three labeled examples from the same community before evaluation.

System Prompt:

> *You are an expert at understanding online community dynamics and what content resonates with different communities.*

User Prompt Template:

> *You are evaluating whether a comment would be highly upvoted in the Reddit community r/subreddit_name.*
>
> ***Community Description:*** *subreddit_description*
>
> *Here are some examples of comments from this community and whether they received high approval:*
>
> ***Example 1:***
>
> *Comment: example1_comment*
>
> *Approval: example1_label*
>
> ***Example 2:***
>
> *Comment: example2_comment*
>
> *Approval: example2_label*
>
> ***Example 3:***
>
> *Comment: example3_comment*
>
> *Approval: example3_label*
>
> *Now evaluate this comment:*
>
> ***Comment to Evaluate:*** *target_comment*
>
> *Would this comment likely receive high approval (top 5% of upvotes) from this community?*
>
> *Respond with exactly one of: HIGH or LOW*

A.7.4 Condition 4: Value-Augmented Prompting

Provides the Schwartz Basic Human Values framework (Schwartz, 1992; Schwartz et al., 2012) as an analytical lens for evaluation, based on the framework utilized by Epstein et al. (2026).

System Prompt:

> *You are an expert at understanding online community dynamics and human values. You use established psychological frameworks to analyze what content resonates with different communities.*

User Prompt Template:

> *You are evaluating whether a comment would be highly upvoted in the Reddit community r/subreddit_name.*
>
> ***Community Description:*** *subreddit_description*
>
> *To help with your evaluation, consider the Schwartz Basic Human Values framework. These are universal values that motivate human behavior:*
>
> 1. ***Self-Direction****: Independent thought and action (creativity, freedom, curiosity)*
> 2. ***Stimulation****: Excitement, novelty, and challenge*
> 3. ***Hedonism****: Pleasure and sensuous gratification*
> 4. ***Achievement****: Personal success through demonstrating competence*
> 5. ***Power****: Social status, prestige, control over resources*
> 6. ***Security****: Safety, harmony, stability of society and relationships*
> 7. ***Conformity****: Restraint of actions that might harm others or violate norms*
> 8. ***Tradition****: Respect and commitment to cultural/religious customs*
> 9. ***Benevolence****: Concern for the welfare of close others*
> 10. ***Universalism****: Understanding, tolerance, and protection of all people and nature*
>
> ***Comment to Evaluate:*** *target_comment*
>
> *Consider:*
>
> - *Which values does this comment express or appeal to?*
> - *Which values are likely important to this specific community?*
> - *Does the comment align with or violate values this community cares about?*
>
> *Based on this value-based analysis, would this comment likely receive high approval (top 5% of upvotes) from this community?*
>
> *Format your response as:*
>
> *PREDICTION: [HIGH or LOW]*
>
> *VALUES_EXPRESSED: [List the 1-3 most relevant Schwartz values this comment expresses]*
>
> *COMMUNITY_VALUES: [List the 1-3 Schwartz values you think this community prioritizes]*
>
> *ALIGNMENT: [Brief assessment of alignment]*

B Computational Resources

All experiments using open-source models were run on organization GPU servers equipped with 3xNVIDIA A40. The API costs for running OpenAI models were around $250.