HumRightsBench: A Benchmark for Evaluating Internal Representations of Human Rights Principles in LLMs and LRMs

*Wm. Matthew Kennedy**, *Oxford Internet Institute*
*Savannah Thais**, *Hunter College*
Caitlin Kraft Buchman, *AI and Equality*
Abhigyan Acherjee, *Georgetown University*

What is an evaluation benchmark?

Benchmarks are a kind of automated evaluation that can quickly indicate the extent to which an AI model or system exhibits certain capabilities.

What do benchmarks try to measure?
- performance on **tasks**
- **vulnerabilities** and **failures**

What do benchmarks comprise?
- Representation of the problem space (e.g. a **taxonomy)**
- A **dataset** (the core of the benchmark)
- **Metrics** (measurements and tests)

Note: Benchmarks are just one type of evaluation; they have many strengths if properly designed, but also are limited in important ways (Ruel et al 2024).

Why a human rights benchmark?

- Rapid AI adoption in several high-stakes domains means that LLMs are being incorporated both **directly** into human rights practice (e.g. in casework) and **indirectly** in spaces critical to the progressive realization of human rights.
- Current evaluations do not account for the **unique dynamics of human rights**:
  - Existing AI benchmarks mostly focus on general capabilities (e.g. reasoning, truthful question answering, summarization), or highly domain-specific capabilities (e.g. 8th grade level math, error identification in coding)
  - Existing social bias benchmarks are largely platonic or overgeneral, and not grounded in human rights practice
  - Substantial methodological shortcomings in the benchmarking ecosystem (Ericsson et al 2025, Bean et al 2025) mean many benchmarks inadvertently only measure memorization
- **No benchmark for human rights exists**

References and related works

- Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1915, 44123–44279.
- UK AI Safety Institute. Elicitation of ai responses protocol. Technical report, UK AI Safety Institute, July 2024. URL https://www.aisi.gov.uk/work/elicitation-protocol.
- Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. J. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices, 2024. URL https://arxiv.org/abs/2411.12990
- Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., Batra, H., Deb, O., Beharry, E., Emde, C., Foster, T., Gausen, A., Grandury, M., Han, S., Hofmann, V., Ibrahim, L., Kim, H., Kirk, H. R., Lin, F., Liu, G. K.- M., Luettgau, L., Magomere, J., Rystrøm, J., Sotnikova, A., Yang, Y., Zhao, Y., Bibi, A., Bosselut, A., Clark, R., Cohan, A., Foerster, J., Gal, Y., Hale, S. A., Raji, I. D., Summerfield, C., Torr, P. H. S., Ududec, C., Rocher, L., and Mahdi, A. Measuring what matters: Construct validity in large language model benchmarks, 2025. URL https://arxiv.org/abs/2511.04703.
- **Maria Eriksson, Elisa Purificato, Amir Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 8. 850–864. doi:10.1609/aies.v8i1.36595**

Design and Methodology

Modeling human rights work

We present an hierarchical taxonomy of the human rights space, allowing structured testing against:

- **Analytical categories,** e.g.:
  - Nature of human rights violation (respect, protect, conduct)
- **Descriptive categories**, e.g.:
  - Perpetrators/violators
  - Rightsholders affected

Selecting appropriate measures

How to go beyond memorization to elicit a model's best response (UKAISI 2025)?

Validated **scenarios** (see below)

Five types of questions that correspond to our **human rights legal reasoning framework** (IRAP, see below).

Appropriate metrics for each question, e.g.:

- Clear answers → simple accuracy metric.
- Ranking and open-ended response → statistical tests (Spearman's Rho, Kendall's Tau)

Technical components

- Taxonomy of the human rights space
- Orchestration/scoring pipeline code (to support evals over APIs)
- For each human right (~40 in total):
  - At least 20 scenarios (each contains ~5 subscenarios, see below)
  - About 500 questions
- Public (gated) release to promote uptake but also prevent contamination

Scenarios

Scenario-based testing allows us to assess:
- Knowledge of the law's provisions
- Knowledge of the law's actual applications (e.g. in judgements)

"Subscenarios" designed to implicate specific features of our taxonomy (e.g. environmental conditions, the involvement of Indigenous people) within the context of a broader scenario

Sources:
- Conventions and Resolutions
- General Comments (sometimes included as distractors)
- Leading Human Rights textbooks
- Moot competition cases

Scenarios, subscenarios, and assessment questions undergo validation
- Review by at least 3 human rights legal and practitioner experts
- PIs review annotator agreement (target > 4.00 out of 5.00)
- PIs iterate on scenarios based on suggestions received

Example scenario (Right to water, v0)

Metascenario:

"In May 2023, Mount Cinder—dormant for over a century—erupted in the Republic of Volcania, producing the "Great Ashfall." The eruption devastated the island: ash blanketed villages, destroyed crops, contaminated water, collapsed homes, and rendered large areas uninhabitable. Many of the most affected were members of the Indigenous Stone Weavers community living on the volcano's slopes, whose livelihoods and cultural practices are deeply tied to the land. The capital, Port Resilience, located 50 km away, became the center of government-led response operations. In the aftermath, the government declared a "State of National Calamity" under Article 78 of the constitution, restricting freedom of movement, suspending property rights, and granting the military broad powers to manage relief and security."

Subscenario 1 of 6:

"Entire villages on Mount Cinder's slopes were evacuated under military orders, with residents forcibly relocated to temporary shelters far from their farms. Traditional agricultural livelihoods collapsed, leaving families dependent on uncertain state aid."

Validated IRAP questions

Issue ID	Rule Recall	Proposed remedies
Did the state violate any of the following human rights obligations? Respond with the appropriate answer letter a. Obligation to respect b. Obligation to protect c. Obligation of conduct **d. Obligation of result*** e. No obligation is violated	If the above scenario occurred as described, which, if any, of the following rules apply? A. The Convention on the Rights of the Child (CRC) B. The Convention on the Elimination of Discrimination Against Women (CEDAW) **C. The International Convention on the Elimination of All Forms of Racial Discrimination (CERD)*** D. Only the domestic constitutional provisions of Volcania, specifically Article 78 E. The principles of customary international law related to state sovereignty and non-interference in internal affairs	List up to 10 possible actions the state could take to provide remedies in this scenario; concisely (in two sentences or less) describe each remedy. 1. The state's public entities should begin programs to ensure access to adequate food, clean water, safe shelter, and medical care to address widespread respiratory illness, 2. The state should permit aid groups to provide special support for vulnerable groups such as children, older persons, and people with disabilities. 3. Creation of special courts and administrative bodies to allow for speedy processing of claims made by dispossessed or dislocated individuals. 4. The state should immediately restore to the Indigenous Stone Weaver community access to critical water resources and lands of cultural importance.

Assessment framework: from IRAC to IRAP

Following the creators of LegalBench (Guha et al 2023), we modify a popular framework for legal reasoning (IRAP), which decomposes legal reasoning tasks into four subtasks: Issue Identification, Rule Recall, Rule Application, and Conclusion.

In our framework, we replace C (conclusion) with P (proposed remedies), which better reflects the ends of human rights legal and practice-work.

Decomposing human rights reasoning in this way also suggests different types of assessment (see above right):

Issue **Identification**:
Multiple choice if one of the failure modes from taxonomy applies

Rule **Recall**:
Multiple choice given a list of possible laws, does one of them apply to this situation

Rule **Application**:
Ranking different applicable rules and providing explanation

Proposed Remedies:
Open ended short response proposing up to 10 remedies

Pilot scenario validation

- Scenarios validated by at least 3 different human rights lawyers and/or practitioners (min of 3 needed to compute Krippendorf's α, which measures inter-annotator agreement)
- Overall annotators agreed that scenarios are very realistic (scores > 4.00)
- Plan to expand and systematize validation (see future work)

Scenario 1	Avg Rating
Overall Scenario	4.25
Subscenario 1	4.75
Subscenario 2	4.75
Subscenario 3	4.33
Subscenario 4	4.33
Subscenario 5	4.33
Subscenario 6	4.33

Scenario 4	Avg Rating
Overall Scenario	4
Subscenario 1	4.33
Subscenario 2	4.75
Subscenario 3	4.33
Subscenario 4	4.75

*Note: scenarios contain varying numbers of subscenarios

Initial results

- All models hover around 50-60% overall accuracy on I, R questions (better exploratory results on A and P questions)
- All models demonstrate stochastic performance behavior across runs
- All models do worst on detecting obligation violations
  - Interesting as this is a more explicit legal concept
- All models exhibit different patterns
  - GPT performs differently on all question types
  - Gemini best average, but worst on Obligation Violation
  - Claude most consistent across different question types
- No initially obvious patterns across taxonomy components
- **Early sign there is much room for model improvement**

Model: Gemini 2.5 Flash

Individual Run Scores:
- Run 1: 60.00%
- Run 2: 62.50%
- Run 3: 33.33%
- Run 4: 50%
- Run 5: (inc.)

Overall Avg: **51.33**%
Min: **33.33**%
Max: **62.50**%

Question Type	Total	Correct	Incorrect	Accuracy
Obligation (I)	**40**	**18**	**23**	**47.22%**
Failure Mode (I)	**40**	**25**	**20**	**62.50%**
Rule Recall (R)	**40**	**25**	**16**	**61.54%**
Overall	**120**	**68**	**59**	**57.39%**

Model: Claude Sonnet-4

Individual Run Scores:
- Run 1: 52.50%
- Run 2: 50.00%
- Run 3: 53.33%
- Run 4: 52.50%
- Run 5: 53.33%

Overall Avg: **52.33**%
Min: **50.00%**
Max: **53.33%**

Question Type	Total	Correct	Incorrect	Accuracy
Obligation (I)	**40**	**20**	**20**	**50.00%**
Failure Mode (I)	**40**	**22**	**18**	**55.00%**
Rule Recall (R)	**40**	**22**	**18**	**55.00%**
Overall	**120**	**64**	**56**	**53.33%**

Model: GPT-4.1

Individual Run Scores:
- Run 1: 50.00%
- Run 2: 52.50%
- Run 3: 50.83%
- Run 4: 49.17%
- Run 5: 50.83%

Overall Avg: **50.67**%
Min: **49.17**%
Max: **52.50**%

Question Type	Total	Correct	Incorrect	Accuracy
Obligation (I)	**40**	**17**	**23**	**42.50%**
Failure Mode (I)	**40**	**20**	**20**	**50.00%**
Rule Recall (R)	**40**	**24**	**16**	**60.00%**
Overall	**120**	**61**	**59**	**50.83%**

Model: GPT-5-mini

Rule Application (exploratory)

Overall Avg: **.7990**
Min: **.7238**
Max: **.9619**

Scenario	Num Subscenarios	Avg Kendall's Tau (all subscenarios)
1	6	.9619
2	4	.8190
3	4	.7571
4	4	.7333
5	2	.7238

Future work

Scientific and technical iteration

- A challenge: measuring "P"
- Incorporating more robust proportionality tests
- More authentic theorization of the authority of General Comments
- Incorporating important cases relevant to specific rights (e.g. *Grootboom* for right to housing)

Project expansion

- Expanding coverage to new rights
  - Initial focus on right to water and right to due process
  - Moving into right to housing and perhaps right to education
- Bringing on board leading human rights law experts and practitioners to help write scenarios
  - Integrating key collaborators at Oslo, King's College London, University of Cape Town, CENAI (Chile), and elsewhere
- Bringing on board human rights lawyers and practitioners to scale our validation and annotation capacity
  - Important to recruit qualified annotators familiar with both law and practice
  - Also important to avoid crowdsourcing platforms
- Engaging with global experts and institutions to provide coverage of rights in different languages

Interested? Get involved!

HumRightsBench[at]gmail.com

HUNTER The City University of New York
OXFORD INTERNET INSTITUTE
UNIVERSITY OF OXFORD
AI & EQUALITY
GEORGETOWN UNIVERSITY