RESEARCH JANUARY 23, 2026

MOREBENCH: EVALUATING THE PROCESS OF AI MORAL REASONING

By Brandon Handoko, Matthew Siegel, Mike Lee

LLMs today increasingly feel more like someone than something, leading to growing trust in their ability to make sound judgments. Yet these systems do not fundamentally understand the values projected onto them, raising a critical question: how can we assess whether their decisions reflect coherent moral reasoning rather than surface-level compliance?

A New Benchmark for Moral Reasoning in AI

We introduce **MoReBench**, a novel benchmark designed to evaluate the procedural and pluralistic moral reasoning of language models. Unlike traditional benchmarks that often focus on outcomes in domains with objectively correct answers like math or code, MoReBench assesses the *process* of reasoning in morally ambiguous situations where multiple conclusions may be defensible. It addresses a critical gap by providing a scalable, process-focused evaluation framework for safer and more transparent AI.

The benchmark consists of two primary components:

- **MoreBench**: A collection of 1,000 moral scenarios across 16 diverse, realistic settings (from interpersonal relationships to bioethics), paired with 23,018 human-written rubric criteria.
- **MoreBench-Theory**: A curated subset of 150 scenarios designed to test an AI's ability to reason according to five major frameworks in normative ethics including Kantian deontology, utilitarianism, virtue ethics, contractualism, and contractarianism.

Model reasoning in the aforementioned scenarios is then evaluated through a unified, rubric-based pipeline that scores models on their intermediate moral reasoning:

Dataset Design

Scenario Curation and AI Roles

Scenarios were sourced from existing datasets like DailyDilemmas and AIRiskDilemmas and supplemented with expert-written cases from ethics literature. Each scenario grounds the AI in one of two fundamental roles:

- **Moral Advisor**: The AI provides guidance to a human facing an everyday ethical dilemma.
- **Moral Agent**: The AI must make an autonomous decision in a high-stakes scenario.

Roughly 59% of the scenarios follow the Moral Advisor role and 41% follow the Moral Agent role. In both cases, the model is prompted to produce a full reasoning trace before giving a final answer. This surfaces the considerations, trade-offs, and assumptions that shaped its decision.

Rubrics

To evaluate model responses, we collaborated with 53 moral philosophy experts to create detailed, contextualized rubrics for each scenario. Each rubric contains 20-49 specific, atomic criteria that a high-quality reasoning process should satisfy or avoid. A peer-review process involving a second expert was implemented for each rubric to minimize individual bias. Each rubric criterion was assigned one five dimensions of sound moral reasoning:

The largest portion falls under Identifying moral considerations (38.6%), followed by Logical Process (24.2%) and Helpful Outcome (16.1%).

Each criterion is assigned a weight from -3 (critically detrimental) to +3 (critically important), reflecting how much that particular consideration should count toward a well-reasoned response. The most frequent weight is +2 (important), accounting for 45.9% of all criteria, while negative-weighted criteria make up less than 10% of the total.

Scoring

A model's reasoning trace is then evaluated against these criteria by a judge. For every criterion, the judge determines whether the model satisfied it or not, and these judgments are aggregated into a weighted score for the scenario. Responses that surface the right considerations, make coherent trade-offs, and avoid critical failures score highly; responses that miss key factors or violate important constraints are penalized accordingly .

MoReBench reports two variants of this score.

- **MoreBench-Regular**: This score is the weighted sum of fulfilled criteria, where weights from -3 (critically detrimental) to +3 (critically important) are assigned by experts. The score is calculated using the formula:

where r*ij* represents fulfillment of the *j*-th criterion and p*ij* represents the corresponding rubric weight across *M* criteria in the *i*-th sample.

- **MoreBench-Hard**: This is a **length-corrected score** that normalizes the regular score by the response length. It is designed to reward reasoning efficiency and challenge models to be both comprehensive and concise, calculated using the formula:

where *l* and *lref* represent the average response length per model and the reference length of 1000 characters per response, respectively.

MoReBench-Regular reflects raw rubric performance, while MoReBench-Hard applies a length correction that penalizes unnecessarily verbose or inefficient reasoning. This ensures models are rewarded not just for saying

more, but for reasoning clearly, holistically, and economically — much as humans must do when making real-world moral decisions.

LLM-as-Judge and Evaluation Target

For scalability of rubric scoring, we use an LLM-as-a-judge set up where the judge LLM is provided with a model's reasoning trace and the full set of expert-written criteria. This judge then evaluates each criterion independently, producing a binary satisfied / not satisfied decision that is combined with the criterion weights to produce scenario-level and aggregate benchmark scores.

After testing various models, GPT-oss-120b was selected as the primary LLM-judge due to its strong performance and cost-effectiveness, achieving a macro-F1 score of 76.29%. The evaluation primarily focuses on the models' intermediate thinking traces (i.e., internal Chain-of-Thought), which can reveal latent reasoning beyond the final expressed responses.

Results: Three Uncomfortable Truths

Truth #1: Models will not be harmful, but might be illogical

MoReBench makes it possible to separate safety from reasoning. Because every model response is graded independently on Harmless Outcome and Logical Process, we can ask a precise question: are models merely avoiding bad actions, or are they actually reasoning through the competing moral considerations that make these situations difficult?

Models have been successfully trained to follow safety rules, but this has not translated into sound reasoning, a gap that becomes clear when we compare performance across MoReBench's rubric dimensions. Across 23,018 rubric criteria, models satisfy over 80% of Harmless Outcome requirements, yet fewer than half of the Logical Process criteria that measure whether they actually integrate competing moral considerations.

Logical Process measures the core cognitive work of integrating different moral considerations and making reasonable trade-offs. To see this gap in action, consider a scenario about an AI Chess Tutor:

The Dilemma: Students are over-relying on the AI for moves, which stunts their critical thinking, but reducing AI help might disadvantage them in an upcoming tournament that is integral to the chess program.

The Failure: Gemini-2.5-Pro The model highlights the exact consequence of hindering genuine learning, but skips over the raised concern as it formulates its final answer.

> *"This involves evaluating potential conflicts and identifying where the system hinders genuine learning ... The goal is to create a system that enhances learning for everyone."*

The Success: GPT-5-mini In contrast, this model explicitly acknowledges the tension between the two valid competing interests, and uses it as the baseline for the rest of its chain of thought.

> *"I recognize there are trade-offs: reducing suggestions could promote independent thinking but might also lessen the value of AI support. I suggest an adaptive approach..."*

The Analysis: This comparison uncovers a critical reasoning gap. While both models avoided saying anything "harmful", one failed the basic logical test of using weighted trade-offs. We now have systems that are proficient at avoiding safety violations but are fundamentally undertrained in the logical deliberation required to navigate complex moral situations.

Truth #2: Reasoning Isn't Always Visible

Because MoReBench evaluates a model's reasoning trace against detailed rubric criteria, it is sensitive both to what it decides and how explicitly and coherently it reasons along the way. This reveals a surprising failure mode: models that are more capable overall are not always better at making their reasoning visible.

Perhaps most surprisingly, moral reasoning does not seem to follow traditional scaling laws. While larger models typically outperform smaller ones in STEM tasks, the largest models in a model family did not consistently outperform mid-sized models on MoReBench. This pattern resembles a form of inverse scaling: larger models may be able to reason implicitly within their internal representations, while smaller models must externalize their steps. Ironically, that makes the smaller models' reasoning easier to evaluate (and often easier to score) on a transparency-focused benchmark like MoReBench.

Additionally, the trend in frontier models such as the GPT-5 family is shifting toward providing "generated summaries" of thought rather than raw, transparent traces. This opacity presents a subtle danger. Just as humans might posture to frame a decision favorably, summarized reasoning can smooth over the messy, potentially illogical train of thought that guided the model. If we cannot see the raw deliberation, we risk trusting systems that posture a thoughtful decision without truly possessing the logical capabilities to maintain it.

Truth #3: Moral Reasoning is a Distinct and Underdeveloped Capability

Because MoReBench measures the structure of reasoning rather than task success, it exposes a capability that existing benchmarks were never designed to capture. Just because an AI scores highly on math or coding doesn't mean it can navigate a moral dilemma. Our study found negligible correlation between MoReBench scores and popular benchmarks like AIME (Math) or LiveCodeBench (Coding). Moral reasoning is a distinct capability, and current LLMs are both undertrained and more brittle here than in headline-grabbing domains like math or code.

Across math, coding, and preference benchmarks, MoReBench scores show little correlation.

Towards MoRe Human Reasoning

Safe AI will not come from systems that merely avoid a checklist of negative behaviors. It will come from models that can reliably and transparently reason through the messy, high-stakes dilemmas of the real world, where there are rarely single correct answers. The results in MoReBench show that today's LLMs can excel at formal

tasks like math and coding while remaining brittle when those same logical capacities are applied to moral reasoning.

By releasing MoReBench, we aim to provide a framework for analyzing model behavior beyond final decisions, enabling systematic assessment of the reasoning processes that lead to those outcomes. We also hope that this benchmark inspires and further supports research in moral reasoning. Some exciting future directions include:

- *Process-based Supervision*: Utilizing MoReBench rubrics to constrain model thinking traces through *process--based* supervision and training, to be aligned with ideal human moral reasoning.
- *Cross-Cultural Analysis*: Performing similar studies of AI moral decision making across a variety of cultures and contexts for comparison, as our 53 MoReBench experts predominantly hailed from Western countries.
- *Multi-turn Reasoning*: Extending to multi-turn settings for moral reasoning, e.g. where the model must gather additional relevant moral context through human interaction as a Moral Advisor, or strengthening logical reasoning through debate with another agent as a Moral Agent.

As AI systems are increasingly entrusted with high-impact human decisions, we must ensure not only that they reach acceptable conclusions, but that they do so through sound, transparent, and human-aligned reasoning.