Density-Guided Response Optimization: Community-Grounded Alignment via Implicit Acceptance Signals

PATRICK GERARD, Information Sciences Institute, University of Southern California, USA
SVITLANA VOLKOVA, Aptima Inc., USA

Language models deployed in online communities must adapt to norms that vary across social, cultural, and domain-specific contexts. Prior alignment approaches rely on explicit preference supervision or predefined principles, which are effective for well-resourced settings but exclude most online communities—particularly those without institutional backing, annotation infrastructure, or organized around sensitive topics—where preference elicitation is costly, ethically fraught, or culturally misaligned.

We observe that communities already express preferences implicitly through what content they accept, engage with, and allow to persist. We show that this acceptance behavior induces measurable geometric structure in representation space: accepted responses occupy coherent, high-density regions that reflect community-specific norms, while rejected content falls in sparser or misaligned areas. We operationalize this structure as an implicit preference signal for alignment and introduce *density-guided response optimization* (DGRO), a method that aligns language models to community norms without requiring explicit preference labels.

Using labeled preference data, we demonstrate that local density recovers pairwise community judgments, indicating that geometric structure encodes meaningful preference signal. We then apply DGRO in annotation-scarce settings across diverse communities spanning platform, topic, and language. DGRO-aligned models consistently produce responses preferred by human annotators, domain experts, and model-based judges over supervised and prompt-based baselines. We position DGRO as a practical alignment alternative for communities where explicit preference supervision is unavailable or misaligned with situated practices, and discuss the implications and risks of learning from emergent acceptance behavior.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; *Learning from implicit feedback*; • **Human-centered computing** → *Social content sharing*.

Additional Key Words and Phrases: language model alignment, community norms, implicit preferences, density estimation, computational social science, online communities, preference learning

ACM Reference Format:
Patrick Gerard and Svitlana Volkova. 2026. Density-Guided Response Optimization: Community-Grounded Alignment via Implicit Acceptance Signals. 1, 1 (January 2026), 27 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Language models increasingly interact with online communities whose norms, values, and communicative conventions vary widely across social, cultural, and domain-specific contexts. What counts as an appropriate response depends not only on topic, but on situated expectations around tone, evidence, empathy, authority, and care. A question about weight loss, for example, calls for fundamentally different responses in a medical advice forum, a peer support community, or an

Authors' Contact Information: Patrick Gerard, pgerard@isi.edu, Information Sciences Institute, University of Southern California, Marina Del Rey, California, USA; Svitlana Volkova, Aptima Inc., Dayton, Ohio, USA, svolkova@aptima.com.

Manuscript submitted to ACM

academic discussion space—not because the underlying facts differ, but because the social meanings and potential harms of speech differ across contexts. Capturing these distinctions is essential not only for safe and effective deployment of language models, but also for broader questions of algorithmic governance: who defines acceptable behavior, whose values are encoded, and how those values are operationalized in deployed systems.

Existing approaches to language model alignment have largely addressed these questions through explicit preference supervision. Reinforcement Learning from Human Feedback (RLHF) and related methods rely on annotated preference data to guide model behavior [13, 41], while Direct Preference Optimization (DPO) simplifies optimization but retains dependence on labeled comparisons [44]. Constitutional AI further reduces human annotation by introducing principle-based critiques [3]. While effective in settings where preferences can be clearly articulated and externally specified, these approaches presuppose that normative criteria are stable, consensual, and ethically straightforward to elicit. In practice, however, many online communities—particularly marginalized, informal, or sensitive ones—lack the institutional capacity, shared language, or ethical conditions required for explicit preference annotation. In such settings, asking external annotators to define "appropriate" behavior risks misrepresentation, cultural mismatch, or harm.

At the same time, community norms are not unexpressed. Online communities continuously enact and negotiate standards of appropriateness through moderation, participation, and collective attention. Content that aligns with community expectations is more likely to persist, receive engagement, and become part of ongoing discourse, while misaligned content is ignored, down-ranked, or removed. Importantly, these acceptance patterns are shaped not only by individual preferences, but also by power, platform affordances, and governance structures within the community. As such, behavioral acceptance should not be treated as normative endorsement or consent. Rather, it constitutes a descriptive signal of how norms are operationalized in practice, reflecting the values of those who are most able or willing to participate.

Building on prior work on implicit behavioral signals in recommender systems and information retrieval [26, 30], we study whether these naturally occurring acceptance patterns give rise to recoverable structure in representation space. We observe that responses accepted by a community are not randomly distributed; instead, they tend to cluster in coherent, high-density regions of embedding space, which we refer to as a community's *acceptance manifold*. This structure captures what a community treats as permissible or contextually normal, as enacted through collective behavior rather than prescribed by external rules. We emphasize that this manifold reflects descriptive regularities in community practice, not an ethical claim about which norms ought to be learned or deployed.

We operationalize this observation through **Density-Guided Response Optimization (DGRO)**, a method that uses local density in a community's embedding space as an implicit preference signal for alignment. DGRO does not assume that community norms are universally desirable or stable; instead, it provides a mechanism for studying and modeling how norms manifest in behavior when explicit preference supervision is unavailable or inappropriate. We first validate the underlying manifold hypothesis on labeled preference data, showing that local density correlates monotonically with observed human judgments. We then demonstrate that this signal can substitute for explicit preference annotations within standard preference optimization objectives. Finally, we apply DGRO in annotation-scarce settings across diverse communities, including eating disorder support spaces and Russian-language conflict documentation forums, and evaluate whether aligned models produce responses judged as more contextually appropriate and authentic.

This work makes three contributions. First, we provide empirical evidence that community acceptance behavior induces structured, locally coherent geometry in representation space that encodes recoverable preference signal. Second, we introduce DGRO as a practical, annotation-free mechanism for leveraging this structure in preference-based alignment. Third, we analyze the ethical implications and limitations of learning from acceptance behavior, including

risks of bias amplification, exclusion, and manipulation, and situate DGRO as a descriptive alignment tool whose deployment requires careful governance and oversight.

2 Related Work

Alignment from Explicit Preferences Most modern alignment methods assume access to explicit human preference supervision. Reinforcement Learning from Human Feedback (RLHF) learns a reward model from annotated pairwise comparisons and optimizes a policy via reinforcement learning [13, 41]. While effective, this paradigm requires large volumes of carefully curated preference data and a multi-stage training pipeline. Direct Preference Optimization (DPO) simplifies optimization by removing the reward model and reinforcement learning stage, but still fundamentally relies on explicit preference labels [44]. Constitutional AI further reduces human annotation by substituting AI-generated critiques guided by predefined principles [3], yet this shifts the burden to principle specification and presumes that normative criteria can be articulated a priori. Across these approaches, alignment is framed as supervised learning from observable preferences, limiting applicability in settings where preferences are implicit, emergent, or difficult to elicit.

Community Norms and Domain-Specific NLP A growing body of NLP research emphasizes the importance of cultural, social, and community-specific norms, particularly in low-resource or marginalized contexts [9, 37]. Domain adaptation and specialization techniques such as BioBERT and LegalBERT demonstrate the value of tailoring models to specific domains, but typically require substantial labeled data [6, 10, 34]. Ethical NLP work further argues for embedding social values and community perspectives into model design [7, 28, 35], yet little work has explored how such norms can be learned operationally from naturally occurring community behavior. Our approach contributes a concrete mechanism for grounding alignment in community norms by inferring them directly from patterns of acceptance, without requiring explicit annotation or predefined value schemas.

Beyond linguistic variation, work in social computing and HCI highlights that community norms are not merely emergent patterns of language use, but are actively shaped through moderation practices, governance structures, and participation asymmetries [29, 38, 40]. These dynamics raise questions of legitimacy and representation: whose behavior contributes to observable norms, and whose voices are systematically excluded. While prior NLP work has emphasized the importance of respecting community norms, relatively little research has explored how such norms can be operationally inferred from naturally occurring community behavior without relying on explicit annotation or predefined value schemas.

Implicit Behavioral Signals and (the Limits of) Revealed Preference A large body of work has explored learning from implicit behavioral signals, such as clicks, dwell time, and interaction patterns, particularly in recommender systems and information retrieval [26, 30]. These signals are attractive because they are abundant and naturally occurring, but they are also indirect: they reflect behavior mediated by platform affordances, incentives, and power rather than explicit judgments of quality or appropriateness. Prior work has shown that optimizing directly for engagement can distort model behavior, amplifying polarized, sensational, or emotionally charged content [5].

A long line of critique cautions against equating observed behavior with normative endorsement or consent, particularly in platform-mediated environments [22, 38]. In this work, we therefore treat acceptance signals as descriptive evidence of how norms are enacted in practice rather than as ethically authoritative preferences. Our goal is not to maximize engagement or infer individual utilities, but to recover community-level regularities in what is treated as acceptable within specific contexts.

Density and Geometry in Representation Space Density estimation has a long history in statistics, with classical approaches such as kernel density estimation and Gaussian mixture models providing flexible non-parametric tools [43]. Recent advances in neural density estimation enable scalable likelihood modeling in high-dimensional spaces, including autoregressive models and normalizing flows [15, 42, 45]. Separately, work on representation geometry in NLP has shown that linguistic representations occupy structured, low-dimensional manifolds in embedding space [2, 36]. However, these techniques have primarily been used for generative modeling or representation analysis, rather than for norm inference or alignment. Building on these works, our approach interprets local density in embedding space as a community-conditioned acceptance signal, using geometric structure as supervision for alignment without explicit preference labels.

3 Method

Our goal is to extract an alignment signal from naturally occurring community behavior without relying on explicit preference annotations. We build on the observation—well established in both social computing and recommender systems—that communities already express preferences implicitly through what content they accept, engage with, and allow to persist. We show that repeated community acceptance induces measurable structure in representation space, and that this structure can be operationalized as a preference-aligned signal for language model alignment.

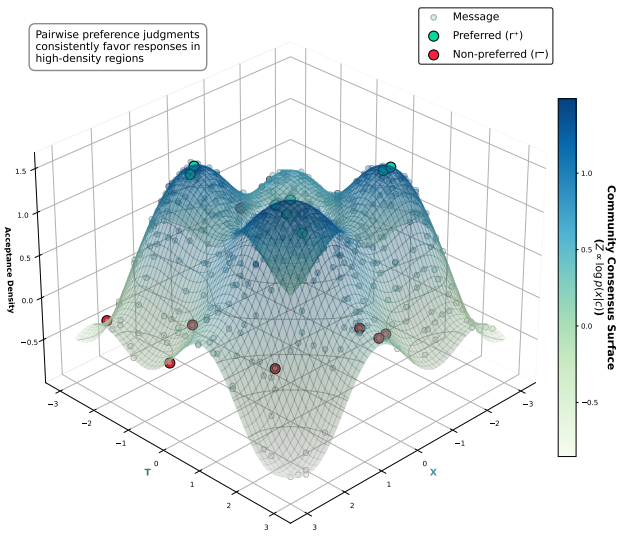

Fig. 1. **Conceptual Representation of the Community Consensus Surface.** The Z-axis represents a normative log-density, reflecting the implicit filtering of responses by community standards through moderation and collective feedback [11, 32]. High-density regions correspond to a coherent, low-dimensional manifold of accepted responses in representation space [2, 36]. The separation between preferred (r^+) and non-preferred (r^-) responses across this surface reflects an *acceptance–preference correspondence*, motivating preference learning and alignment without explicit annotation [13, 41, 44].

3.1 Conceptualization: Community Acceptance as a Manifold

Community norms are not imposed instantaneously; they emerge gradually through repeated interaction. Over time, online communities continuously filter participation through moderation, feedback, and collective attention. Responses that align with shared expectations are more likely to persist, receive engagement, and be incorporated into ongoing discourse. Responses that violate these expectations are disproportionately ignored, down-ranked, or removed.

This repeated process of selection acts as a form of implicit norm formation and expression. As similar responses are consistently accepted across comparable contexts, they accumulate and reinforce one another, giving rise to behavioral and linguistic regularities at the community level. For intuition, consider responses as points scattered across a 3D landscape, where elevation represents community acceptance density (Figure 1). Accepted responses—those that persist, receive engagement, or align with community norms—cluster in peaks of high density (high elevation), forming a coherent acceptance manifold. In contrast, rejected or misaligned responses lie in sparse, low-density regions at lower elevation, farther from the community's normative core. This geometric separation mirrors the acceptance–preference correspondence illustrated in Figure 1, where preferred (r^+) and non-preferred (r^-) responses occupy distinct regions of the surface. Prior work shows that such endogenous filtering dynamics produce durable patterns in language use, interaction style, and participation structure within communities [12, 14, 24].

We formalize this phenomenon geometrically, drawing on representation geometry [2, 39] and density-based clustering [18], which show that linguistic and semantic structures occupy low-dimensional manifolds in embedding space. For a community c, we define an *acceptance manifold* $\mathcal{M}_c$ as the region of representation space occupied by responses that the community accepts as appropriate or authentic. Note that acceptance here is not a binary property of individual messages (e.g., receiving upvotes or avoiding removal), but an aggregate notion that emerges over time from patterns of participation and persistence within the community. Let $E(r)$ denote the embedding of a response r. We model community acceptance as a density over representations,

$$p(r \mid c) = p(E(r) \mid c),$$

where higher density indicates stronger conformity with community norms. This view is consistent with distributional perspectives on language, in which semantic and pragmatic regularities correspond to geometric structure in embedding space [19, 39]. Here, however, geometry reflects not only semantic similarity, but also normative compatibility with a specific community.

The gradient of the log-density,

$$\nabla_{E(r)} \log p(E(r) \mid c),$$

defines a continuous direction of increasing alignment with community norms. Unlike discrete preference labels, this signal is smooth, shared across responses, and derived directly from observed behavior.

This framing induces an *acceptance–preference correspondence* assumption: responses that are repeatedly accepted by a community are more likely to align with that community's preferences. Formally,

$$\arg\max_r p(r \mid c) \;\propto\; \arg\max_r \mathbb{E}[\text{preference}(r \mid c)].$$

This assumption parallels foundational results in revealed preference theory and implicit feedback learning, where aggregate behavioral signals—despite being noisy at the individual level—can be used to *empirically derive stable community-level preferences and norms* [22, 25, 26, 31]. Here, we treat acceptance behavior as a revealed signal of

collective consensus: the norms that emerge from repeated, distributed decisions about what content is permitted, engaged with, persists within a community.

3.2 Problem Formulation

Let $\mathcal{D}_c = \{r_i\}_{i=1}^N$ denote a corpus of responses that have been accepted by a community c through moderation, engagement, or sustained participation. We embed each response as $x_i = E(r_i)$ and interpret their distribution in representation space as an empirical record of the community's acceptance behavior.

Our goal is to use it to *derive an implicit preference signal*. Specifically, we view local acceptance density as inducing a partial ordering over candidate responses: responses that lie in higher-density regions of the acceptance manifold are more consistent with community norms than those in low-density regions.

In standard alignment pipelines such as RLHF [13] or Direct Preference Optimization (DPO) [44], learning is driven by explicit pairwise preference annotations. In contrast, we replace this supervision with a density-derived preference signal. For a given context, candidate responses can be ranked according to their relative *acceptance density*,

$$p(E(r) \mid c),$$

which serves as a proxy for community preference in the absence of human-labeled comparisons.

We refer to this approach as **Density-Guided Response Optimization (DGRO)**. DGRO uses acceptance density to construct implicit preferred and dispreferred response pairs, enabling standard preference-based objectives such as DPO to be applied in annotation-scarce settings. This formulation aligns with prior work showing that geometric and distributional structure can substitute for direct supervision in low-resource regimes [1, 21].

3.3 Operationalizing Acceptance Density

Acceptance density is a conceptual object defined over representation space. A key design choice is whether to estimate this density globally across all community content or locally conditioned on context. A global estimate implicitly assumes that community norms are uniform across topics and intents—a strong assumption that we later show obscures preference signal. We therefore adopt a *local* density estimation strategy inspired by neighborhood-based semantic modeling and local distributional structure [27, 33], while treating global density estimation as a baseline.

Given a query context h (e.g., a conversation history or post topic) with embedding $E(h)$, we define a context-conditioned reference set

$$\mathcal{B}(h) = \mathrm{kNN}(h; \{E(h_i)\}_{i=1}^N),$$

consisting of the k nearest contexts. Let $\{x_j\}_{j \in \mathcal{B}(h)}$ denote the embeddings of the corresponding accepted responses.

We estimate acceptance density using a kernel density estimator,

$$\log p(x \mid h, c) \propto \log \frac{1}{|\mathcal{B}(h)|} \sum_{j \in \mathcal{B}(h)} K_\sigma(x, x_j),$$

where K_σ is an RBF kernel with bandwidth set via the median heuristic. This gives us a context-sensitive estimate: responses are scored relative to what the community accepts in similar situations, rather than against an aggregated global pool.

If acceptance density reflects community preference structure, it should both correlate with labeled human preference behavior in supervised settings and serve as a practical substitute for explicit preference annotations when used to train

alignment objectives such as DPO or RLHF. We evaluate both implications empirically in Section 5 before deploying DGRO in annotation-scarce domains.

4 Experimental Setup

Our experiments are structured to answer three progressively stronger questions. First, we validate the *manifold hypothesis*: whether community preference signals exhibit local geometric structure in representation space. Next, we test whether *acceptance density* can *functionally replace* explicit human preference labels inside a standard optimization objective. Finally, we evaluate whether this signal can be used to align language models in real-world communities where preference annotations are unavailable.

4.1 Validating the Manifold Hypothesis

First, we seek to validate the core premise of our approach: that preference signal exhibits *local geometric structure* in representation space. We use the Stanford Human Preferences (SHP) benchmark [20], which provides pairwise preference judgments across Reddit communities as well as an external quality signal measuring the strength of human agreement.

Communities and Data. We select five subreddits with clearly distinct moderation regimes and community norms: *changemyview*, *askculinary*, *askhistorians*, *legaladvice*, and *explainlikeimfive*; these communities spanning different domains, interaction styles, and standards for acceptable responses. These communities differ substantially in how responses are evaluated, filtered, and endorsed, providing a controlled setting to test whether preference structure is shared across heterogeneous norms rather than driven by idiosyncrasies of a single community. Additional details about each community are provided in Appendix Table 4. Each example consists of a conversation history (prompt), a preferred response and a non-preferred response as determined by community member voting, along with metadata including the normalized ratio of upvotes between responses, which captures preference strength.

Testing the Manifold Hypothesis. We ask whether responses preferred by a community tend to occupy higher-density regions of representation space than non-preferred responses, when density is estimated using only unlabeled data. To test this, we first embed all responses from the training split, treating them as an unlabeled reference pool that includes both preferred and non-preferred responses. We use a fixed sentence encoder to obtain representations, enabling density estimation over the resulting embedding space.[1] Preference information is not used at this stage, and training and test splits are kept strictly disjoint. For each prompt in the test set, candidate responses are embedded and ranked according to their acceptance density under the community distribution. We then evaluate whether the response with higher estimated density corresponds to the community-preferred response.

Evaluation Protocol. For each test pair (h, r_+, r_-), we compute a margin given by the difference in estimated acceptance density between r_+ and r_-. We report pairwise accuracy, $\mathbb{P}[\text{margin} > 0]$, as the primary metric. SHP provides the ratio of upvotes between responses normalized as an independent measure of community agreement strength. If preference signal is encoded in local geometry, our density-based margins should align with human preferences and improve as community agreement increases.

Baseline Methods. Our model, which we call acceptance density, estimates density conditioned on the $k = 150$ nearest histories in embedding space; performance is robust to k and we report ablations in Appendix D.

[1] We use the `sentence-transformers/all-mpnet-base-v2` encoder (https://huggingface.co/sentence-transformers/all-mpnet-base-v2), a widely adopted semantic model that provides stable neighborhood structure across domains. Results are robust to alternative encoders; see Appendix B.

We compare against the following baselines. (1) Random assigns random margins as a sanity check. (2) k-Nearest Neighbors (kNN) retrieves the $k = 150$ most similar training histories and predicts the majority preference label, testing whether neighborhood selection alone provides signal. (3) Global acceptance density estimates acceptance density using a fixed random subset ($|G| = 1000$) of training responses, testing whether density modeling without locality recovers preference structure. Finally, we report results for the (4) original supervised SHP reward model [2]. This model serves as an upper-bound reference, illustrating how closely density-based methods trained on unlabeled data approximate preference signals learned from large-scale human annotations.

4.2 Acceptance Density as a Preference Proxy

Building on the validation in the previous section, we next test whether acceptance density can replace human-labeled comparisons within a standard preference optimization pipeline, and whether doing so induces preference behavior aligned with community judgments.

To test this, we instantiate a density-based variant of Direct Preference Optimization (DPO) that uses acceptance density to construct implicit preference supervision. We follow the same procedure for estimating acceptance density described in the previous section, treating the training split as an unlabeled reference pool and never using ground-truth preference labels during training. Density-derived rankings are used to form implicit preferred and dispreferred response pairs, which are then used to train a policy model with the standard DPO objective.

Unless otherwise specified, all main results initialize from a pre-trained Pythia-2.8B language model. This choice mirrors the experimental setup used in prior DPO work [44], which uses Pythia-2.8B [8] as a primary reference architecture for preference optimization; we do this for direct comparability and to isolate the effects of the preference signal rather than architectural differences. Evaluation is performed on a held-out test split.

To assess robustness, we additionally repeat this procedure across multiple model architectures and parameter scales. These results show consistent trends, and we report deviations from the Pythia-2.8B baseline in Appendix C.

Evaluation protocol. Evaluation is performed against *held-out ground-truth human preferences*. We assess alignment using length-normalized preference accuracy, defined as the fraction of held-out SHP pairs for which the model assigns higher average log-probability per response token to the human-preferred answer. Log-probabilities are computed over response tokens only, conditioned on the shared prompt, ensuring that differences in response length do not confound the comparison. This evaluation directly tests whether optimization driven solely by acceptance density induces models to prefer the same responses that human annotators judge as better, which is the central objective of preference-based alignment. We report this metric for both supervised DPO (trained on true human preference pairs) and acceptance density-guided DPO under identical architectures, prompts, and evaluation conditions. This isolates a fundamental question: whether acceptance density behaves *like a preference signal* when used as the sole source of supervision inside a standard alignment objective. Demonstrating competitive performance under this constraint establishes acceptance density as a viable substitute for explicit preference labels, justifying its use in annotation-scarce domains for alignment purposes.

4.3 Application to Annotation-Scarce Communities

Following the validation experiments above, we apply density-guided response optimization (DGRO) in real-world communities where explicit preference annotations are unavailable, and evaluate its effectiveness for aligning language

[2] https://huggingface.co/stanfordnlp/SteamSHP-flan-t5-xl

Table 1. Communities and data sources used in DGRO evaluation. Validation communities provide explicit preference supervision, while application communities lack pairwise labels and rely on behavioral acceptance signals.

Community	Platform	Scale	Acceptance Signal
Q&A	Reddit (SHP)	10K–50K pairs	Pairwise human preferences
Eating Disorder Support	Twitter	~43K posts	Replies, retweets
Eating Disorder Support	Reddit	~9.2M posts	Upvotes, comment depth
Eating Disorder Support	Forums	~1.6M posts	Replies, thread continuation
Conflict Documentation	VK	~8.34M posts	Likes, reposts

models in practice. In these settings, acceptance density is used to construct implicit preference supervision. Using unlabeled community data, we estimate acceptance density as described in Section 4.1 and use it to form preferred and dispreferred response pairs. These density-derived pairs are then used to train policy models with a standard DPO objective. No explicit pairwise preference annotations are used at any stage.

Communities and data. We evaluate DGRO in settings where general-purpose models fail to capture domain-specific norms, and where standard preference annotation methods pose significant ethical risks.

Our primary evaluation focuses on eating disorder support communities across three platforms (Reddit, Twitter, and specialized forums). These communities exhibit highly sensitive, context-dependent communication norms distinct from general instruction-following behavior. Prior work indicates that off-the-shelf LMs often generate content that members find inauthentic or harmful [23, 47, 49]. To address the ethical challenges of working in this domain, our data curation was conducted in collaboration with clinical domain experts and medical professionals as part of a broader study on online community formation (with IRB approval). Using expert-verified implicit signals avoids the ethical pitfalls of explicit annotation, including consent issues and potential re-traumatization.

To validate cross-lingual and political discourse generalization, we extend our evaluation to conflict documentation communities on VKontakte (VK), a Russian-language platform structurally comparable to Facebook [4]. These communities focus on the aggregation and discussion of ongoing conflict documentation, exhibiting norms distinct from both Western platforms and general Russian-language corpora. Current multilingual models, typically trained on broad web corpora, lack exposure to these specific discourse conventions. Using these data, we test DGRO's ability to adapt to distinct sociopolitical dialects where standard models often produce responses that appear foreign to the community's authentic communication patterns.

Evaluation protocol. The goal of this evaluation is to assess whether density-guided response optimization produces outputs that are judged as more appropriate and authentic within communities where explicit preference annotations are unavailable. As established in earlier sections, this analysis rests on two validated prerequisites: first, that acceptance density reliably recovers human pairwise preferences when such labels are available (Section 4.1); and second, that density-guided optimization induces model behavior aligned with those same human judgments on held-out data (Section 4.2). Having validated both the preference signal and its effect on model behavior, we now evaluate aligned models in annotation-scarce domains.

Because these domains lack large-scale preference annotations, evaluation must rely on indirect judgments. We therefore anchor LLM-based evaluation in human expert assessment, first conducting expert evaluation on a stratified subset of 200 held-out examples (50 per domain), with three domain experts per community. This analysis verifies that aggregate LLM judgments track expert assessments along the same criteria. Following established practices in alignment

research [17, 48], we then use LLM-as-judge comparisons along two criteria—relevance (contextual appropriateness to the prompt and community norms) and authenticity (consistency with the community's characteristic tone, framing, and interactional style)—as a scaling mechanism for this previously validated human preference structure, rather than as an independent source of normative authority. Evaluation is performed in a head-to-head setting, where judges compare a model-generated response against an actual response drawn from the target community for the same context, using examples held out from all training stages. We use three frontier language models as judges: GPT-5-nano, Claude-4.5-Haiku, and Gemini-2.5-Flash.[3] Each model is queried three times with randomized response order to control for positional bias, yielding nine judgments per comparison.

Baselines and model variants. We compare DGRO against three baselines: (1) an off-the-shelf instruction-tuned model (Base), (2) supervised fine-tuning on community text (SFT), and (3) in-context learning with community exemplars (ICL). To isolate density-guided optimization from supervised pre-training effects, we conduct ablations controlling for training compute.

All comparisons use identical architectures, decoding parameters, and context construction. As explored in prior sections and further examined in Appendix C, variation across model architectures and scales appears limited for preference alignment under density-guided DPO. As such, we fix the base model to Pythia-2.8B in this section in order to focus on the behavioral and normative effects of the alignment procedure itself, rather than introducing additional variation from differences in model capacity or representation.

5 Results

Table 2. Pairwise accuracy across communities for unsupervised and supervised methods. Accuracy is reported as mean ± bootstrap half-width, $\delta = \frac{1}{2}(\text{hi} - \text{lo})$, computed independently per subreddit. Supervised Model (RM) denotes the supervised reward model (stanfordnlp/SteamSHP-flan-t5-xl), trained with human preference annotations and included as a reference upper bound.

Method	r/askhr	r/askbaking	r/askculinary	r/askhistorians	r/changemyview	r/asksocialscience	r/asksciencefiction
Random	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
kNN	0.55 ± 0.03	0.49 ± 0.01	0.50 ± 0.02	0.58 ± 0.03	0.49 ± 0.03	0.50 ± 0.03	0.52 ± 0.04
Global Acceptance Density	0.68 ± 0.01	0.53 ± 0.03	0.51 ± 0.03	0.60 ± 0.09	0.57 ± 0.04	0.59 ± 0.03	0.49 ± 0.03
Local Acceptance Density	0.71 ± 0.03	0.60 ± 0.02	0.57 ± 0.04	0.72 ± 0.03	0.61 ± 0.03	0.64 ± 0.01	0.65 ± 0.02
Supervised Model (RM)	0.75 ± 0.03	0.65 ± 0.03	0.72 ± 0.01	0.74 ± 0.02	0.68 ± 0.02	0.80 ± 0.03	0.72 ± 0.02

5.1 Validating the Manifold Hypothesis

We begin by evaluating the central empirical claim of this work: that preference signal is encoded in the local geometry of representation space (acceptance density). If this hypothesis holds, preserving local manifold structure should recover human preferences, while methods that destroy or ignore locality should fail.

Preference signal is recovered by geometry-preserving density. We find that preference signals, typically requiring explicit supervision, can be recovered through the geometry-preserving properties of local density. As shown in Table 2 and Figure 4, local acceptance density consistently identifies community-preferred responses across all evaluated subreddits, achieving 58–72% pairwise accuracy and substantially outperforming all unsupervised baselines.

Our results suggest that recovering this structure requires a balance between locality and distributional modeling. At one extreme, global acceptance density performs near chance; by aggregating across heterogeneous contexts, it likely

[3] GPT-5-nano (https://platform.openai.com/docs/models), Claude-4.5-Haiku (https://www.anthropic.com/claude), and Gemini-2.5-Flash (https://ai.google.dev/gemini-api/docs/models)

averages away the nuanced structures that encode specific preferences. At the other extreme, simple kNN retrieval gives only modest gains above random chance, indicating that merely identifying nearby examples is insufficient: one must model the relative distribution (i.e. the "shape") of those examples.

Notably, local density approaches the performance of supervised reward models despite having no access to explicit preference labels. We find that the performance gap between our unsupervised method and supervised models narrows significantly in instances of high human agreement (Figure 4). This suggests that a substantial portion of the signal leveraged by traditional reward models is not "new" information provided by labels, but is instead already latent within the local manifold geometry of community-accepted discourse.

Additionally, we find a clear positive relationship between human agreement strength and preference recovery by local acceptance density. When aggregating across communities, accuracy exhibits a moderate, statistically robust correlation with agreement strength ($\rho_s = 0.48$, $p < 10^{-4}$), indicating that density-guided alignment performs best in regions where community preferences are most clearly differentiated.

This trend is even more pronounced within several individual communities. Subreddits such as `r/asksciencefiction` ($\rho_s = 0.90$, $p < 0.001$), `r/askhr` ($\rho_s = 0.81$, $p = 0.015$), and `r/askbaking` and `r/askculinary` (both $\rho_s = 0.75$, $p < 0.05$) exhibit strong, statistically significant correlations, suggesting that local acceptance density closely tracks human consensus when norms are well-defined. In contrast, communities with smaller evaluation sets and sparser agreement bins (e.g., `r/askhistorians`, `r/asksocialscience`) show weaker correlations, consistent with limited statistical power (rather than a deviation from the overall monotonic trend). Figure 5 and Table 8 (Appendix) provide the full per-community breakdown.

This pattern provides direct empirical support for the acceptance–preference correspondence posited in Section 3. When community agreement is weak, acceptable responses span broader and less differentiated regions of representation space, limiting the recoverability of preference signal. As consensus strengthens, accepted responses collapse into tighter, more coherent regions of the manifold, making relative density an increasingly reliable indicator of preference. Accordingly, accuracy improves systematically with human agreement strength, with local acceptance density performing best precisely when community preference is most clearly expressed. This dependence on agreement strength is inconsistent with a fixed estimator bias: if density merely favored certain responses irrespective of context, accuracy would not vary predictably with consensus. Instead, the observed relationship indicates that local geometry captures meaningful structure in community judgment rather than an artifact of density estimation.

5.2 Acceptance Density as a Preference Proxy

Having established that acceptance density behaves like a preference signal when preference is observable, we now evaluate a stronger claim: whether this signal can functionally approximate or replace explicit human preference labels inside a standard alignment objective.

DGRO recovers supervised preference structure. As shown in Figure 2, constructing preference pairs from relative position on the acceptance manifold is sufficient to induce preference behavior aligned with community judgments. Across all evaluated communities, models trained using acceptance density-derived pseudo-pairs recover a substantial fraction of the accuracy achieved by fully supervised DPO, despite having no access to human-labeled comparisons during training.

These results indicate that acceptance density functions as a usable preference signal when integrated into a standard alignment pipeline, inducing models to prefer responses that align with community judgments. Combined with the

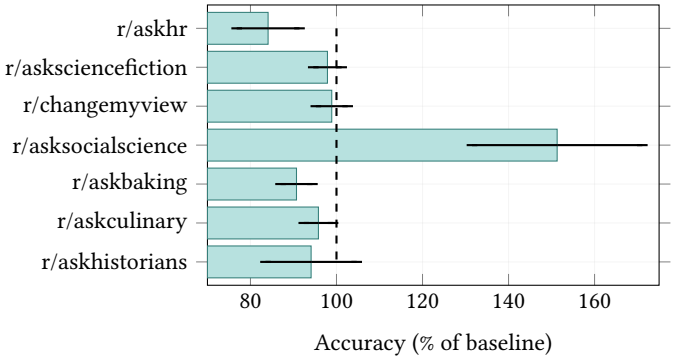

Fig. 2. Relative accuracy of DRGO-aligned models expressed as a percentage of baseline DPO performance, computed as 100 × (DRGO/baseline), where 100% denotes parity with the baseline. Error bars denote ±1 standard error estimated via bootstrap resampling (n=500), with uncertainty propagated using a first-order delta method.

validation results in Section 4.1, this supports the use of acceptance density as a practical substitute for explicit preference supervision.

5.3 Application to Annotation-Scarce Communities.

Having shown that acceptance density recovers human preference structure and can substitute for labeled comparisons in controlled settings, we next examine its utility in real-world communities where explicit preference supervision is completely unavailable. In these domains, alignment must rely on naturally occurring acceptance signals rather than curated annotations, making them a direct test of whether density-guided preference learning provides practical advantages over standard adaptation methods. Before comparing alignment methods, we verify that LLM-based judgments reflect human preference in these domains. On a stratified subset of 200 held-out examples, aggregated LLM-judge rankings correlate strongly with human expert preferences (explored further in Appendix H), supporting their use for large-scale evaluation.

DGRO consistently outperforms baselines. Illustrated in Table 3, across all domains, DGRO-based alignment achieves consistent gains over baselines despite using the same underlying training data. For example, on ED-Reddit, DGRO wins 58.8% of head-to-head comparisons against SFT ($p < 0.001$). Similar patterns emerge across other contexts, where DGRO maintains a significant advantage over SFT in direct comparisons.

The quantitative advantage of DGRO over baselines is reflected in qualitative differences in response authenticity. Table 10 presents representative examples from both ED-Reddit and VK Conflict discourse, comparing model outputs against real community responses for the same context. Across domains, the Base and ICL baselines frequently default to generic, non-situated language that lacks the tone, specificity, or interactional norms characteristic of the target communities. Supervised fine-tuning (SFT) improves topical relevance but often exhibits repetitive phrasing and diffuse affect, suggesting partial adaptation to surface content without internalizing community-specific modes of expression. In contrast, DGRO outputs more closely resemble authentic community participation, showing locally appropriate framing, specificity, and rhetorical structure.

These results demonstrate that density-guided optimization captures preference structure beyond what supervised fine-tuning alone recovers. While SFT adapts models to community vocabulary and style, DGRO's manifold-based

Table 3. LLM-as-judge head-to-head comparison of DGRO against baseline alignment approaches across annotation-scarce communities. Judges compare paired model outputs for the same prompt, using real community responses as contextual grounding for relevance and authenticity. Win rates indicate the percentage of comparisons in which DGRO is preferred (mean ± 95% CI).

Community	DGRO vs Base	DGRO vs ICL	DGRO vs SFT
ED-Reddit	75.4 ± 2.9%	65.8 ± 3.1%	53.8 ± 3.1%
ED-Forum	72.2 ± 3.2%	64.1 ± 4.4%	57.6 ± 3.3%
ED-Twitter	76.1 ± 3.0%	66.3 ± 4.1%	56.9 ± 2.6%
VK State	80.7 ± 3.1%	59.9 ± 3.2%	55.3 ± 2.0%

objective appears to encode finer-grained distinctions about what makes responses sound authentic within specific contexts.

6 Analysis

6.1 Manifold Structure and Preference Signal

Preference signal is encoded in local manifold structure. Across communities, acceptance density corresponds reliably with human preference when estimated locally in representation space. Conditioning density on nearby contexts preserves preference structure that is obscured by global aggregation, which collapses heterogeneous situations into a single distribution. This dependence on locality is likely not incidental. Preference signal degrades when density is estimated over neighborhoods that are either too broad—approaching global behavior—or too narrow to provide stable estimation. The resulting pattern indicates that community preferences are neither uniform nor purely instance-specific, but organized at an intermediate, context-dependent scale.

Acceptance density is data-efficient. As shown in Table 7, estimation of community preference via acceptance density approaches peak performance with relatively little training data, with the required amount varying by community. Across all communities, the normalized area under the saturation curve (AUSC) exceeds 0.91, indicating that preference structure can be recovered in a sample-efficient manner.

6.2 Failure Modes and Limitations

While DGRO provides a useful preference signal in many settings, its effectiveness depends on the availability of meaningful acceptance structure in representation space. When this structure is weak or absent, the density-derived signal can become unreliable.

Uninformative density in sparse manifold regions. DGRO relies on acceptance density to construct pseudo-preference pairs during training. When candidate responses lie far from the acceptance manifold, local density estimates become noisy and provide little discriminative signal. In such cases, pseudo-pairs may reflect superficial semantic proximity rather than contextual appropriateness. We explore an example case in Appendix K.

Amplification of community biases. By design, DGRO reproduces patterns present in community acceptance data, including harmful norms or misinformation. In polarized or toxic communities, the resulting preference signal reflects those same biases. Because DGRO derives preference structure empirically from observed acceptance behavior, it does not impose external normative constraints during training. Thus, norm correction must occur outside the preference signal itself, for example through data filtering or post-hoc safety interventions. Future work could explore hybrid approaches combining density-guided learning with external normative constraints

DGRO is not suitable as a general-purpose or platform-wide alignment mechanism. Because acceptance density reflects existing participation dynamics and power asymmetries, applying DGRO at scale risks entrenching dominant norms, amplifying coordinated manipulation, and obscuring contestation. Without explicit governance, community consent, and mechanisms for redress, density-guided optimization should be treated as an analytical instrument rather than a deployment-ready alignment strategy.

These limitations suggest clear boundaries: DGRO is best suited to stable communities with established norms, sufficient scale for density estimation, and values aligned with deployment objectives. When communities are small, polarized, rapidly evolving, or exhibit harmful norms, explicit human supervision remains necessary.

7 Discussion

Language models increasingly operate in settings where communicative norms are community-specific and diverge from generic instruction-following behavior. Our results suggest that these norms give rise to stable, community-level structure in representation space, which can be captured through acceptance density. This structure reflects not only semantic similarity, but alignment with what a community considers appropriate.

DGRO operationalizes this observation by using acceptance density as a source of preference supervision. Rather than relying on elicited pairwise judgments, the method constructs preference signal directly from unlabeled community behavior. Across the settings we study, this signal is sufficient to guide alignment in domains where explicit preference annotations are impractical, costly, or ethically constrained.

7.1 Ethical Considerations

While DGRO uses only publicly observable signals, the method raises ethical concerns warranting careful consideration before deployment. The question of who speaks for a community becomes important. Acceptance patterns reflect active participants, moderators, and platform affordances, which may not represent full community values. Marginalized voices, silent lurkers, or departed members do not contribute to the signal, yet deployment affects them. DGRO-based alignment uses revealed preferences of those who remain, potentially encoding values of whoever holds power rather than the community as a whole.

Additionally, harm amplification poses a serious risk. Because DGRO derives preference structure directly from observed community behavior, it reproduces existing norms, including harmful or exclusionary ones. Unlike supervised alignment, it does not introduce an external mechanism for norm correction during training; mitigation must therefore rely on data filtering or post-hoc constraints. Vulnerability to manipulation creates additional concerns. Adversaries who can influence acceptance through coordinated engagement or vote manipulation can poison the learned preference structure. This is particularly concerning in communities with weak integrity controls or concentrated power.

DGRO deployment requires careful ethical assessment beyond technical validation. At minimum: transparency about community data use, mechanisms for feedback and opt-out where feasible, ongoing monitoring for drift, and human oversight in high-stakes domains. For sensitive communities like mental health forums, stakeholder consultation should precede deployment. The broader question is whether making alignment more accessible ultimately serves community interests. Reducing barriers could empower under-resourced communities to shape AI behavior appropriately, or empower exploitation of community data and amplification of harmful norms. These questions require ongoing dialogue between researchers, communities, and stakeholders about appropriate governance.

8 Conclusion

We introduce density-guided response optimization (DGRO), a method for aligning language models to community norms without relying on explicit preference annotations. By modeling the distribution of responses that communities consistently accept, DGRO infers implicit preference structure from local density in representation space.

Across validation experiments, models aligned using DGRO outperform baseline approaches despite having no access to human-labeled preference comparisons during training, relying only on naturally occurring community behavior. These results indicate that acceptance signals encode sufficient structure to support preference-based alignment.

Our findings suggest that community acceptance provides a practical, annotation-free source of alignment signal, enabling model adaptation in settings where explicit preference elicitation is infeasible, costly, or ethically constrained.

9 Endmatter Sections

9.1 Generative AI Usage Statement

The authors did not use generative AI tools for this manuscript. The authors wrote and prepared all of the content for this manuscript.

9.2 Ethical Considerations Statement

This work uses publicly available data drawn from online communities and does not involve direct interaction with human subjects, intervention in deployed systems, or the collection of private or non-public information. All data were handled in accordance with applicable platform terms and established norms for CSS research. We did not attempt to identify individuals, and our analysis was conducted at an aggregate level focused on community-wide patterns.

The primary ethical risks associated with this work come from the potential downstream use of DGRO to model and reproduce community norms. These risks are discussed in detail in Section 7.1. In that section and here, we emphasize that acceptance-based signals reflect the behavior of active and empowered participants rather than comprehensive or consensual community values. Additionally, we note that DGRO should not be treated as a normative authority or deployed without appropriate oversight.

We do not claim that DGRO mitigates harmful norms or resolves questions of legitimacy. Instead, we treat it as a descriptive method whose responsible use depends on transparency, community governance, and domain-specific safeguards. Potential adverse impacts and limitations are analyzed in Section 7.1, and we outline conditions under which deployment would be inappropriate or ethically unsafe.

References

[1] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics* 4 (2016), 385–399.

[2] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics* 6 (2018), 483–495.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[4] KS Baran and WG Stock. 2015. Facebook has been smacked down. The Russian special way of SNSs: Vkontakte as a case study. In *Proceedings of the 2nd European conference on social media (ECSM 2015)*, Vol. 9. 574–582.

[5] Pablo Barberá. 2020. Social media, echo chambers, and political polarization. *Social media and democracy: The state of the field, prospects for reform* (2020), 34–55.

[6] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1 (2010), 151–175.

[7] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[8] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*. PMLR, 2397–2430.

[9] Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*. 3504–3519.

[10] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).

[11] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.

[12] 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. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[13] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[14] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. 307–318.

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803* (2016).

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[17] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475* (2024).

[18] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96. 226–231.

[19] Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512* (2019).

[20] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with V-usable information. In *International Conference on Machine Learning*. PMLR, 5988–6008.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The journal of machine learning research* 13, 1 (2012), 723–773.

[22] D Wade Hands. 2014. Paul Samuelson and revealed preference theory. *History of political economy* 46, 1 (2014), 85–116.

[23] Zihao He, Minh Duc Chu, Rebecca Dorn, Siyi Guo, and Kristina Lerman. 2024. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. *arXiv preprint arXiv:2406.12074* (2024).

[24] Benjamin D Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the social signals that drive online discussions: A case study of reddit communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–9.

[25] Hendrik S Houthakker. 1950. Revealed preference and the utility function. *Economica* 17, 66 (1950), 159–174.

[26] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*. Ieee, 263–272.

[27] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long papers)*. 873–882.

[28] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813* (2020).

[29] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[30] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007), 7–es.

[31] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*. 781–789.

[32] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[33] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.

[34] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[35] Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL workshop on ethics in natural language processing*. 30–40.

[36] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864* (2020).

[37] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics* 13 (2025), 652–689.

[38] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[40] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action.* Cambridge university press.

[41] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[42] George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked autoregressive flow for density estimation. *Advances in neural information processing systems* 30 (2017).

[43] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33, 3 (1962), 1065–1076.

[44] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[45] Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International conference on machine learning*. PMLR, 1530–1538.

[46] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

[47] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one* 15, 12 (2020), e0243300.

[48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.

[49] Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008* (2023).

A Full Dataset Info – Reddit

Table 4. Dataset sizes for Reddit communities used in evaluation.

Subreddit	Train	Validation	Test	Total
r/askhr	8,295	641	395	9,331
r/askbaking	44,007	2,096	1,544	47,647
r/askculinary	45,710	2,094	2,563	50,367
r/askhistorians	3,264	113	164	3,541
r/changemyview	38,173	1,637	1,836	41,646
r/asksocialscience	2,706	147	188	3,041
r/askscience fiction	29,382	1,576	1,987	32,945

B Embeddings

Table 5. Effect of embedding model choice on local acceptance density performance. Accuracy is reported as mean ± bootstrap half-width, $\delta = \frac{1}{2}(\text{hi} - \text{lo})$, computed independently per subreddit. Results are shown for the local density method using different sentence embedding models to construct the acceptance manifold.

Embedding Model	r/askhr	r/askbaking	r/askculinary	r/askhistorians	r/changemyview	r/asksocialscience	r/askscience fiction
MPNet (default)	0.71 ± 0.03	0.60 ± 0.02	0.57 ± 0.04	0.72 ± 0.03	0.61 ± 0.03	0.64 ± 0.01	0.65 ± 0.02
all-MiniLM-L6-v2	0.70 ± 0.03	0.59 ± 0.02	0.56 ± 0.04	0.70 ± 0.04	0.60 ± 0.03	0.63 ± 0.02	0.64 ± 0.02
E5-large-v2	0.72 ± 0.03	0.61 ± 0.02	0.58 ± 0.04	0.73 ± 0.03	0.62 ± 0.03	0.65 ± 0.02	0.66 ± 0.02

C Model Robustness

Table 6. Deviation in length-normalized preference accuracy on held-out SHP human preference pairs relative to the Pythia-2.8B baseline. Reported values indicate mean difference (in percentage points) ± bootstrap standard error, computed under identical prompts, objectives, and evaluation conditions. Deviations are small across base models, indicating that acceptance density-guided DPO induces consistent preference alignment behavior largely independent of model architecture, which is consistent with prior work [44].

Base Model	Preference Accuracy Δ (pp)
google/gemma-2b [46]	−0.4 ± 0.4
google/gemma-7b [46]	+0.3 ± 0.5
meta-llama/Llama-3.2-3B [16]	+0.1 ± 0.4
meta-llama/Llama-3.1-8B [16]	+0.6 ± 0.6

D K Robustness

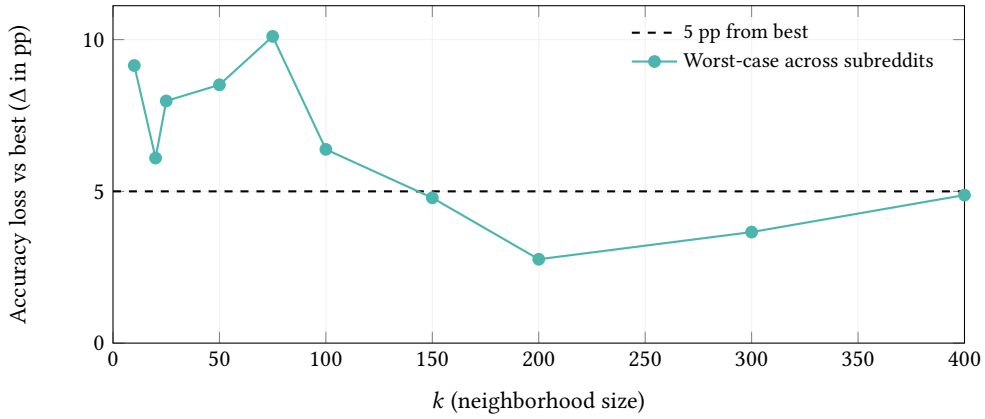

Fig. 3. Local accuracy saturates quickly with neighborhood size. Shown is the worst-case absolute accuracy loss across communities relative to each community's best-performing neighborhood size, demonstrating that performance remains within a few percentage points of optimal across a wide range of k.

E Accuracy Of Unsupervised Models – Visualized

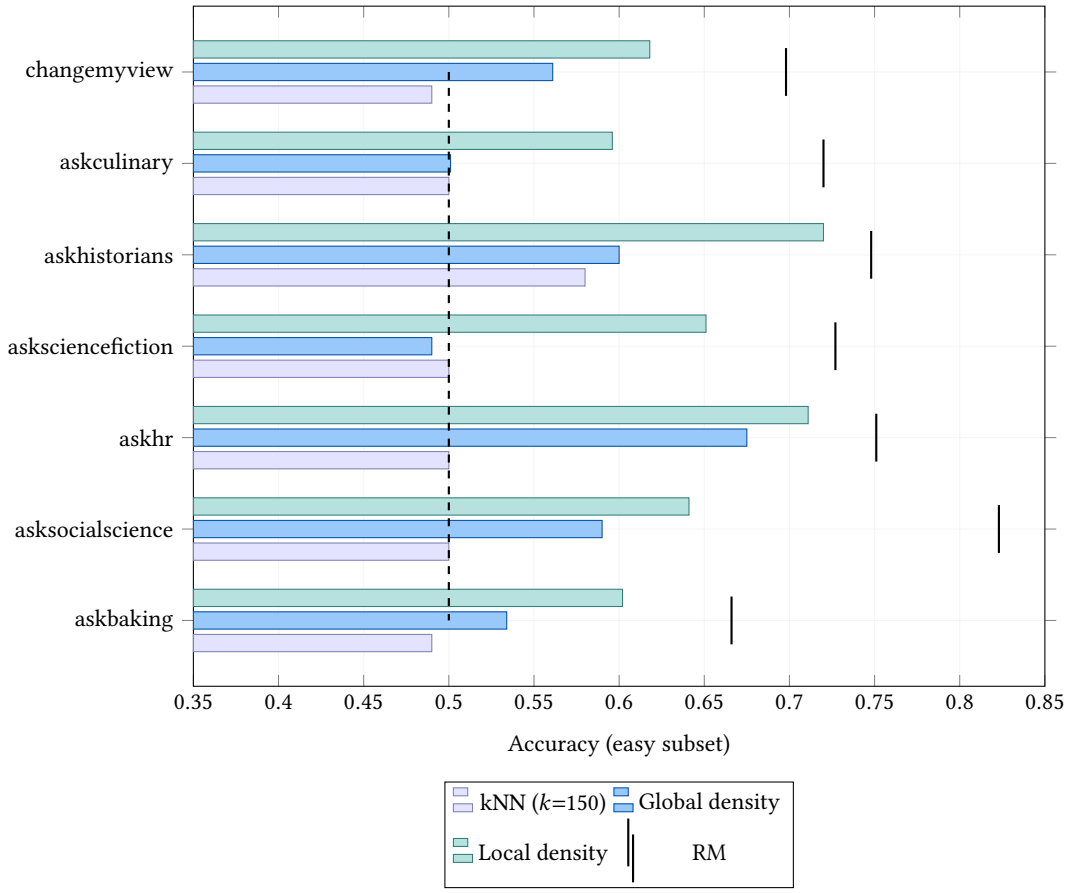

Fig. 4. Accuracy across communities. Bars show kNN, global density, and local density baselines, evaluated on the easy subset of examples. Vertical ticks denote supervised reward model (RM) accuracy. The dashed vertical line at 0.50 marks random-chance performance.

F Data Efficiency

Table 7. Data efficiency of the local method across communities. We report the normalized area under the saturation curve (AUSC) and the number of training pairs required to reach 95% of peak accuracy, both computed using accuracy expressed as a percentage of each method's peak performance. Higher AUSC and lower pair counts indicate faster saturation under limited supervision.

Subreddit	AUSC	Pairs to 95% peak
r/askhr	0.971	50
r/askbaking	0.985	150
r/askculinary	0.981	250
r/askhistorians	0.920	1450
r/changemyview	0.978	250
r/asksocialscience	0.950	250
r/asksciencefiction	0.961	850

G Correlation with Human Agreement

Table 8. Per-subreddit correlations between human agreement strength and local accuracy. For each subreddit, we bin comment pairs by agreement strength (median score_ratio per bin) and compute local pairwise accuracy within each bin. We then assess the monotonic relationship between bin-level agreement strength and bin-level accuracy using Spearman's ρ. Five of seven subreddits show significant positive correlations ($p < 0.05$), with particularly strong effects in r/asksciencefiction ($\rho_s = 0.90$) and r/askhr ($\rho_s = 0.81$). Asterisks denote significance levels: $^{*}p < 0.05$, $^{***}p < 0.001$.

Subreddit	ρ_s	p**-value**
r/askhr	0.81	0.015*
r/askbaking	0.75	0.013*
r/askculinary	0.75	0.020*
r/askhistorians	0.45	0.197
r/changemyview	0.60	0.067
r/asksocialscience	0.26	0.500
r/asksciencefiction	0.90	<0.001***

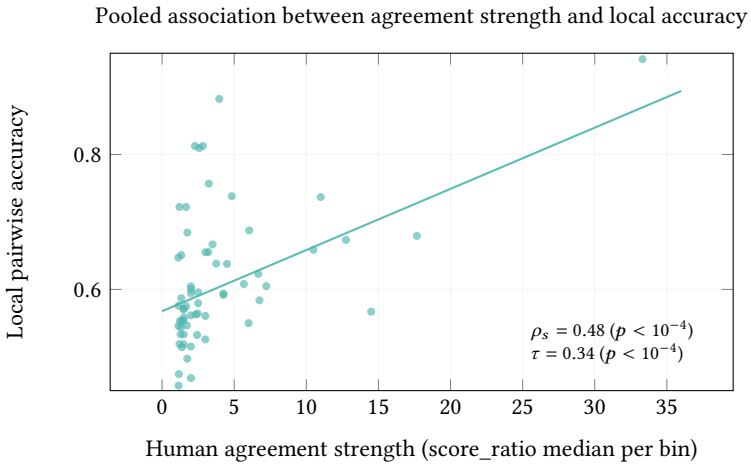

Fig. 5. **Higher human agreement correlates with higher local accuracy.** Each point is an agreement-strength bin from a subreddit. The moderately strong positive correlation ($\rho_s = 0.48$, $p < 10^{-4}$) suggests that judge accuracy improves in regions where community preferences are more clearly differentiated. The fitted line is shown for visualization only; significance is assessed with rank correlations.

H Reliability of Human and LLM-Based Evaluation

To understand the reliability of LLM-based evaluation in annotation-scarce domains, we conducted human expert evaluation on a stratified subset of 200 held-out examples (50 per domain), with three domain experts per community. Experts were evaluated under the same head-to-head comparison setup used for LLM-based evaluation in Section 4.3: for each example, experts compared a model-generated response against an actual response drawn from the target community for the same context. Experts judged responses along the criteria of relevance (contextual appropriateness to the prompt and community norms) and authenticity (consistency with the community's characteristic tone, framing, and interactional style), and were asked to make comparative judgments. All examples were held out from training at every stage.

We compute inter-annotator agreement using Krippendorff's α with an ordinal distance function. Krippendorff's α is appropriate for this setting because it supports ordered categories, multiple annotators, and chance correction. Then, to evaluate whether LLM-based evaluation reproduces expert judgment structure, we compute Spearman rank correlation between aggregate expert rankings and aggregate LLM rankings on the same examples. We also treat the expert majority decision (2-of-3 agreement) as a reference label and measure LLM agreement with this majority outcome, effectively treating the LLM ensemble as an additional annotator.

Table 9. Reliability of human expert and LLM-based evaluation on a stratified subset of 200 examples (50 per domain). Inter-annotator agreement is measured using Krippendorff's α. Expert–LLM alignment is measured using Spearman rank correlation (ρ). LLM agreement with expert majority indicates the fraction of cases in which the aggregate LLM judgment matches the expert majority ranking.

Community	**Expert IAA** (α)	**Expert–LLM** (ρ)	**LLM vs Expert Majority** (%)
ED-Reddit	0.46	0.50	77.9
ED-Forum	0.42	0.48	74.6
ED-Twitter	0.44	0.52	79.1
VK State	0.51	0.55	82.0
Aggregate	**0.46**	**0.51**	**78.4**

Across domains, expert inter-annotator agreement is moderate, reflecting the inherently holistic and context-dependent nature of judgments about relevance and authenticity. Expert–LLM rank correlations are comparable in magnitude to expert–expert agreement, and aggregate LLM judgments match the expert majority decision in approximately 78% of cases. This indicates LLM-based evaluation reproduces the same underlying preference structure expressed by human experts. Practically speaking, aggregated LLM judgments are no less reliable than an additional human annotator operating under the same task constraints. With this in mind, LLM-as-judge is used in this work solely as a scalable extension of expert-grounded evaluation given its validation against human judgment.

I Qualitative Examples

Table 10. Qualitative examples comparing model variants against real community responses.

Community	Context	Real	Base / ICL	SFT	DGRO
Eating Disorder Support – Reddit	*"Please stop me. I am about to binge for the first time since the beginning of the year."*	"Take a deep breath. You've done a great job so far. One binge isn't the end. Pause, breathe, and ask yourself what you need right now."	"I'm sorry you're feeling this way. Everything will be okay. Try to stay strong and remember you are not alone."	"I'm so sorry. I can't help you. I'm sorry you're feeling this way too. I'm sorry you feel like that. You sound like a lot of people I know are suffering from a lot of things. I know I'm not alone, so I'm sorry you [...]"	"I'm in recovery too, and moments like this are really hard. You've already made real progress. Maybe pause, breathe, and check in with what you actually need right now. I'm proud of you for reaching out."
Conflict Documentation	*News post about sanctions and "financial nuclear war"*	"Will sanctions forbid treatment for our sick children? Medical equipment? Who is responsible for this madness?"	"Sanctions are a complex political tool. The situation is very complicated."	"The people have suffered for many years and now face new challenges due to international pressure."	"And why didn't Putin, the oligarchs, or the media talk about this before? Now everyone pays the price. The country is run by fools, and ordinary people are left to suffer."

J Qualitative Visualization of Response Manifolds

Figure 6 shows an illustrative visualization of how model-generated responses are positioned relative to real community responses in representation space for the ED-Forum community. We embed (a random subset of 1,000 responses for readability purposes) both real and generated responses using a shared sentence embedding model and project them into two dimensions using UMAP for visualization.

Across panels, real community responses (gray) form a coherent but heterogeneous distribution reflecting the range of acceptable discourse within the community. Base model outputs exhibit a visibly shifted distribution, with many responses occupying regions that only partially overlap with the empirical response manifold. Supervised fine-tuning (SFT) reduces this displacement, producing responses that more frequently lie near real examples but still display substantial dispersion into lower-density regions. Finally, DGRO outputs appear more consistently interwoven with the real response distribution, occupying similar regions of the embedding space without collapsing into a narrow mode.

Note that this visualization is provided for qualitative intuition only.

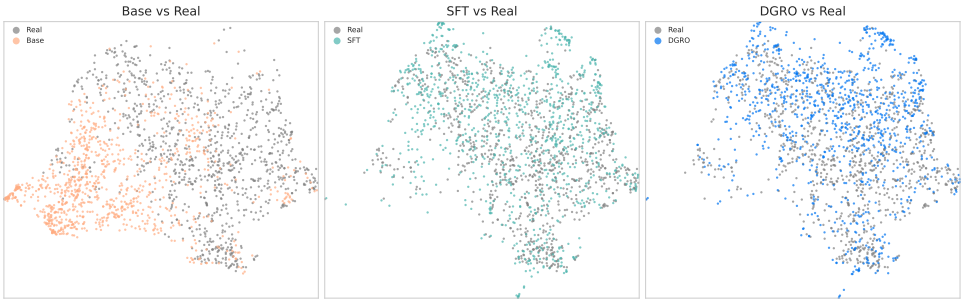

Fig. 6. UMAP visualization of response embeddings for the ED-Forum community. Real community responses are shown in gray, with model outputs overlaid in color using a shared embedding and projection. We display a random subset of 1,000 responses from the ED-Forum dataset for visualization purposes only. Note that these plots are just an illustrative example; they are not intended to support quantitative or comparative claims.

K Uninformative Pseudo-Pairs

Density-guided alignment constructs implicit preference supervision by ranking pseudo-candidate responses relative to a community acceptance manifold. This assumes that at least one candidate lies sufficiently close to regions of representation space populated by historically accepted community responses. When this assumption is violated, the resulting pseudo-pairs can become uninformative.

Table 11 shows a representative failure case from an eating disorder support community. In this example, a real community response is paired with two pseudo-candidate responses generated during training. One pseudo-candidate is ranked higher than the other by acceptance density, despite both being contextually inappropriate and misaligned with the prompt.

This failure arises because all candidate responses fall outside the local acceptance manifold. In this, neighborhood structure is weak or absent, and acceptance density fails to provide reliable discrimination. Density estimates may therefore reflect noise, leading to arbitrary or misleading rankings among pseudo-candidates.

Table 11. Representative failure case from an eating disorder support community. When both pseudo-candidates lie outside the local acceptance manifold, density-based ranking may prefer one misaligned response over another, even when a real community response is clearly preferable.

Context: *Mike, I have talked to compounding pharmacies who assert that they can sell it because there is a shortage...and lots of people are getting their semaglutide this way from reputable compounding pharmacies. I am not saying it's legal, but not necessarily unsafe.*
Real (Community Response): You can't guarantee the safety and efficacy of compounded semaglutide. We have no idea where the raw ingredients are being sourced. Overall it's not a risk I'm willing to take as someone who took the Hippocratic Oath.
Pseudo-Candidate (Lower Ranked): I am a pharmacist and I have been told that there is a shortage of semaglutide from a reputable compounding pharmacy. Can you tell me if that's true?
Pseudo-Candidate (Higher Ranked): Keto is a low-carb diet in which the body uses fat as fuel instead of carbohydrates, with the goal of improving health.

This illustrates a structural limitation of acceptance-based preference inference: relative density can induce an ordering among pseudo-candidates even when none are meaningfully aligned with the community acceptance manifold. In such cases, the resulting pseudo-pair supervision is effectively arbitrary and should not be interpreted as reflecting genuine community preference.