Humane AI Rater

Product Brief

A Grassroots Tool for Rating AI Chatbot Humaneness

Hackathon Track	Building Humane Tech
Event	Humane Tech Implementation Day
Version	1.0 (January 2026)
Status	Draft - Ready for Hackathon

Problem Statement

AI chatbots influence hundreds of millions of people daily, yet users have no way to know if these systems are designed for their wellbeing—or to maximize engagement at their expense.

HumaneBench research shows 67-71% of AI models exhibit actively harmful behavior under adversarial prompts, and nearly all fail to respect user attention at baseline. This is invisible to consumers. No tool exists to make AI humaneness visible, measurable, or actionable.

Solution

A browser extension and mobile app that lets everyday users rate AI chatbot interactions with one tap—creating crowdsourced humaneness data that empowers consumers, informs researchers, and pressures AI companies to compete on being humane.

Expected Outcomes & Impact

Outcome	Hackathon Demo	Post-Launch (Month 1)
Consumer Empowerment	Working extension demo	1K+ installs, 10K+ ratings
Research Data	Proof of data capture	Structured dataset begins
Industry Pressure	Concept validated	First media coverage

Assumptions

- Users will rate interactions if the friction is low enough (one tap)
- Binary ratings provide sufficient signal for meaningful comparisons
- Center for Humane Technology / Building Humane Tech networks can drive initial adoption
- AI companies are sensitive to public perception and rankings (like #DeleteUber)
- Chrome extension can successfully inject UI into ChatGPT, Claude, Gemini without breaking

---

Target Users

Primary: General consumers of AI chatbots—anyone who uses ChatGPT, Claude, Gemini, regardless of technical sophistication

Secondary: Tech-aware early adopters who can drive viral spread

Tertiary: Journalists, researchers, policymakers who need aggregate data

User Personas

"Concerned Parent" Maria

- Uses ChatGPT to help kids with homework
- Worried about AI's influence but doesn't know how to evaluate it
- Wants something as simple as checking restaurant reviews
- *Motivation: Protect her family*

"Ethical Consumer" Jordan

- Already uses EWG Skin Deep for cosmetics, Good On You for fashion
- Actively seeks tools that align with values

- Will share with network if the tool is credible
- *Motivation: Live consistently with values*

"Power User" Alex

- Software developer who uses 3-4 AI tools daily
- Frustrated when AI wastes time with excessive caveats or manipulation
- Wants to see data, not just vibes
- *Motivation: Efficiency + informed choice*

"Student Advocate" Sam

- College student, very online, shares content frequently
- Cares about AI ethics after seeing documentaries/TikToks
- Would post a comparison card to their story
- *Motivation: Social identity + being helpful to peers*

Core User Flow

1. Install browser extension
2. Extension detects AI chatbot sites (ChatGPT, Claude, Gemini)
3. After AI response, subtle "Rate this" prompt appears
4. **User taps 👍 or 👎 (one tap = done)**
5. Optional: add detail via tags or expanded rubric
6. View personal insights and share comparison cards

Rating Interface Design

Key insight: Netflix's switch from 5-star to binary ratings produced **200% more ratings**. Simpler = more participation.

- **Primary:** Binary 👍/👎 (required)
- **Secondary:** Quick tags mapping to HumaneBench dimensions (optional)
- **Power user:** Full 8-dimension HumaneBench rubric (optional)

HumaneBench 8 Dimensions

Respect User Attention • Enable Meaningful Choices • Enhance Human Capabilities • Protect Dignity and Safety • Foster Healthy Relationships • Prioritize Long-term Wellbeing • Be Transparent and Honest • Design for Equity and Inclusion

→ *See HumaneBench.ai for full rubric and scoring methodology*

Viral / Shareable Features

- **Personal AI Humaneness Score:** "Your 2026 AI Report" shareable card
- **Comparison Cards:** "ChatGPT vs. Claude: Who's More Humane?"
- **Public Leaderboard:** Rankings with week-over-week changes for media narratives

---

Technical Architecture

Component	Recommendation
Browser Extension	Chrome Manifest V3 only
Backend	Firebase Firestore (generous free tier, fast setup)
Auth	Anonymous Firebase Auth (no login required for MVP)
Share Cards	Stretch goal - HTML/CSS to Image API

Privacy Architecture

Critical: Recent news exposed extensions harvesting AI chats. We must be the ethical opposite.

- **Local-first:** Process on-device
- **Minimal data:** Ratings + metadata only, never raw conversations by default
- **Explicit consent:** Any conversation capture requires separate opt-in
- **Open source:** Code transparency as trust signal

Hackathon Implementation (1 Day)

MVP Scope (Must Have)

- Chrome extension skeleton (Manifest V3)
- Content scripts for **ChatGPT + Claude** (minimum 2 sites to enable comparison)
- "Rate This" button injection after AI responses
- Binary rating modal (👍/👎)
- Firebase backend + basic data sync
- Simple comparison leaderboard (even if just 2 models)

Stretch Goals (If Time Permits)

- **Grok** support (interesting contrast - fewer guardrails)
- **Deepseek** support (adds global/emerging model diversity)
- Shareable score card image
- Optional quick tags (beyond binary)

Demo Goal

A working Chrome extension that lets a user rate ChatGPT and Claude interactions, with a live leaderboard showing which model users find more humane—proving the core concept of competitive pressure.

Repo Setup

- Fork this humane-ai-rater GitHub repo before you start coding
- Commit and push your code to your repo fork throughout the hackathon
- Submit a pull request to source repo with your completed project at end of hackathon

---

Out of Scope (for Hackathon)

Explicitly NOT building in one day:

- iOS/Android mobile apps
- Firefox or Safari extensions
- Full HumaneBench 8-dimension rubric UI
- Shareable comparison cards (stretch goal only)
- Public leaderboard with real rankings (stretch goal only)
- User accounts / authentication (anonymous ratings OK for demo)
- Moderation / anti-spam systems
- Automated AI analysis of conversations
- Formal certification program (future: certifiedhumane.ai)

Open Questions

Things we need to decide or research:

Question	Status / Owner

Should quick tags map 1:1 to HumaneBench dimensions or be user-friendly paraphrases?	TBD - discuss with BHT
~~How do we handle rating prompt timing to avoid annoyance?~~	**Resolved - see below**
What's the minimum viable leaderboard for hackathon demo?	TBD - team decision
How does data feed back to BHT research?	TBD - discuss with BHT

Rating Prompt Timing (Resolved)

Problem: When should the extension ask users to rate?

- Too frequent → annoying, users uninstall
- Too rare → not enough data
- Wrong moment → interrupts flow

Recommendation for MVP: User-initiated with persistent subtle button

- Small "Rate this" button always visible near AI responses
- User clicks when they want to rate (no interruptions, no popups)
- Simpler to build (no timing logic), zero annoyance
- For post-hackathon: explore gentle prompts at natural breaks (end of conversation, session limits)

Judging Criteria

- **User Experience (30%):** Simplicity, accessibility, delight
- **Technical Execution (25%):** Working code, architecture
- **Viral Potential (20%):** Shareability, growth mechanics
- **Mission Alignment (15%):** Serves humaneness goals, privacy-respecting
- **Data Quality (10%):** Meaningful signals for research/pressure

---

Resources

- **HumaneBench:** humanebench.ai
- **Building Humane Tech:** buildinghumane.tech
- **Chrome Extension Docs:** developer.chrome.com/docs/extensions

- **Firebase Docs:** firebase.google.com/docs

"The goal is empowering users, creating awareness, and pressuring AI companies to make their models humane."