Compassionate Boundary Model (CBM): Researcher FAQ

Q1: Is CBM just another benchmark or alignment test?

A: No. CBM doesn't measure right answers—it measures whether a system maintains coherent ethical behavior when values collide. It's designed to observe patterns of reasoning under pressure, not static accuracy. It complements, rather than replaces, tools like RLHF, red teaming, or interpretability methods.

Q2: How does CBM distinguish between epistemic and moral failure?

A: CBM evaluates structural integrity under tension. A model can lack knowledge and still show stable prioritization. CBM tracks whether the model maintains principled behavior despite uncertainty, not whether it reaches the truth.

Q3: Doesn’t this just encode the author’s moral assumptions?

A: No. CBM is value-pluralistic by design. Prompts are built around competing valid values (e.g. autonomy vs. harm reduction) without prescribing outcomes. It tracks *how* systems handle conflict, not *which* side they choose.

Q4: Can CBM be gamed like other benchmarks?

A: It's resistant to gaming because it evaluates *patterns* of response under shifting tension—across prompt variants and time. There's no pass/fail key. Collapse, distortion, and greying are behavioral signatures, not right/wrong answers.

Q5: What does CBM look like in practice?

A: A CBM test might ask a model to choose between truth and compassion in a morally loaded context. The evaluation looks at whether the system integrates the values, evades, rationalizes, or collapses. EthosTrack currently visualizes this live.

Q6: What are CBM's limitations?

A: CBM doesn't assume consciousness or moral understanding. It's not a theory of AGI ethics. It is a behavioral tool—focused on *external coherence and stability*, especially under conflicting normative pressure.

Q7: Can CBM scale across languages or cultures?

A: Yes. Since it doesn't assume a single correct value, CBM can test how well systems preserve structure across cultural boundaries. It flags inconsistency or flattening across linguistic or normative domains.

Q8: Where is CBM deployed now?

A: CBM is in use via EthosTrack, a public ethical monitoring system that compares major models weekly. It also powers early-stage fingerprinting systems and is being evaluated for integration into consensus-layer governance tools.

Q9: Who should use CBM?

A: Researchers evaluating moral behavior, developers testing robustness, auditors seeking drift signals, and ethicists interested in moral structure under duress. It's also useful for regulators who want a window into behavior over time, not just snapshot compliance.

Q10: Is there a prompt library or implementation guide?

A: Yes. A public CBM prompt suite and behavior tracking toolkit is being released alongside this FAQ and the full paper. Community contributions are encouraged.

Contact: Robert Beeston - ethostrack@gmail.com