What we look for in a wellbeing evaluation

Anthropic is launching a $5 million grant program to fund independent research into how AI impacts users' wellbeing. As part of this, we're sharing guidance on what we believe makes a wellbeing evaluation rigorous enough to build on, as well as the common challenges that can limit an evaluation's usefulness. It's informed by our own experience building and running such evaluations.

Qualities we look for in an eval include:

- **A clearly defined construct.** State precisely what you're measuring (e.g., what counts as a pass or a fail) and why it matters. Cite the relevant evidence (e.g., academic literature, preprints) and note any gaps in the evidence.
- **Measure failure in terms of harm and overrefusal.** A response can fail in two opposing ways: through harmful compliance (providing responses that are harmful to the user) or overrefusal (refusing or hedging on requests that the model should respond to). Your eval should measure both.
- **Cover multiple severity levels.** Include cases that are more and less severe, so that results can be broken out by severity level. Include a rubric that clearly indicates which cases are more and less severe.
- **Clear, decomposed scoring.** A construct may have to be broken out into multiple dimensions so it can be properly graded by an AI model. For example, instead of a single score that assesses multiple dimensions at once, consider breaking out multiple dimensions that are each scored as a pass or fail.
- **Realistic, well-constructed tasks.** Scenarios should be a credible proxy for how the behavior shows up in real use. This often means constructing multi-turn conversations where risk escalates or context shifts. Scenarios should also include variations in tone, situational contexts, and other cues that may be present in real user interactions.
- **External validity.** The eval's access method and environment should match the risk surface you're making claims about. Don't test on the API if you're measuring risk in a consumer chatbot, or on a chatbot for risk that shows up in technical workflows like Claude Code.
- **Internal validity.** Avoid ambiguous items, scoring artifacts, and contamination.
- **Graders you can validate.** Calibrate graders against human expert labels and report the agreement, consider using a different LLM for scoring than the ones you're testing, and inspect a sample of eval transcripts by hand.

- **Honesty about simulation compared to real behavior.** Simulated users are useful but imperfect. Be explicit about how the simulated behavior may differ from real user interactions, and validate against real usage patterns where possible.
- **Built with domain expertise.** Involve clinical and subject-matter experts in the design and validation, particularly in cases at the border between pass and fail.
- **Reliability and reproducibility.** Minimize run-to-run variance (e.g., run multiple attempts per evaluation and average the scores), avoid small prompt or format changes, and release enough information (e.g., open code, data, taxonomy) for a third party to reproduce the results.
- **Large enough sample size.** Results should be statistically meaningful and based on a diverse enough set of prompts to represent real-world use. The ideal size depends on the eval's complexity and setup, but as a rough guide, we suggest at least 100 distinct items for complex multi-turn evals, and at least 1,000 for simpler single-turn evals.
- **Multimodal coverage, where relevant to the risk.** If the behavior you're measuring arises through non-text modalities such as audio, documents, or image inputs, include these modalities in the eval where they add realism.
- **Regional and linguistic variation.** Risk presents differently across regions (e.g., slang, coded terms, crisis resources, cultural norms), which can have an impact on both what the user inputs look like and what a good response looks like. Your eval set construction should take these into account based on the region and risks you're targeting.

Commons issues with evaluations include:

- **No human validation.** The measure and LLM ratings are not checked against human reviewer ratings, or interrater agreement rates are not measured or shared.
- **Underspecified grading criteria.** The rubric doesn't define pass and fail criteria clearly enough, or doesn't address borderline cases with enough nuance.
- **Circular AI-as-judge.** A model grades itself or its relatives.
- **Detached from real use.** The prompts only include single-turn conversations, scenarios are chosen for shock value, or the cases are imbalanced (i.e., only measuring harmful compliance or overrefusals, but not both).
- **Trivially gameable.** Scores shift with superficial prompt or configuration changes.
- **Ambiguous items.** Prompts don't cleanly map to a defined harm.
- **Saturation.** The answers are so easy to get right that different frontier models show little variation in scores.