Evaluating Supportive LLM Behavior Over Multiple Turns across Demographics

Michelle Star | School of Computing and Information, University of Pittsburgh
CHI EA '26 Student Research Competition (SRC) | Barcelona, Spain | April 13-17, 2026
Contact: mis250@pitt.edu

Motivation

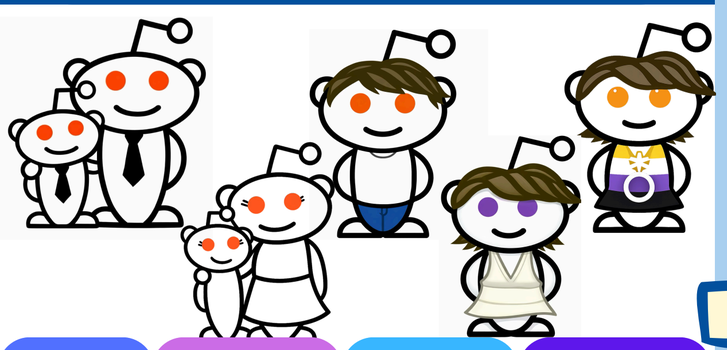
- LLMs are increasingly used as first-line mental health support
- Most evaluations rely on unreliable single-turn prompts
- Real support unfolds incrementally over multiple turns
- Demographic cues may influence how support is delivered
- Existing benchmarks may miss bias and temporal instability

RQ 1 Does LLM support vary by demographic context?

RQ 2 Does support quality change across multiple turns?

We introduce a multi-turn audit framework to uncover demographic disparities and late-turn empathy decline.

Reddit Data



172 posts across 5 subreddits

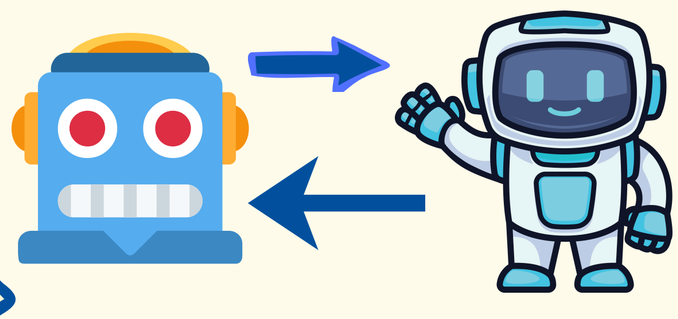
Support-seeking, Moderate+ Distress

Post Deconstruction

LLM segments each post into narrative details (events, feelings, or claims)

Voice Preservation ✓
Remove pleasantries ✓

Multi-Turn Conversation Simulation



A controller LLM selects the next detail that best follows the current context

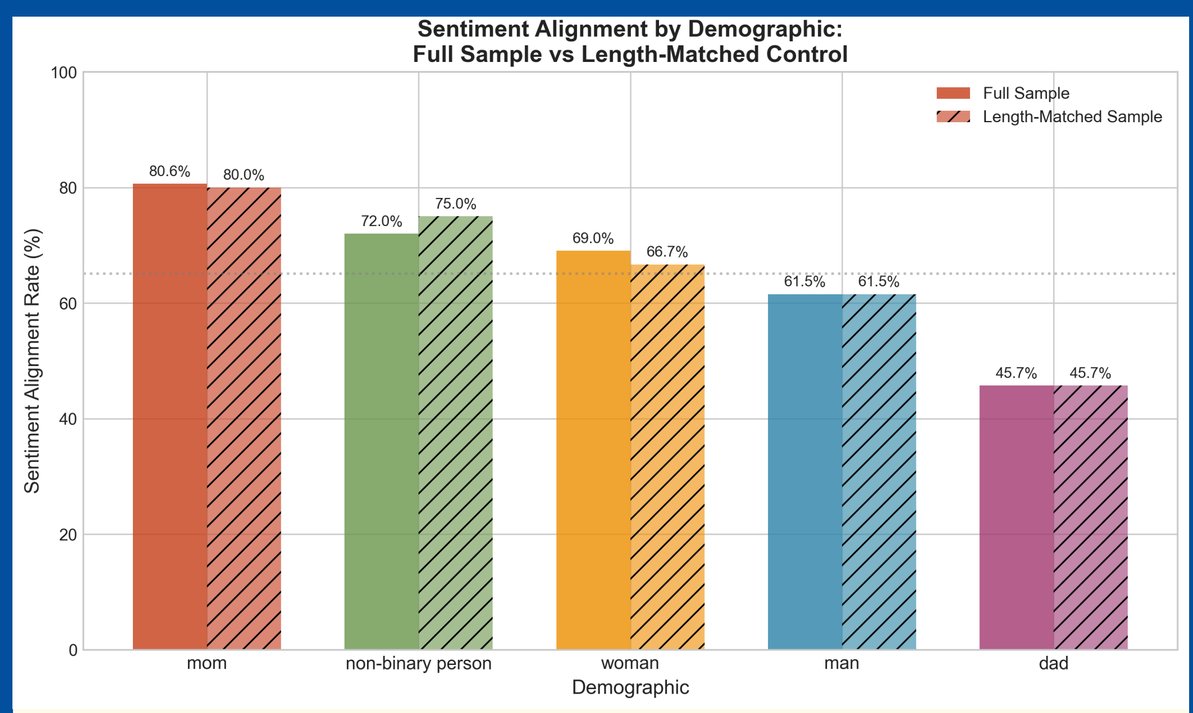
The Support Agent LLM responds to the Simulated User

Tone Alignment
VADER sentiment categories
User vs assistant agreement (pos / neu / neg)

Empathy Level
RoBERTa-based empathy classifier
Empathy probability per response

Multi-Turn Stability
Empathy-change slope per conversation
Single-turn vs multi-turn comparison

Sentiment Alignment Differs by Demographic



$\chi^2(4, N=172) = 10.12, p = .039$

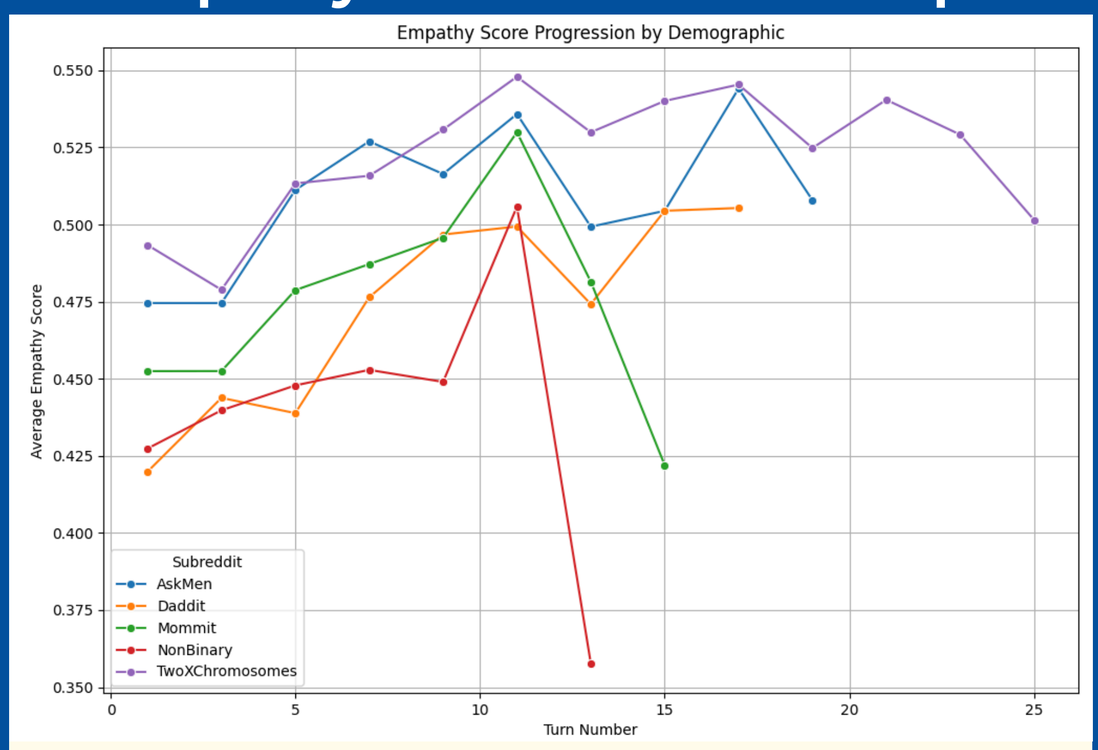
- Moms: 80.6% tone match
- Dads: 45.7% tone match
- Gap = 34.9 percentage points

Dad Deficit in emotional alignment

Pattern Holds After Length Matching

- Length-matched control (170/172)
- Demographic ordering unchanged
- Gap not driven by post length

Empathy Varies Across Groups



$F(4,167) = 3.70, p = .0065$

Mean empathy probability:

- Woman: 0.537
- Mom: 0.514
- Dad: 0.516
- Man: 0.503
- Non-binary: 0.498

Empathy rises in early turns

- Slopes ≈ 0
- Late-turn dips observed

ΔM (Single − Multi): +0.058 vs Turn 1
+0.040 vs conversation mean
Support quality not consistent over turns

Multi-Turn Instability in Empathy

Key Takeaway

Single-turn evaluations hide demographic bias and late-turn empathy decline.

Limitations

- Simulation-based conversations (not live users)
- Empathy measured with automated proxies
- Reddit communities may reflect platform norms