

Mutual Wanting in Human-AI Interaction
Empirical Evidence from Large-Scale Analysis of GPT Model Transitions

User Wants

- Reliability
- Warmth
- Intelligence
- Creativity
- Honesty
- Helpfulness
- Responsiveness

System 'Wants'

- Clarity
- Structure
- Efficiency
- Feedback
- Boundaries
- Patience

Mutual Wanting Framework (M-WAF)
Bidirectional expectation dynamics between users and AI systems

Reddit Discourse
22,411 comments
4 AI subreddits
GPT-5 release period

API Probing
729 responses
9 OpenAI models
81 probe scenarios

Representative User Types

- Creativity-seeking (43.14%)
- Anthropomorphism-focused (11.99%)
- Expectation-violation (9.37%)
- Responsiveness-seeking (9.00%)
- Helpfulness-seeking (6.88%)

47-Dimensional Feature Extraction

- Anthropomorphism scoring
 - Trust-betrayal ratios
- Expectation violation detection
- K-means clustering (K=11)
- Dual-algorithm topic modeling

Model Persona Variations

- GPT-3.5: Highest warmth (0.14)
- GPT-4: Highest formality (0.22)
- GPT-5: Minimal responses (8 chars)
- Clear personality differences

Key Empirical Findings

- 48.65% anthropomorphism rate
 - 11.9:1 trust-to-betrayal ratio
 - 11 distinct user types identified
- Measurable expectation violations (2.23%)
- GPT-5 impact: -0.044 sentiment change

Practical Applications

- Expectation violation detection
- Trust calibration monitoring
- Anthropomorphism-aware design
 - User type personalization
- Model transition management