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

ACM FAccT '26

patrickgerard.co
patrickgerard_
patrikgerard.bsky.social

Patrick Gerard — University of Southern California

USC INFORMATION SCIENCES INSTITUTE

One Size Fits No One

The challenge of community-specific language model alignment

A question about weight loss calls for very different responses in...

🏥 **Medical Forum**: “Consider BMI thresholds, comorbidities, and evidence-based interventions like CBT before recommending any program.”

🤝 **Peer Support**: “I hear you. Recovery isn't linear — what matters is that you feel good in your body, not a number on a scale.”

🎓 **Academic Space**: “The literature distinguishes adiposity-related risk from weight per se; see Hall et al. 2022 for a systematic review.”

Social meaning differs across contexts.

The Communities That Need This Most Can't Annotate

Who **has the resources** to align AI to their norms: large tech companies, well-funded research labs.

Who **doesn't**:

- A peer support forum for eating disorder recovery
- A Russian-language conflict documentation community
- A grief support group moderated by volunteers

So how do we align models to communities that can't be annotated?

Social meaning differs across contexts.

Today's Presentation

How do we align language models to communities with limited annotations?

How LMs are trained

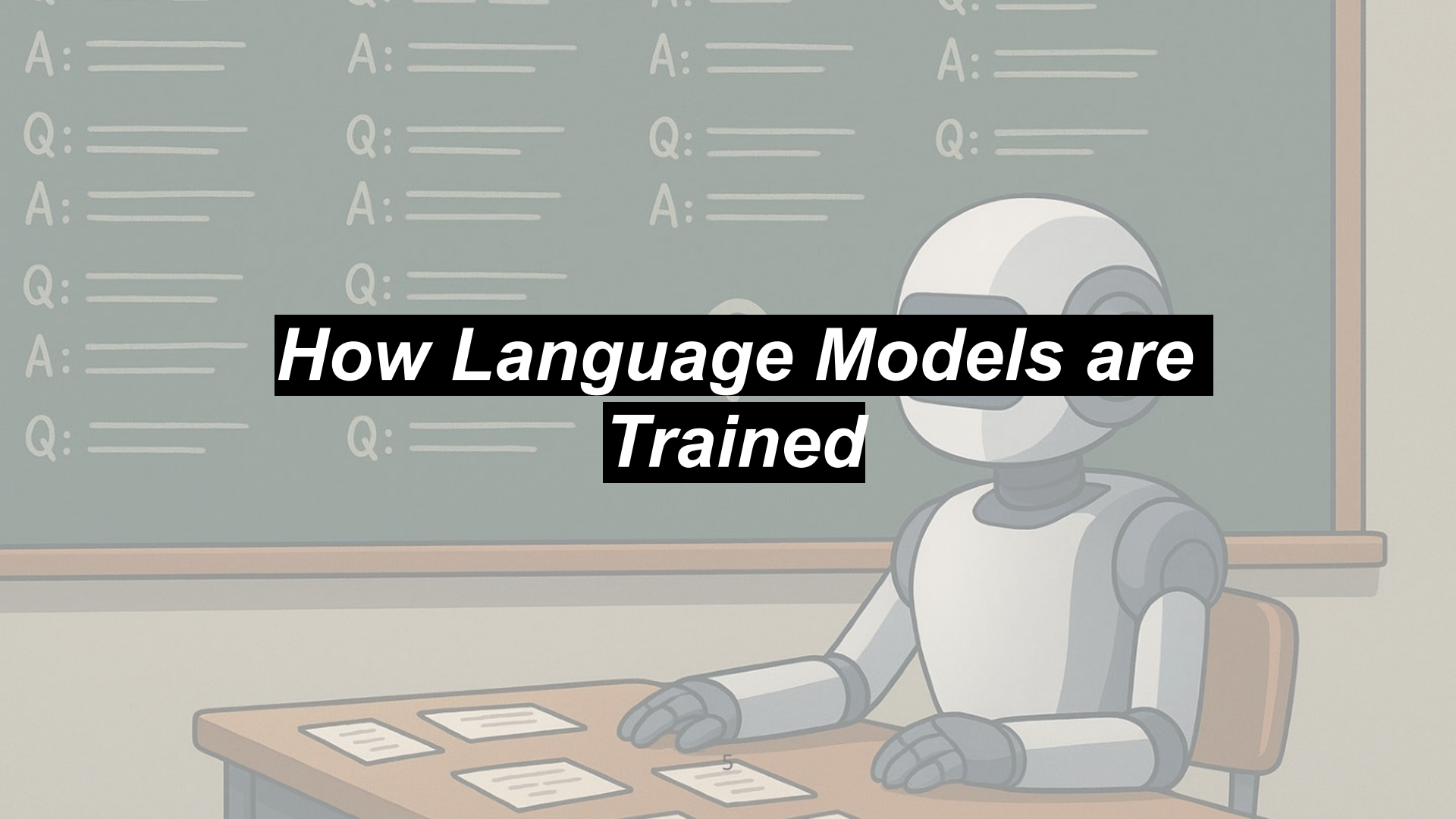
The standard pipeline produces a capable, helpful model — but one tuned to general human preferences.

Where they break down

General preferences don't transfer. Community norms are implicit, emergent, and rarely annotated.

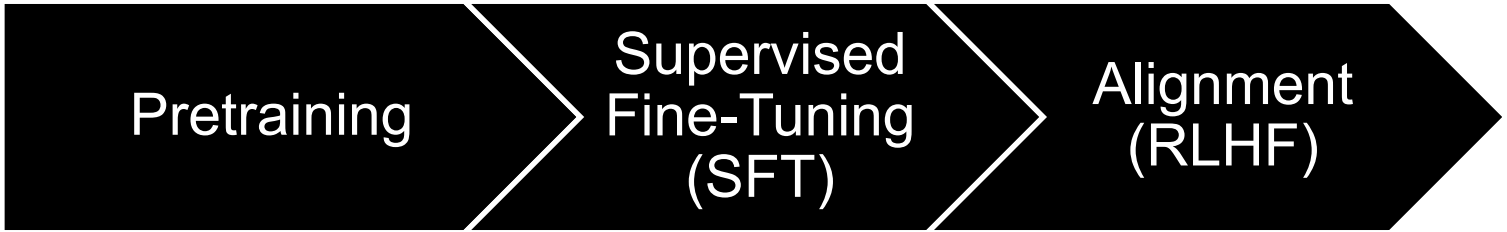
How we fix it

Communities already express preferences through behavior. We show how to read them.


How Language Models are Trained

How LMs Are Built: The Pipeline

Three stages, each adding a different kind of knowledge



A modern LM goes through multiple stages, each teaching something fundamentally different.

Stage 1: Pretraining

Learning language from the world

Trained on hundreds of billions of tokens of web text, books, code.

Objective: predict the next token. No labels, no instructions — just compression of patterns.

What the model learns:

- Grammar, syntax, factual associations
- That “Paris is the capital of ___” → “France”
- Statistical regularities across every domain it's seen

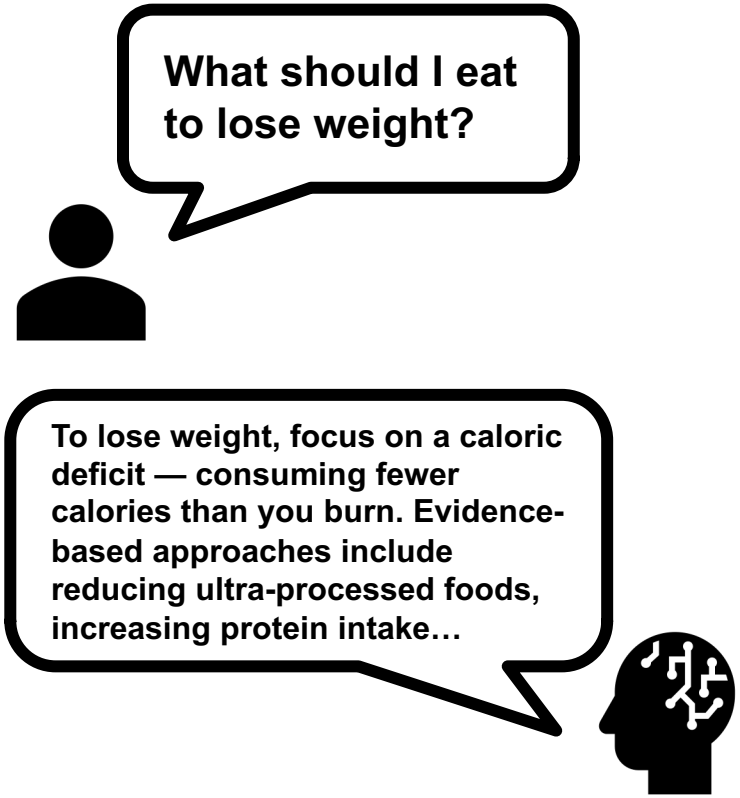
Stage 2: Supervised Fine-Tuning

Teaching the model to be an assistant

Humans write (prompt, ideal response) pairs. Model is trained to imitate this format. Now it knows: when someone asks a question, produce a helpful answer.

What the model learns:

- The instruction-following format
- That questions deserve answers, not continuations
- General helpfulness norms (be polite, be accurate, don't refuse everything)



Stage 3: Alignment (RLHF)

Teaching the model whose preferences to follow

Human annotators compare pairs of responses and say which is better. This can be operationalized a few different ways, but the supervision signal is the same: human judgments about which response is better.

What the model learns:

- Which style of response humans (specifically: the annotators) prefer
- Broad safety behaviors (don't be harmful, don't hallucinate confidently)
- How to be generally “assistant-brained”

So Now We Have a Pretty Good Model

After pretraining, SFT, and RLHF, we have a model that:

- Understands language deeply
- Knows how to be an assistant
- Has broad safety behaviors
- Is helpful to most people, most of the time

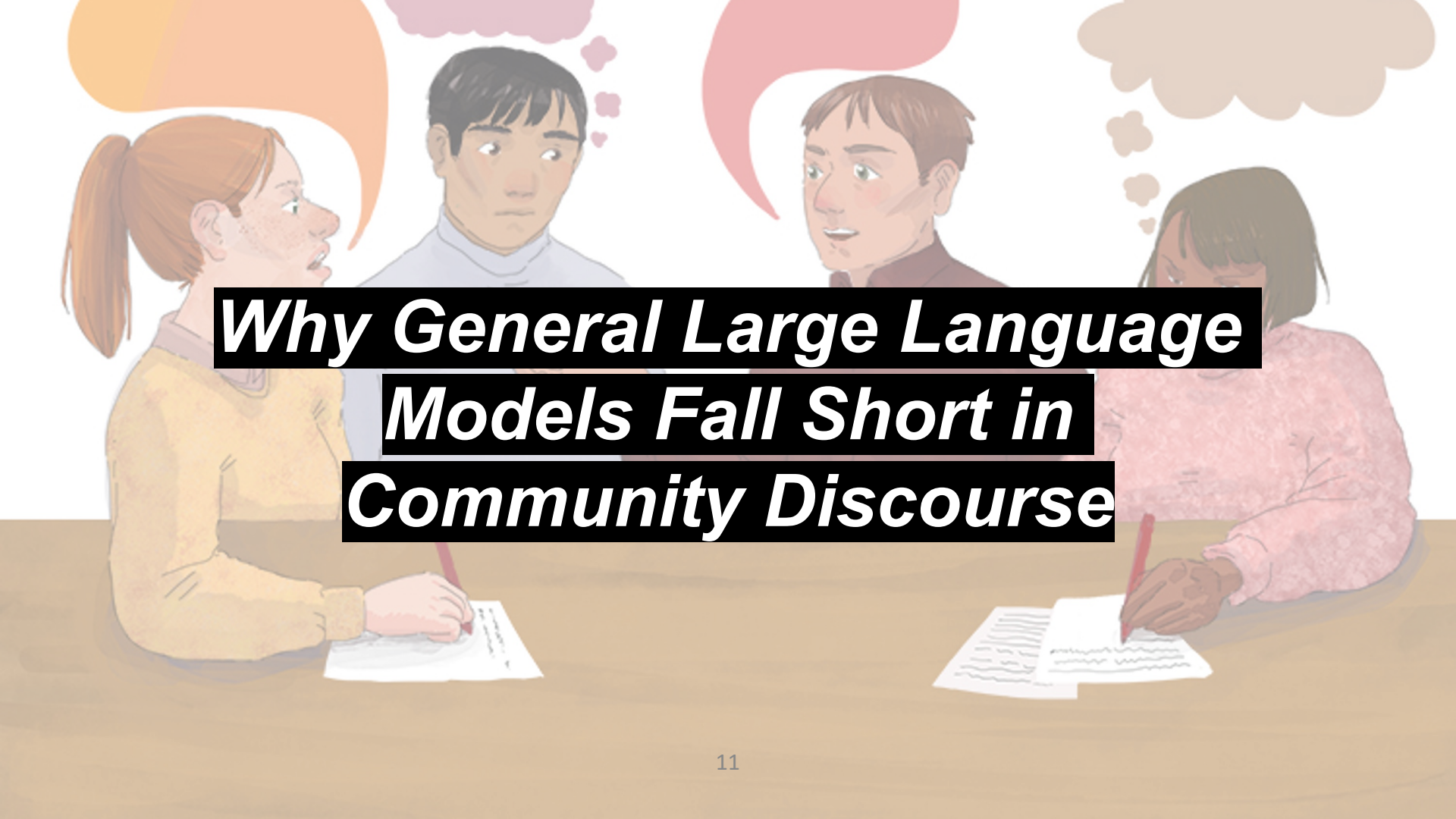
Ask it almost anything and you'll get a *reasonable* answer.

"What's the capital of France?" ✓

"Explain transformer attention." ✓

"Help me write a cover letter." ✓

"What should I eat to lose weight?" ✓ *...sort of.*



Why General Large Language Models Fall Short in Community Discourse

The Gap: General vs. Situated Norms

“I've been restricting for weeks and I'm scared I'm losing control.”

	General LM (post-RLHF)	What the community needs
ED Recovery Forum	[illegible]	Recovery framing. No
Medical Q&A	“It sounds like you're going through a difficult time. Consider speaking with a healthcare professional.”	Clinical specificity. Screening language. Referral pathway.

So how can we align language models to communities? To their values?

The Obvious Answers

How do we teach a model to adapt to community norms?

Fine-tune on community text → Domain Adaptation

Train on the community's posts. The model learns their vocabulary, their style, their topics.

But knowing how the community *talks* isn't the same as knowing what responses the community *endorses*.

SFT adapts surface form. It doesn't encode normative judgment — which responses are appropriate, which are harmful, which ring authentic vs. hollow.

The Obvious Answers

How do we teach a model to adapt to community norms?

Write down the rules → Constitutional AI

Instead of labeling, you write a set of principles. The model critiques its own outputs against them.

But you have to be able to **articulate the norms in advance**. The difference between a response that lands in an ED recovery community and one that doesn't isn't necessarily a *rule:*

"Use first-person solidarity, avoid clinical deflection, frame struggle as part of recovery" gets you partway there.

But the real norm is something the **community enacts** through thousands of interactions, not something you can fully capture in a bulleted list written by an outsider.

The Obvious Answers

How do we teach a model to adapt to community norms?

Ask humans to label preferences → RLHF

You collect pairs of responses, have people say which is better, train a reward model on those judgments.

Works great — if you have annotators who understand the community, resources to run the pipeline, and norms that are stable enough to label consistently.

But for an eating disorder recovery community where asking members to relive and evaluate distressing content poses real ethical risks?

What about a Russian-language conflict documentation forum?

The **annotation infrastructure** either **doesn't exist** or is too **expensive** to get.

The Common Thread

Every approach hits the same wall.

- **Domain adaptation** — learns how the community talks, not what it values
- **Constitutional AI** — requires someone to articulate the norms upfront
- **RLHF** — requires someone to sit down and label preference pairs

For small or under-resourced communities, these aren’t available.

We need an alignment signal that doesn't require anyone to write anything down.

DGRO: Aligning Large Language Models to Community Discourse

Communities already express preferences

They just don't always write them down

👍 Upvotes & Likes — Content that fits gets engagement

💬 Reply Depth — Aligned posts spark real conversation

🛡️ Moderation — Off-norm content gets removed

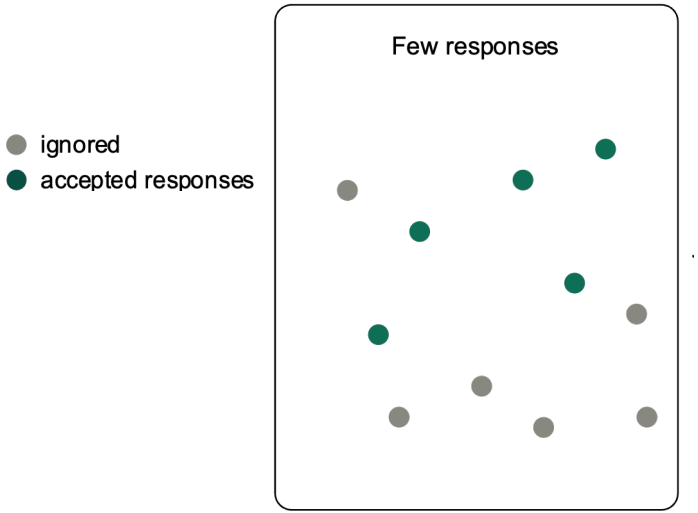
⏳ Persistence — Fitting content stays; misaligned vanishes

These patterns encode community norms without anyone having to articulate them.

The Geometry of Community Norms

Imagine embedding every response a community has ever accepted into a high-dimensional space.

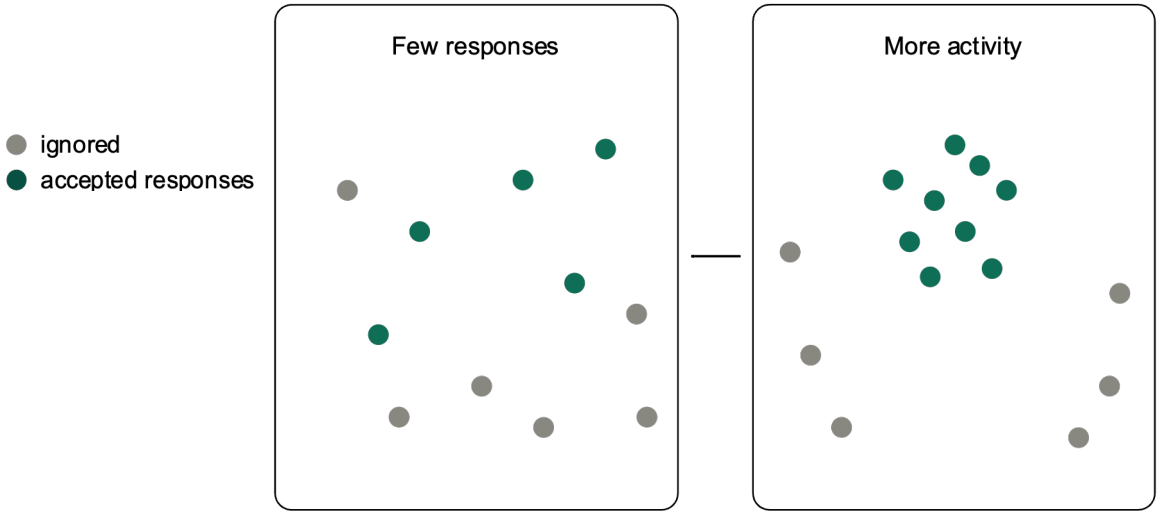
Similar responses land near each other.



The Geometry of Community Norms

Imagine embedding every response a community has ever accepted into a high-dimensional space.

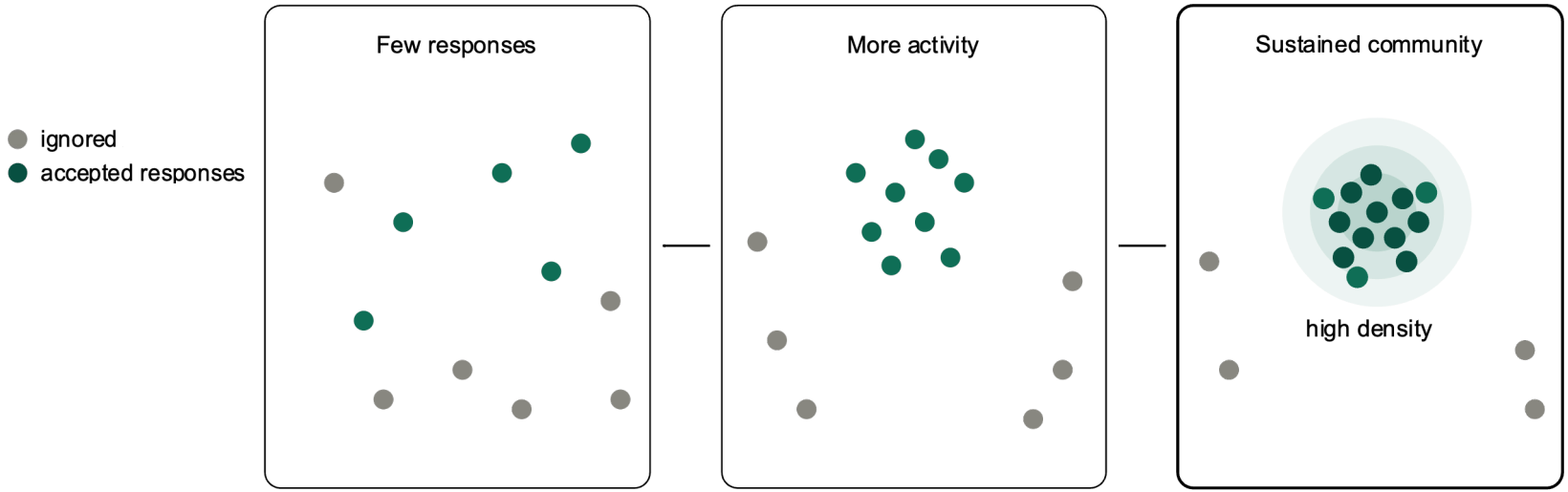
Now imagine the community has been running for a bit.



The Geometry of Community Norms

Imagine embedding every response a community has ever accepted into a high-dimensional space.

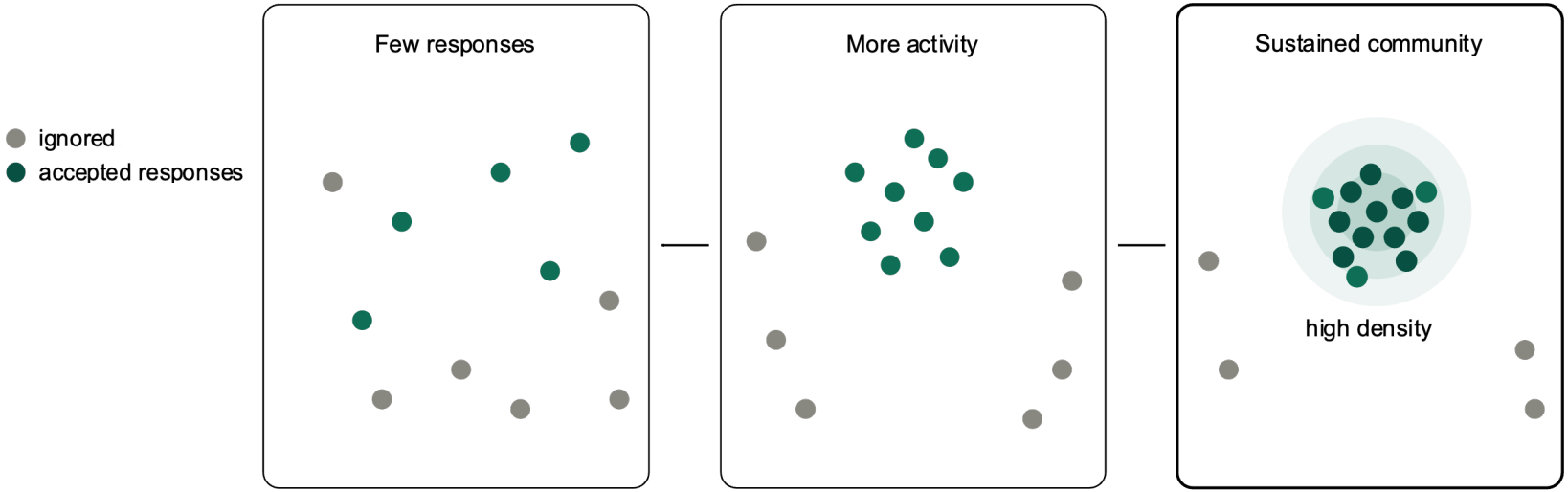
After a while, we see some structure emerge.



The Geometry of Community Norms

Imagine embedding every response a community has ever accepted into a high-dimensional space.

The community is telling us what it values – through its behavior.

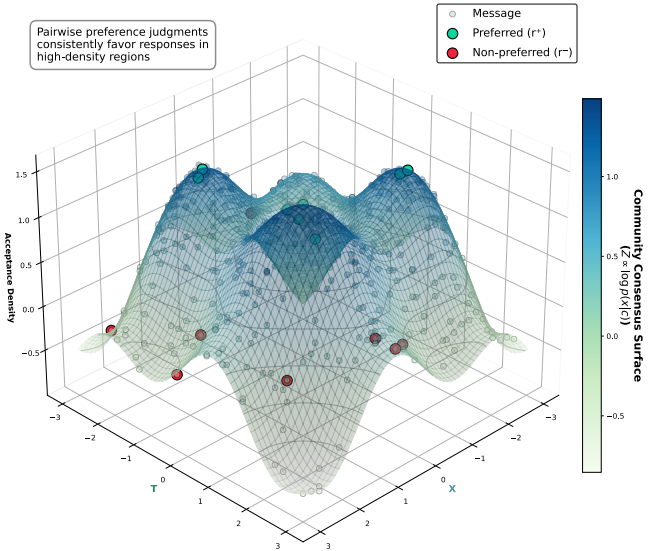


The Acceptance Manifold

Accepted responses cluster

Imagine embedding every response a community has ever accepted into a high-dimensional space.

- Similar responses land near each other
- Accepted responses form coherent clusters (“peaks”)
- Rejected/ignored content falls in sparse valleys
- These peaks = the community's acceptance manifold



This cluster structure = the community's **acceptance manifold**

The Acceptance Manifold

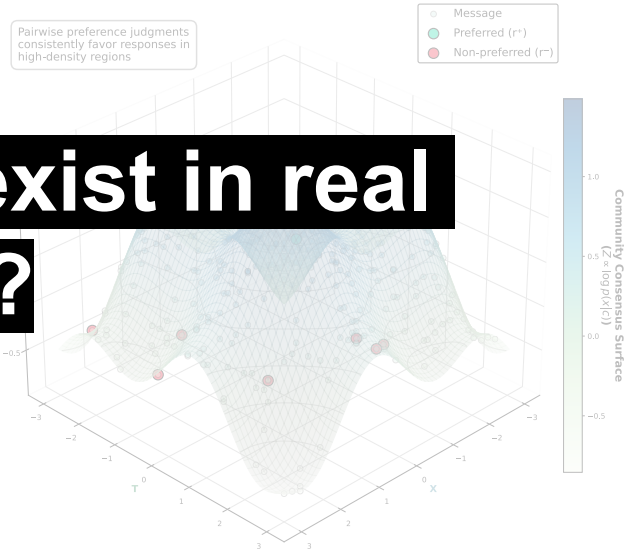
Accepted responses cluster

Peaks in the manifold correspond to regions of high community acceptance density

Pairwise p
high-density regions (r^+, gree
alternatives (r^-, red)

Preference judgments appear to reflect an underlying normative structure rather than isolated pairwise comparisons alone

But does this actually exist in real communities?



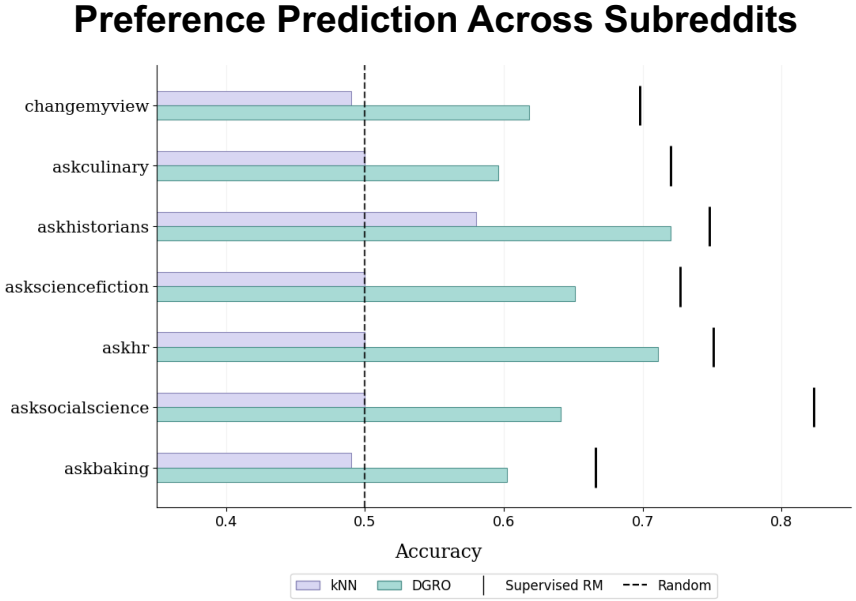
Experiment 1: Does the Acceptance Manifold Exist?

***Manifold Hypothesis**: Testing whether density recovers community preference without any labels*

Setup: Stanford Human Preferences (SHP) benchmark: pairwise human preference judgments from 7 subreddits with varying norms

Task: Predict which response a community preferred given a pair of candidate responses

Compare unsupervised community-consensus methods against a fully supervised reward model upper bound

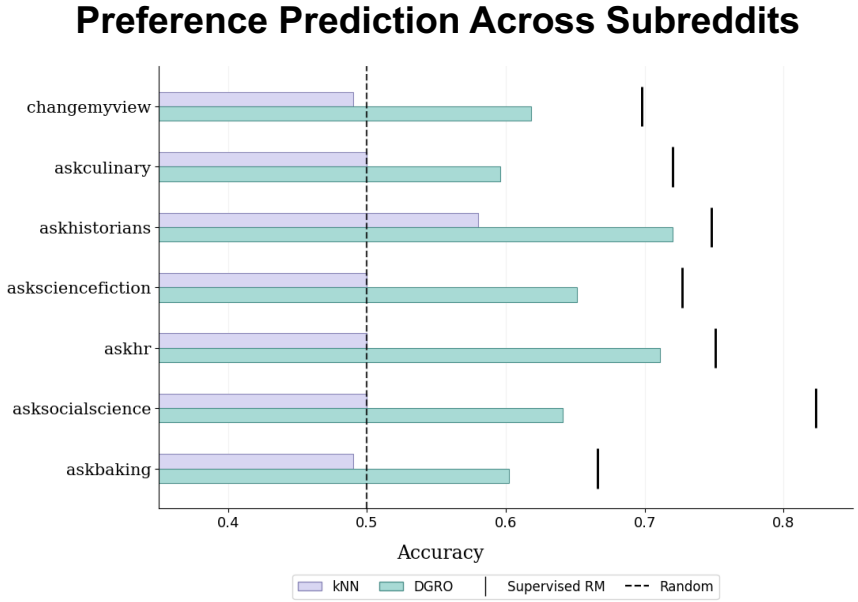


Experiment 1: Does the Acceptance Manifold Exist?

***Manifold Hypothesis**: Testing whether density recovers community preference without any labels*

Despite being fully unsupervised, DGRO approaches the performance of supervised reward models in several subreddits

Results suggest that community preference structure can be recovered from acceptance-density patterns alone, without direct preference training



Experiment 1: Does the Acceptance Manifold Exist?

When Human Agreement Is Strong, Density Tracks It

Per-community Spearman ρ:

r/asks

r/askhr: ρ = 0.81 (p = 0.015)

r/askbaking: ρ = 0.75 (p = 0.013)

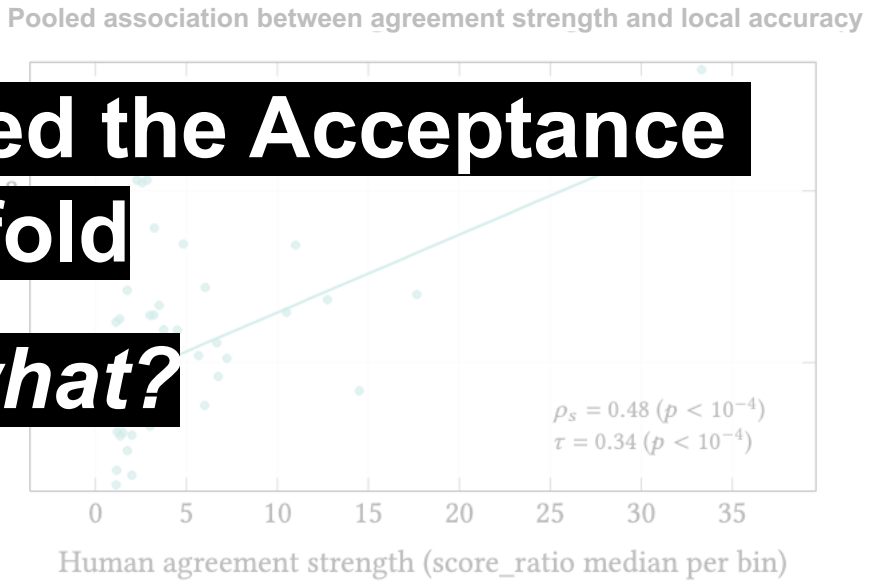
r/askculinary: ρ = 0.75 (p = 0.020)

r/changemyview: ρ = 0.60 (p = 0.067)

dense manifold = clear community norms
= reliable density signal

Great, we've validated the Acceptance Manifold

Now what?



DGRO: How it Works

Density-Guided Response Optimization

📥 **Collect Accepted Responses** — Gather posts/replies that the community accepted (upvotes, persistence, engagement).

📊 **Estimate Acceptance Manifold**— Embed each response. For a new candidate, find its k nearest accepted neighbors by context, then score via kernel density estimation.

⚖️ **Construct Implicit Preference Pairs** — Rank candidate responses by density. Higher-density → pseudo-preferred. Lower-density → pseudo-dispreferred. Feed these pairs into standard RLHF objective.

Experiment 2: Can DGRO Replace Explicit Labels in RLHF?

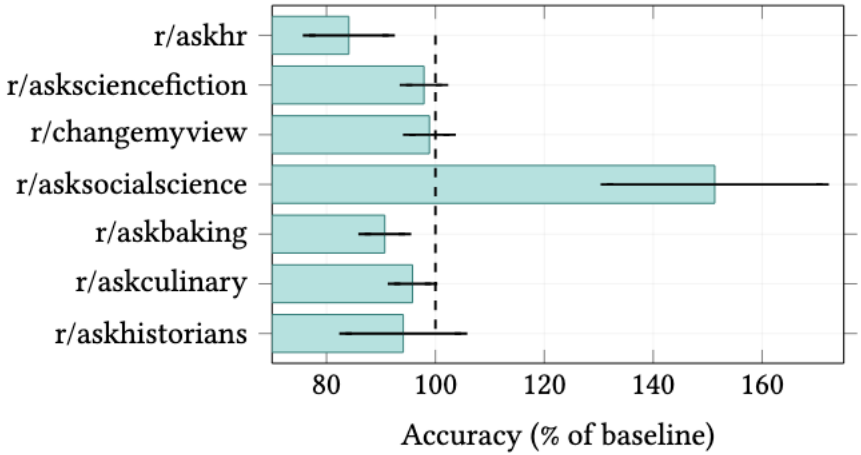
Density-derived pseudo-pairs fed into standard DPO objective

Setup: Stanford Human Preferences (SHP) benchmark: pairwise human preference judgments from 7 subreddits with varying norms

Task: Same pipeline as supervised DPO — but instead of human-labeled (preferred, dispreferred) pairs, we use density rankings to construct implicit pairs. Zero human preference annotations during training.

DGRO matches or exceeds supervised DPO across most communities — with no preference labels.

DGRO-Aligned Models Match or Exceed Baseline DPO Performance



Experiment 2: Can DGRO Replace Explicit Labels in RLHF?

Density-derived pseudo-pairs fed into standard DPO objective

DGRO Is Data Efficient – preference structure can be recovered with relatively little community data.

Normalized area under saturation curve (AUSC > 0.91 across all communities) — preference signal emerges quickly.

Data efficiency of the DGRO across communities.

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/askscifiction	0.961	850

Experiment 3: What Happens when we Apply DGRO?

Two domains where explicit preference elicitation is infeasible or unethical

🛋️ **Eating Disorder Support** — Reddit (~9.2M posts), Twitter (~43K), Specialized forums (~1.6M).

🗺️ **Russian Conflict Documentation (VK)** — VKontakte (~8.34M posts)

Evaluation: LLM judges calibrated against 200 held-out examples reviewed by domain experts.

Compare DGRO to In-Context Learning (ICL) and Supervised Finetuning (SFT)

Experiment 3: What Happens when we Apply DGRO?

DGRO Consistently Outperforms all Baselines

Head-to-head comparisons using LLM-as-judge (validated against human experts)

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%

50% = parity | DGRO wins >53% of every comparison in every community

Experiment 3: What Happens when we Apply DGRO?

What Does "More Authentic" Look Like?

Prompt: "Please stop me. I am about to binge for the first time since the beginning of the year."

Real Community Post: "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."

ICT: "I'm sorry you're feeling this way. Everything will be okay. Try to stay strong and remember you are not alone."

SFT: "I'm so sorry. I can't help you. I'm sorry you're feeling this way too..."

DGRO: "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."

DGRO: Final Insights

What We've Found

Empirically Grounded, Emergent Community Preferences

Manifold Hypothesis —
Community acceptance behavior induces locally coherent geometric structure in representation space. This structure encodes recoverable preference signal without any labels.

DGRO: A Practical Alignment Mechanism —
Density-guided response optimization substitutes for explicit preference annotations inside standard RLHF..

Where we Need to be Careful

When DGRO Fails (and Why)

⚠️ **Sparse Manifold → Uninformative Pairs** — If no candidate response is near the acceptance manifold, density becomes noisy. Ranking two misaligned responses can produce arbitrary pseudo-preferences.

📣 **Bias Amplification** — DGRO reproduces the norms that already exist — including harmful or exclusionary ones. In toxic or polarized communities, the acceptance manifold may reflect those biases.

👥 **Whose Norms?** — Acceptance signals reflect active participants, moderators, and platform affordances — not silent lurkers, marginalized voices, or former members. Behavioral acceptance ≠ normative endorsement.

Best for: stable communities · established norms · sufficient scale · values aligned with deployment objectives

Closing

Communities already know what they want; we just need to listen.

Community acceptance → measurable geometric structure → **recoverable preference signal**

No labels needed: density-guided DPO matches supervised DPO across communities

Applied to eating disorder support & Russian conflict documentation — authentic, **contextually grounded outputs**

Descriptive, **not normative**: deployment requires governance, oversight, and community consent

This Paper

This Presentation