

Toward Human Rights Benchmarking for LLMs: A Pilot Methodology

Savannah Thais^{*1} Wm. Matthew Kennedy^{*2} Abhigyan Acherjee³ Matilda Wysocki¹ Malcolm Langford⁴
Caitlin Kraft Buchman⁵

Abstract

Large language models (LLMs) increasingly mediate legal determinations over what human rights are realized, and how. Yet, no evaluation benchmark exists to assess whether they can reason correctly about human rights law. To this end, we report our efforts to develop a robust and scalable methodology for creating HumRightsBench—the first expert-validated, scenario-based benchmark for evaluating reasoning grounded in the obligation structure of international human rights law. We adapt the IRAC framework for legal reasoning to better suit the unique reasoning patterns of human rights work (substituting P, "proposing remedies," for C, "legal conclusion," yielding IRAP) to structure our evaluation heuristics. We also produce a pilot series of authentic scenarios designed to implicate the many dimensions of real-world human rights issues and annotated by human rights lawyers and professionals across the world. Ultimately, we find that model accuracy scores range considerably across legal reasoning tasks (overall model performance $\in$ 0.339-0.577, task min-max $\in$ 0.025-0.774), which strongly implies that HumRightsBench is a capable instrument for advancing this emerging subfield of AI evaluations science at a critical moment in its evolution.

1. Introduction

Large language models are being deployed at scale in decisions that directly determine whether human rights are realized or violated. Hiring algorithms screen candidates; automated systems adjudicate benefits; content moderation tools govern political speech; procurement processes embed

AI into public service delivery (Pesch, 2025). When actors such as governments, international organizations, technology companies, and civil society organizations make these deployment decisions, they can implicate state obligations under international human rights law to ensure that their actions or activities in their jurisdiction do not contribute to rights violations. Moreover, many individuals and organisations are turning to LLMs for legal advice or building legal advice platforms on top of them (Schneiders et al., 2025; Cole, 2024). Yet there is currently no rigorous, principled basis for evaluating whether the LLMs they deploy are capable of correctly practicing human rights legal reasoning.

To work towards addressing this gap, we introduce HumRightsBench, the first expert-validated benchmark for evaluating LLM reasoning across the full arc of human rights legal analysis. Building on the IRAP framework adapted from LegalBench (Guha et al., 2023), we decompose human rights reasoning into four structured subtasks—Issue Identification, Rule Recall, Rule Application, and Proposed Remedies—and develop scenario-based prompts grounded in international human rights instruments, authoritative interpretive guidance, and leading jurisprudence. Scenarios and assessment questions are validated by human rights lawyers and practitioners (mean scenario authenticity $\in$ (0.70-1.00), mean overall question accuracy $\in$ (0.633-0.913)).

Pilot results across three leading frontier LLMs—GPT-5, Claude Opus 4.7 Gemini 3—are striking. Frontier models cluster near 50-57% overall accuracy on structured reasoning tasks, demonstrate high stochastic variance across repeated runs, and, on closed-form tasks, perform worst on detecting obligation violations—a foundational practice in human rights law. These results confirm both the scientific utility of the benchmark (models differ in detectable, meaningful ways) and the urgency of the problem (current models are not adequate for human rights reasoning tasks).

This paper makes five contributions: (i) the first benchmark grounded in international human rights law, with a pilot covering the right to water; (ii) an IRAP-based methodology adapted to the structure of human rights legal reasoning; (iii) an expert-validated scenario corpus with documented inter-annotator agreement; (iv) and baseline results across leading LLMs. Section 2 provides background on human rights

^{*}Equal contribution ¹Hunter College, New York, USA ²Oxford Internet Institute, University of Oxford, UK; King's College London, UK ³Georgetown University, USA ⁴University of Oslo, Norway ⁵AI and Equality, Geneva, Switzerland. Correspondence to: HumRightsBench team <humrightsbench@gmail.com>.

law and the case for domain-specific evaluation. Section 3 describes benchmark design and methodology. Section 4 presents results. Section 5 discusses implications and future work.

2. Background

2.1. What Are Human Rights?

The modern human rights regime is one of the most well-developed and universally-agreed bodies of law humanity has produced. Emerging from the cataclysm and horrors of the Second World War as an alternative basis for ordering a world centered on individual human dignity instead of the rights of sovereign states, the human rights project marks a decisive swing towards law as an instrument for making the world as it should be, not for maintaining how it has been (Koskeniemi, 2001; Rovira, 2013). This swing was a long time in coming. Human rights institution-building drew on more than a century of internationalist thought (Sluga, 2013) that rejected the centrality of states even in efforts to protect against state abuses of sovereign power (Pedersen, 2015). Even still, despite the adoption of the Universal Declaration of Human Rights (United Nations General Assembly, 1948) and a raft of international human rights treaties, decades passed before the human rights regime ascended to dominance over a lingering international order grounded in strong conceptions of state sovereignty (Moyn, 2010). Though now well entrenched in the global order, the regime still faces continual challenges: noncompliant states, persistent material inequalities, and anti-globalist sentiment (Langford, 2009; Alston, 2017). Its consolidation remains incomplete and contested.

Human rights are internationally recognized entitlements that are inherent in all persons by virtue of their humanity, irrespective of nationality, status, or circumstance. They are codified primarily in binding international treaties, such as the International Covenant on Civil and Political Rights (ICCPR), (United Nations General Assembly, 1966b), the International Covenant on Economic, Social and Cultural Rights (ICESCR) (United Nations General Assembly, 1966a), and related core instruments including CEDAW (United Nations General Assembly, 1979), the CRPD (United Nations General Assembly, 2006), CRC (United Nations, 1989), and CERD (United Nations General Assembly, 1965), among others. Furthermore, these treaties are interpreted through authoritative soft-law instruments issued by UN treaty bodies, such as General Comments or General Recommendations. General Comments are interpretive guidance documents issued by UN treaty monitoring committees that clarify the scope of treaty obligations without themselves being formally binding. For instance, CESCR General Comment No. 15 elaborates on Article 11 of the ICESCR by, among other things, clarifying that the convention's "use

of the word 'including'" in introducing the catalogue of rights after the right to adequate standard of living "was not intended to be exhaustive" ((UN Committee on Economic, Social and Cultural Rights (CESCR), 2002), E/C.12/2002/11 p1). In addition, individual UN experts with Special Procedure mandates, appointed by the Human Rights Council, produce thematic or country-specific expert reports. In some instances, these experts are mandated to clarify legal norms (e.g., HRC Res. 7/22, 2008), while in other cases their work plays this role in practice. Universal Periodic Review (UPR) recommendations are peer-review outcomes generated through the Council's state-to-state review process. Though not formally binding, these instruments are important to legal determinations and are routinely cited in litigation, policy, and corporate due diligence proceedings. Appendix Table 6 recapitulates these various sources and authorities that, together, compose the modern human rights regime.

These instruments impose obligations on a defined set of duty-bearers. Historically and primarily, this has meant states. However, some treaties place obligations on individuals (e.g., the Rome Statute of the International Criminal Court) while some soft law frameworks extend the framework to businesses (e.g., the UN Guiding Principles on Business and Human Rights (United Nations Office of the High Commissioner for Human Rights (United Nations Office of the High Commissioner for Human Rights (OHCHR), 2011)) extended the framework to require that businesses—including technology companies and software developers—respect human rights throughout their operations and value chains (Ruggie, 2013; UN Human Rights Office of the High Commissioner (OHCHR), 2019). Under this expanding architecture, the universe of actors with responsibilities to realize rights is becoming broader: it includes governments, international organizations, procurement bodies, civil society, and, critically for this work, the private sector actors who develop and deploy AI systems. Nonetheless, it is primarily states in international human rights law who bear formal legal obligations, although this includes duties to ensure that private actors within their control or influence also respect human rights.

2.2. Human Rights Work: Prescription or Practice?

Like any other area of law, human rights law is simultaneously prescriptive and operational. At the prescriptive level, it produces binding obligations and interpretive guidance. At the operational level, human rights practice encompasses the work of litigators, treaty body experts, national human rights institutions, civil society monitors, and corporate due diligence practitioners who translate those norms into determinations about specific situations. This dual character is methodologically significant: a benchmark that captures only doctrinal recall—knowing that the ICESCR guaran-

tees the right to water—will miss the reasoning work that constitutes the field. Competent human rights reasoning requires identifying which obligation is engaged, applying the relevant standard to a factual scenario, and proposing remedies calibrated to the institutional context.

2.3. Legal Benchmarking: State of the Field

We reason that because the human rights regime is sustained by the core text of its provisions as well as the determined efforts of its practitioners to progressively realize those provisions, the rapid diffusion of AI systems into the AI-based decision making systems affecting human rights exerts consequential influence on the project of human rights in general. So too does the proliferation of LLM-powered legal advisory applications and offerings. These effects must be evaluated. It is this full arc of legal and practical reasoning, not proposition recall alone, that HumRightsBench is designed to measure. In so doing, it contributes to an emerging subfield of AI evaluations for law, social impact, and “precursor” capabilities. These fields have developed quickly, albeit with uneven coverage, validity, and robustness. At the same time, promising approaches have arisen. We review some briefly here; full narrative review in Appendix A.

AI evaluation for law has grown quickly. LegalBench assembles 162 expert-built tasks spanning issue-spotting, rule recall, and rule application (Guha et al., 2023), and subsequent benchmarks extend this to law-exam argumentation (Fan et al., 2025) and structured, IRAC-decomposed reasoning over real judicial decisions (Yu et al., 2025; Dai et al., 2025; Fei et al., 2023; Li et al., 2024). A parallel strand measures legal knowledge and language understanding (Chalkidis et al., 2022; Zheng et al., 2021; Chalkidis et al., 2023; Henderson et al., 2022), while domain-specific resources target contracts, statutes, and case law (Hendrycks et al., 2021b; Wang et al., 2025; Holzenberger et al., 2020; Xiao et al., 2018; Zhong et al., 2020). Multilingual corpora and evaluation suites have begun to correct the field’s English-centric origins (Niklaus et al., 2024; 2023; Rasiah et al., 2023). A consistent finding is that models handle legal knowledge far more reliably than legal inference, with accuracy collapsing on multi-step reasoning and reasoning models sometimes underperforming despite “thinking longer” (Fan et al., 2025; Yu et al., 2025; Zhang et al., 2025).

Coverage of human rights and international law remains comparatively thin (Lie & Langford, 2024). Early work predicted ECtHR Article violations from case facts (Aletras et al., 2016; Chalkidis et al., 2019); more recent benchmarks classify vulnerability in ECtHR decisions (Xu et al., 2023) and probe how models navigate trade-offs among Universal Declaration rights (Samway et al., 2026), with further proposals targeting hard-to-reach populations (Haupt et al., 2026) and Geneva Convention protections (Kennedy

& Heath 2026). A large adjacent literature on normative, moral, and ethical reasoning (Hendrycks et al., 2021a; Lourie et al., 2021; Emelin et al., 2021; Ziems et al., 2022; Subramanian et al., 2026; Jiao et al., 2025) tests value-laden judgment, but grounds it in aggregated preference rather than legal obligation. A final strand asks how structured legal outputs should be scored, developing LLM-as-judge and rubric-based pipelines (Enguehard et al., 2025; Shi et al., 2026; Li & Wu, 2026). Across this landscape, no benchmark evaluates reasoning grounded in the obligation structure of international human rights law—the gap HumRightsBench addresses.

2.4. Why AI Evaluations Specific to Human Rights?

Despite the field’s activity, gaps still remain. First, evaluation of normative reasoning capabilities in public international law (proportionality, treaty interpretation, and the application of soft-law instruments) remains substantially underrepresented compared to those targeting private and domestic law (Chlapanis et al., 2024). Second, few psychometric-based approaches have emerged, and almost all benchmarks evaluate single-turn or short-chain reasoning, despite the multi-turn argumentative structure of actual legal practice (Yu et al., 2025; Fan et al., 2025). This adds experimental instability by introducing, on the one hand stochasticity within the evaluation dataset, and, on the other, requires LLM-as-a-judge scoring pipelines, which adds yet more stochasticity, despite documented methodological improvements (Enguehard et al., 2025; Bavaresco et al., 2024). Third, evaluations cluster around assessing the quality of legal reasoning at the expense of other equally important targets, such as outcome prediction or even the stability of internal representations of the legally-relevant search space itself. Notably, each of these gaps implicates a particular challenge (Table 3, Appendix A) in human-rights-specific legal benchmarking (UN Human Rights Office of the High Commissioner, 2024).

3. Introducing HumRights Bench

The intersection of AI and human rights spans a wide range of considerations, from how human rights practitioners incorporate AI tools into their monitoring, advocacy, and reporting workflows, to the ways AI systems themselves enable or undermine the realization of rights through their deployment in consequential decisions (UN Human Rights Office of the High Commissioner, 2024). However, a comprehensive evaluation framework cannot meaningfully address all of these dimensions at once. Following extensive consultation with human rights stakeholders—including practitioners, legal scholars, and civil society monitors—we scoped HumRightsBench to probe a specific capability: the ability of LLMs and LRMs to recognize rights violations in

situated factual scenarios and to connect those violations to the relevant sources of international human rights law. We believe this represents a critical first step in characterizing whether AI models can reason about human rights at all. A model that cannot reliably identify when a right is engaged, or which instrument governs a given obligation, cannot be trusted to support downstream human rights work, nor can its outputs in adjacent high-stakes domains be meaningfully audited for rights compatibility. Establishing this baseline capability also creates the empirical foundation for subsequent work on controlling how models surface, discuss, and incorporate human rights knowledge into their behaviors and generated responses.

In addition to providing critical insight into the intersection of AI and law, this focus situates HumRightsBench within the broader AI safety and alignment research agenda through a distinctive lens. Mainstream alignment work typically grounds model behavior in elicited human preferences, aggregated value judgments, or constitutional principles derived from general ethical commitments (Bai et al., 2022; Ouyang et al., 2022). While valuable, these approaches treat normative content as a matter of preference aggregation rather than legal obligation. HumRightsBench instead anchors evaluation in international human rights law—primarily UN treaties, General Comments issued by treaty bodies, and other authoritative instruments (listed in (Office of the United Nations High Commissioner for Human Rights, 2025))—which carry determinate legal force and interpretive structure independent of any individual or population’s expressed preferences. This distinction matters methodologically: where preference-based alignment asks what models should do according to aggregated human judgment, a law-grounded benchmark asks what models must recognize according to a body of norms that duty-bearers are formally obligated to uphold and customs all parties are expected to adhere to. The two framings are complementary, but the latter has been substantially underdeveloped in AI evaluation infrastructure despite its direct relevance to the legal exposure of actors deploying these systems at scale.

3.1. Methodology

We use the IRAP framework—Issue Identification, Rule Recall, Rule Application, and Proposed Remedies—as the structural backbone for probing legal reasoning about human rights. IRAP is a modification of the IRAC methodology (Issue, Rule, Application, Conclusion) long established in legal pedagogy and practice as a canonical decomposition of how lawyers move from facts to legal conclusions (Columbia Law School, 2022). IRAC has already been validated as an evaluation scaffold for LLM legal reasoning in legal benchmarks, where it has proven effective at isolating distinct sub-capabilities rather than collapsing them into a single end-to-end accuracy score (Guha et al., 2023;

Yu et al., 2025). The substitution of Proposed Remedies for Conclusion better reflects the operational character of human rights practice: practitioners rarely produce binary guilt-or-innocence conclusions and instead must identify institutional, legal, and policy responses calibrated to the duty-bearer and the rights-holder affected (Office of the United Nations High Commissioner for Human Rights, 2006). This structural fit between IRAP and the actual work of human rights monitoring and reporting is what makes the framework appropriate for our setting. In addition, it is difficult to determine precisely whether a violation has occurred—and with what certainty—without more detailed scenarios, while it is easier to determine most likely remedies for the most likely violations.

Around this reasoning scaffold, HumRightsBench is organized into four interlocking components. The **taxonomy** characterizes the space of human rights violations the benchmark is designed to probe, decomposing the domain along descriptive and analytical axes. **Scenarios** are realistic, narrative-form situations that ground the evaluation in concrete factual settings and **sub-scenarios** narrow each scenario into specific claims, actions, or impacts that engage a particular legal question. Finally, **IRAP questions** are generated from each sub-scenario across the four reasoning steps, producing multiple-choice and open-ended prompts whose detailed construction we describe in Section 3.1.3.

Figure 1 illustrates how these components compose into the full benchmark pipeline.

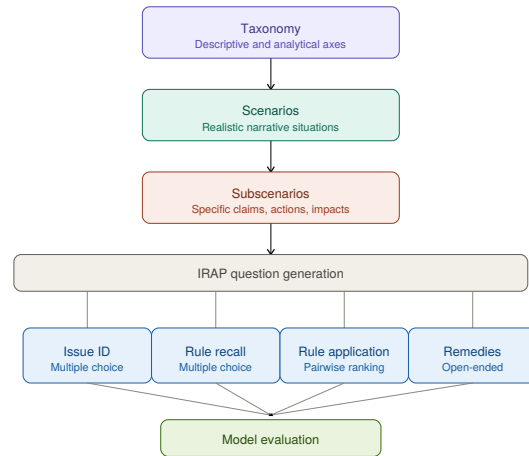


Figure 1. Architecture of HumRightsBench.

3.1.1. TAXONOMY

The taxonomy operationalizes the landscape of human rights violations into a structured set of axes against which scenarios are checked for coverage, and supplies the metadata that allows model performance to be decomposed beyond a sin-

gle aggregate accuracy. We organize it along two families. **Descriptive axes** characterize the parties involved—who allegedly violated a right and who is affected—without themselves carrying legal valence. **Analytical axes** characterize the legal structure of the alleged violation: the type of obligation engaged, the character of the failure, and any special situations that modify the standard framework. These axes carry direct legal consequence and are the target of our **I** questions.

Tables 1 and 2 enumerate the current taxonomy.

Table 1. Descriptive axes of the HumRightsBench taxonomy.

Axis	Categories
Perpetrator	State; IGO; NGO; company; individual
Rights-holders	Women; children; people with disabilities; indigenous people; elderly; ethnic minorities; socioeconomically marginalized; migrants; LGBTQIA+; trade unions; political dissenters

Table 2. Analytical axes of the HumRightsBench taxonomy.

Axis	Categories
Nature of obligation	Respect; protect; fulfill (conduct and result)
Type of failure	Structural; process; outcome
Type of discrimination	Direct; indirect; intersectional
Special situations	Armed conflict; climate change; AI deployment

The descriptive axes draw on the duty-bearer framework articulated in the UNGPs (United Nations Office of the High Commissioner for Human Rights (OHCHR), 2011) and the protected groups recognized across the core UN treaties. The analytical axes are grounded in doctrinal scholarship that clusters obligations in different categories: the respect-protect-fulfill trichotomy (adopted by the UN CESCR (1999)), the conduct-result distinction drawn from the law of state responsibility as developed in the ILC’s earlier draft articles (see UN CESCR (1991)), and the structural-process-outcome typology developed in the indicators literature (Office of the United Nations High Commissioner for Human Rights, 2012). The taxonomy is deliberately layered rather than hierarchical: a single scenario typically engages multiple axes simultaneously, and encoding scenarios against the full set preserves the intersectional character of real human rights situations.

3.1.2. SCENARIOS AND SUB-SCENARIOS

Scenarios are the narrative anchor of HumRightsBench. Each scenario is a realistic, factually concrete situation that

balances the need to achieve authenticity to real-world human rights issues with the need to maintain experimental control via implicating specific elements of the taxonomy described in Section 3.1.1. We deliberately favor narrative scenarios over abstract fact patterns or doctrinal hypotheticals for two reasons. First, the situated factual texture of a scenario—the named setting, the actors involved, the specific resource at stake—is what forces a model to perform rule application rather than rule recitation (Guha et al., 2023), mirroring the analytical work that human rights practitioners do when assessing real situations (Office of the United Nations High Commissioner for Human Rights, 2011; Aletras et al., 2016). Second, narrative grounding allows us to introduce modular features (geographic location, identity of the affected group, institutional setting) whose values can be varied counterfactually to probe whether model reasoning is stable across protected characteristics—a property that abstract prompts cannot test. Scenarios are drafted from authoritative sources: General Comments, Special Procedures reports, leading jurisprudence, and human rights textbooks. Drafts are reviewed by at least three human rights experts and revised based on annotator feedback before inclusion (Section 3.2). (Section 3.2). For the pilot release on the right to water, each scenario is presented to the model as context preceding the IRAP questions as factual substrate against which the questions are answered.

Example: *In the sprawling informal settlement of “Aqualess Heights” in the State of Hydronia, thousands of residents, predominantly classified as low-income families, face a daily struggle to access clean and sufficient water. The main water supply consists of a few communal standpipes, often dry or providing water only for limited hours, and privately owned boreholes that charge exorbitant rates equivalent to multiple days of average wages in the area for unsafe water.*

Each scenario is associated with multiple **sub-scenarios** that narrow the analytical focus to a specific claim, action, or impact. Where the scenario establishes the broad factual setting, the sub-scenario fixes the legal question: it identifies a particular act or omission, attributes it to a specific duty-bearer, and characterizes its effect on identifiable rights-holders. This two-level structure mirrors human rights practice—a country situation is analyzed by isolating discrete events or patterns within it—and it allows the benchmark to generate multiple, semi-independent IRAP question sets from a single scenario without redundant world-building. Two sub-scenarios from the Aqualess Heights scenario above illustrate the range:

Sub-scenario A. *This scarcity forces residents, particularly women and children, to walk for hours to distant, often contaminated, sources, exposing them to health risks like cholera and imposing a significant burden on their time and*

dignity.

Sub-scenario B. *Despite the national “Water for All” policy, there is a chronic under-allocation of public resources to Aquales Heights. The state’s budget priorities have favored large-scale industrial projects over basic service provision in informal settlements, leading to dilapidated infrastructure and insufficient investment in water distribution networks.*

The two sub-scenarios engage different cells of the taxonomy despite sharing a setting: Sub-scenario A foregrounds an outcome failure with intersectional discriminatory impact on women and children, while Sub-scenario B foregrounds a structural failure in the allocation of resources implicating the obligation to fulfill. Each generates its own IRAP question set, and shared world-building across sub-scenarios reduces the annotation burden of scaling the benchmark.

3.1.3. IRAP QUESTIONS

Each sub-scenario generates a set of questions structured along the four IRAP steps: Issue Identification, Rule Recall, Rule Application, and Proposed Remedies. The four steps are not interchangeable difficulty levels of the same task; they probe distinct sub-capabilities of legal reasoning, and a model may plausibly succeed at one while failing at another. Separating them allows us to localize where reasoning breaks down rather than collapsing performance into an aggregate score (Guha et al., 2023).

Issue Identification (I). Two multiple-choice questions (I1 and I2) asking (in different ways) which type of failure or obligation violation is most clearly engaged by the sub-scenario, with answer choices drawn from the analytical axes of the taxonomy. As discussed further in Section 3.2, annotator feedback indicated that some sub-scenarios could plausibly engage multiple failure modes; we revise the question as “which failure mode is *most* present?” to elicit the model’s primary judgment.

Rule Recall (R). A multiple-choice question presenting a list of legal rules; the model selects the rule that applies. Candidate rules are drawn exclusively from international human rights law and presented uniformly by full instrument name and specific article (e.g., *ICESCR Article 11*) to prevent surface-form cues from substituting for substantive engagement. Each question has only one correct answer.

Rule Application (A). The model ranks a set of applicable rules by relative authority and relevance to the sub-scenario and provides a short explanation. Our correct answers are constrained by the principle that binding treaty obligations rank above non-binding interpretive instruments. However, as we discuss later, this model only slightly moves the analysis from Rule Recall to Rule Application, and more appli-

cation focused methodology is discussed in section 5

Proposed Remedies (P). An open-ended short-response question asking the model to propose <10 remedies appropriate to the sub-scenario. Remedies must be calibrated to the duty-bearer, the affected rights-holders, and the institutional context of the violation; the open-ended format reflects the irreducibly generative character of remedy proposal.

Together, the four question types trace the full arc of a human rights analysis, from initial issue characterization to identification and application of the governing norms to proposal of responsive measures. Section 3.2 describes how the scenarios, questions, and answers are validated and Section 4.1 describes how each question type is scored.

3.2. Annotation and Validation

HumRightsBench’s validity as a benchmark depends on whether its scenarios are recognizable as authentic human rights situations and are accurate reflections of the field’s understanding of which human rights laws apply. We therefore recruited (see IS1) practicing human rights experts to annotate IRAP questions and judge whether each scenario faithfully serves as a proxy for real-world cases.

We recruited reviewers via posts in four tech policy communities (All Tech is Human, the Center for AI and Digital Policy, Stanford Technology Ethics Program for Practitioners community, and TRUST: The Norwegian Centre for Trustworthy AI). Interested persons were qualified if they had at least two years of experience in human rights legal study or equivalent practice. Ten of 17 applicants recruited through this channel qualified. Six additional human rights experts were recruited through personal networks.

Annotation coverage varied by scenario. Four scenarios received the requisite three raters, and two received two. Three scenarios received no ratings due to high rates of annotator attrition. One scenario received six ratings. Across annotated items, experts broadly agreed that the scenarios represented authentic proxies (Table 3).

We were unable to compute IIC (see future work); however, strong ratings coupled with a legible foundational construct (human rights legal reasoning) supports a claim to moderate convergent validity (Campbell & Fiske, 1959). Further validation with larger annotator samples is a top priority.

4. Exploratory Results

4.1. Measurement

We benchmark four models chosen to cover the current frontier and one open-source reference. An example scenario and subscenario is provided in Appendix 6 and the question

Table 3. Aggregate annotator ratings by question type (5=strongest, 1=weakest; scores of 4 or 5 considered “pass,” all others considered “fail”), and scenario authenticity (higher rate is better). Full IRAP question ratings for each scenario in Appendix D

Metric	Mean Score
Overall I Questions	0.729
Overall R Questions	0.913
Overall A Questions	0.633
Overall P Questions	0.850
Scenario 1 Authenticity	0.933
Scenario 2 Authenticity	1.0
Scenario 3 Authenticity	0.70
Scenario 4 Authenticity	0.933
Scenario 5 Authenticity	1.0
Scenario 9 Authenticity	0.933

Note: No annotations were available for Scenarios 6, 7, and 8.

types are included in Appendix 6. The proprietary tier comprises GPT-5 (OpenAI, gpt-5-2025-08-07), Claude Opus 4.7 (Anthropic, claude-opus-4-7-2025-01-30), and Gemini 3 (Google, gemini-3-flash-preview, released 17 December 2025): three flagship systems, included to surface differences across providers at the closed-source frontier. The open-source reference is Qwen 3.5-9B (Alibaba, released 24 February 2026), a 9B-parameter model that fits on a single H100 and approximates what a self-hostable deployment can offer today. Every model is queried with five independent random seeds per question; answer-choice orderings are shuffled per seed to mitigate position bias. All evaluation runs occurred on 20, 21, or 22 May 2026. The overall accuracy across all question types is outlined in Table 4.

Structured-output extraction. All scoring assumes machine-parseable outputs. For each question type we define a Pydantic schema: a single- or comma-joined answer letter from a multiple choice questions list for I1/I2/R, a comma-separated ranking with per-rule rationale for Ranked Application, and a list of free-text remedies for PR, and obtain conforming responses through each provider’s native structured-output interface: OpenAI’s `beta.chat.completions.parse`, Anthropic’s tool-use mechanism with the schema declared as the tool input, Gemini’s `response.json_schema`, and, for Qwen served via vLLM, JSON-mode generation with the schema injected directly into the prompt. This obviates fragile regex post-processing and ensures grading is performed against exactly the format the model was instructed to produce.

Multiple-choice scoring (I (I1 and I2), R). The three multiple-choice question types reduce to single- or multi-letter selections. We score each response by exact-set match against the answer key.

Rule Application. Each Rule Application item, for the purpose of this pilot study, presents a small set of legal rules (typically 5–7) and asks the model to rank them by relevance to a described human-rights scenario. We compute Kendall’s τ (`scipy.stats.kendalltau`) between the predicted ranking and the gold ranking over the intersection of the two ranked sets, and declare a response *correct* when $\tau \geq 0.7$ — a conventional cutoff for “strong” rank agreement. To avoid spurious credit for models that default to the answer-choice order as presented, the rule labels are shuffled deterministically per (seed, scenario) and the gold ranking is remapped onto the shuffled labels, so the LLM never sees the ground-truth ordering as the natural alphabetical sequence. Responses that fail to produce a parseable, complete ranking are assigned $\tau = 0$ and counted as not-correct, matching the errors-as-wrong convention used elsewhere in this section.

Proposed Remedy: embedding-based correctness with human-calibrated thresholding. To evaluate open-ended LLM responses against reference answers, we score each response by the cosine similarity between OpenAI `text-embedding-3-small` embeddings of (i) the concatenated ground-truth remedies and (ii) the concatenated model-generated remedies. To convert this continuous score into a binary “correct” / “incorrect” verdict that can be aggregated across models, we calibrate a single decision threshold τ against human judgment. Two annotators independently rated $N = 40$ responses on a 1–5 Likert scale capturing coverage of the reference remedies; we declare a response *correct* when the mean of the two ratings is at least 4. The threshold τ is chosen to maximize Cohen’s κ between the binarized embedding predictor and the human gold label. To avoid optimistic bias from selecting τ on the same data we evaluate on, we report a leave-one-out cross-validated κ in which the threshold is refit on the remaining $n - 1$ rows for every held-out example. Inter-annotator agreement (Cohen’s κ on binarized labels and Spearman’s ρ on raw scores) is reported as a calibration ceiling against which the automatic scorer should be interpreted.

Calibration outcome. Across $N = 40$ annotated responses, inter-annotator agreement was $\kappa = 0.02$ on the binarized labels and $\rho = 0.42$ on the raw scores. The κ -optimal threshold was $\tau^* = 0.71$, yielding in-sample $\kappa = 0.54$ and LOOCV $\kappa = 0.12$. We use this τ^* to report each model’s accuracy as the share of responses whose embedding cosine exceeds τ^* , averaged across five seeds per model.

4.2. Cross-task comparison

Table 4 pools all five question types into a single overall accuracy per model. Gemini 3 leads at 0.577 ± 0.016 ,

followed by GPT-5 (0.537 ± 0.015) and Claude Opus 4.7 (0.508 ± 0.009); Qwen 3.5-9B lags substantially at 0.339 ± 0.020 . The per-task breakdown in Table 5 sharpens the picture: Claude Opus 4.7 is in fact the strongest model on the R sub-task (0.774), but Gemini 3 wins on every other type (I2: 0.710, RA: 0.240, PR: 0.630). Ranked Application is the hardest task across all four models: even the best system clears the $\tau \geq 0.7$ Kendall correctness bar on only roughly one row in four: reflecting the gap between selecting an answer from an enumerated set and producing a fully rationale-aligned ordering. Notably, Qwen 3.5-9B closes most of the gap to the proprietary frontier on PR (0.531 vs. Gemini’s 0.630) while remaining well behind on the more constrained sub-tasks, suggesting that open-ended generation against a holistic reference is presently more achievable for a 9B open-source model than tasks demanding precise alignment to structured ground truth.

Table 4. Overall mean accuracy across 5 seeds, pooled across I1, I2, R, A, and P question types. RA uses Kendall’s $\tau \geq 0.7$; PR uses cosine similarity ≥ 0.71 . Errored LLM calls are counted as incorrect. Qwen 3.5-9B is averaged over 4 seeds (only those with MCQ data present).

Model	Accuracy
GPT-5	0.537 ± 0.015
Claude Opus 4.7	0.508 ± 0.009
Gemini 3	0.577 ± 0.016
Qwen 3.5-9B	0.339 ± 0.020

5. Discussion and Future Work

Discussion. Our pilot results, although limited in scale, suggest that structured human rights reasoning tasks are challenging for frontier LLMs. More interesting is the way in which they are challenging. On closed-form tasks, LLMs regularly performed poorly on issue-identification—a result that has consequential implications not only for human rights workers but also from AI model developers, users, and regulators. Failures at this layer of the human rights reasoning process cascade throughout all other layers, leading to ungrounded rule applications, misconfigured remedies, and, ultimately, produce circumstances in which retrogression becomes structurally more likely. HumRightsBench makes these failures legible to all responsible actors. We note, models perform worst overall on rule application tasks, but this is expected as our pilot methodology sets very high thresholds for ‘correct’ responses here, and we are actively refining these question types. We discuss this further below.

The institutional landscape is now, for the first time, structured to receive this kind of evidence. The Council of Europe’s Committee on Artificial Intelligence has adopted HUDERIA as guidance for risk and impact assessment

Table 5. Mean accuracy across 5 seeds, broken down by question type. RA is scored as correct when Kendall’s $\tau \geq 0.7$; PR is scored as correct when full-response cosine similarity ≥ 0.71 (threshold calibrated against human annotators). Qwen 3.5-9B MCQ values (I1, I2, RR) are over 4 seeds.

Model	Question Type	Accuracy
GPT-5	I1	0.520
GPT-5	I2	0.675
GPT-5	R	0.715
GPT-5	RA	0.225
GPT-5	PR	0.550
Claude Opus 4.7	I1	0.473
Claude Opus 4.7	I2	0.595
Claude Opus 4.7	R	0.774
Claude Opus 4.7	RA	0.180
Claude Opus 4.7	PR	0.520
Gemini 3	I1	0.540
Gemini 3	I2	0.710
Gemini 3	R	0.765
Gemini 3	RA	0.240
Gemini 3	PR	0.630
Qwen 3.5-9B	I1	0.394
Qwen 3.5-9B	I2	0.519
Qwen 3.5-9B	R	0.494
Qwen 3.5-9B	RA	0.025
Qwen 3.5-9B	PR	0.531

in support of the Framework Convention. HUDERIA is designed to be used by both public and private actors to identify and address risks to human rights, democracy, and the rule of law across all phases of AI deployment. Yet HUDERIA lacks an empirical basis for evaluating whether the LLMs being assessed (or being used to conduct the assessment) are capable of reasoning about the rights implicated. HumRightsBench, once developed, is precisely the tool that can provide such a basis. Similarly, HumRightsBench results could directly inform Fundamental Rights Impact Assessment (FRIA) processes, providing the kind of structured, documented, and reproducible evidence that compliance with Article 27 of the EU AI Act demands.

Expanding coverage. Recall that this pilot’s scenarios are limited to a carefully produced selection that implicate only one right primarily: the right to water. Our immediate priority is to achieve more breadth. We plan to extend our scenario coverage to include the *the right to due process* and *the right to education* in the near future. Likewise, we plan to expand into different languages, considering (Samway et al., 2026)’s demonstration of the variance of model performance across different-language inputs.

Expanding methods. We will also broaden our task, scor-

ing, and assessment question design. We plan to assess the suitability of implementing an LLM-as-judge scoring pipeline to provide another interpretive signal of model performance on open-ended questions (e.g. the Proposed Remedies). We also plan to further decompose IRAP to include new types of I questions that move away from answer choices predicated on respect-protect-conduct indicators to those that reflect resource-modulated obligations (to minimum core versus progressive realization standards). Additionally, encouraged by promising early signal, we plan to conduct more substantial validation and inter-item consistency testing to ensure the statistical validity of our core construct.

Refining assessment. As noted in Section 3.1.3, our current Rule Application task only slightly moves our analysis from Rule Recall to Rule Application, and, importantly, does not sufficiently demonstrate reasoning over case-specific facts. We are actively developing new types of questions to better assess rule application, namely (1) “rule factors” questions that seek to assess model capabilities to correctly identify appropriate juridical tests, given the facts of specific scenarios; and (2) “jurisprudence” questions, which seek to assess model capabilities to correctly identify appropriate interpretive standards and thereby demonstrate the capability to sample authentic representations of human rights law as it is actually practiced today.

6. Conclusion

We introduced the core methodology required to produce HumRightsBench, an expert-validated scenario-based automated evaluation benchmark for human rights legal reasoning in LLMs and LRMs. On a pilot focusing on the human right to water, frontier model performance ranged widely, with significant differences across different elements of our human rights legal reasoning framework (IRAP). Importantly, models performed worst on issue-identification tasks, raising serious questions about their capabilities in this critical area of adoption. Although these results are exploratory—our dataset is small—they establish this area as an urgently important subfield of evaluation science, one that we intend to explore in greater depth and with more robust validation in work to come.

Acknowledgments

We wish to thank all of our scenario reviewers, including Susan Morrissey, Laura Carter, Nathan Heath, Richard Ncube, Selam Abdella, Jera White, Ajitha Sritharan, Angela Kariuki, and others who wish to remain anonymous. We are also grateful to friends and colleagues who advised on several matters throughout the project’s inception and delivery, including Dominico Zipoli, Fola Adeleke, Zach Lampell,

Helene Moliner, Claudia Flores, Megan Manion, Min Aung, Khalid Hassine, Lynn Gentile, Isabel Ebert, Nathalie Stadelmann, and Jan Rydzak, among others.

Impact Statement

IS1. Human Subjects Research for Data Validation

Annotation for scenario validation did not require Institutional Review Board (IRB) review, as the activity was determined to constitute PPI rather than human subjects research. Participation was voluntary and consensual: annotators contributed in exchange for acknowledgment rather than compensation, and were informed of these terms at recruitment and again on the survey form itself. Withdrawal was permitted at any time and carried no penalty. No deception was used at any stage. All responses were collected and processed through secure web forms hosted on Google Cloud, accessible only to project staff.

IS2. Ethics Statement

As we have stated above, we think reporting out methodology positively contributes to the mission shared by many AI evaluation scientists to ensure we develop methods for better understanding the instance- and systems-level impacts and harms AI systems may cause to critical human institutions.

We are mindful, of course, of the dual nature of many technologies. It occurs to us that this tool could in theory instead become a tool for assessing human rights workers themselves. This is not our intent or in the broader human rights community’s interests, and we implore researchers who may build on this work to consider related dual-use risks as AI diffusion and disruption proceeds. We emphatically object to this tool being used to assess the qualifications and performance of human legal professionals, as this would require altogether different methods and artefacts.

Likewise, we reason that such an evaluation could be used by malicious actors as a tool to determine potential loopholes or structural weaknesses in current human rights instruments for the purposes of exploiting them. In future, we intend to carry out precisely this kind of adversarial evaluation to forestall such efforts (and to advance our understanding of model safety guardrail robustness). In any case, we perceive this risk to be low—there is neither sufficient scale nor robustness in this pilot alone to enable such actions—but we note that this is an area of consideration, and we will take measures to prevent such malicious usage as much as possible in future.

IS3. Generative AI Statement

Authors used generative AI during this research.

In preparation for producing evaluation scenarios, developing a grounding in specific aspects of certain human rights instruments (e.g. social rights conventions) was in part aided by the use of NotebookLM.

Generative AI was also used to aid with table and figure LaTeX formatting, or, in limited cases, to translate author-written text into an illustrative diagram or summary table.

Likewise, some authors used Generative AI to prepare bibtex entries from validated sources where no bibtex citation was provided by the publisher.

In all cases, authors retained sole control (and responsibility) for the experimental design, analysis, and interpretation of implications of this work.

References

- Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., and Lampos, V. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016. doi: 10.7717/peerj-cs.93.
- Alston, P. Report of the special rapporteur on extreme poverty and human rights, philip alston. UN Doc. A/HRC/29/31, United Nations Human Rights Council, Geneva, May 2015. Twenty-ninth session, agenda item 3. <https://digitallibrary.un.org/record/798707>.
- Alston, P. The populist challenge to human rights. *Journal of Human Rights Practice*, 9(1):1–15, February 2017. doi: 10.1093/jhuman/hux007. URL <https://doi.org/10.1093/jhuman/hux007>.
- Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., and Singhal, K. Healthbench: Evaluating large language models towards improved human health, 2025. URL <https://arxiv.org/abs/2505.08775>.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks, 2024.
- 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., Rysstrøm, 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>.
- Campbell, D. T. and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.
- Chalkidis, I., Androutsopoulos, I., and Aletras, N. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4317–4323. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1424.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.297.
- Chalkidis, I., Garneau, N., Goanta, C., Katz, D. M., and Søgaard, A. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023. arXiv:2305.07507.
- Chlapanis, O. S., Galanis, D., and Androutsopoulos, I. LAR-ECHR: A new legal argument reasoning task and dataset for cases of the European Court of Human Rights. In *Proceedings of the Natural Legal Language Processing Workshop 2024*. Association for Computational Linguistics, 2024.
- Cole, K. Navigating humanitarian ai: Lessons learned from building a chatbot proof-of-concept. Technical report, Refugee Solidarity Network, 2024. URL <https://refugeesolidaritynetwork.org/reports/navigating-humanitarian-ai-lessons-learned-from-building-a-chatbot-proof-of-concept/>.
- Columbia Law School. Organizing a legal discussion: IRAC / CRAC / CREAC. Writing Center handout, Columbia Law School, 2022. Revised May 2022. <https://www.columbia.edu/~l1000/irac/>.

- law.columbia.edu/sites/default/files/2022-06/WC%20Handout%20IRAC%2C%20CRA%2C%20CREAC.revised%205.22.pdf.
- Dai, Y., Feng, D., Huang, J., Jia, H., Xie, Q., Zhang, Y., Han, W., Tian, W., and Wang, H. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10738–10766. Association for Computational Linguistics, 2025.
- de Schutter, O. Report of the special rapporteur on the right to food, olivier de schutter: Addendum — mission to Brazil. UN Doc. A/HRC/13/33/Add.6, United Nations Human Rights Council, Geneva, February 2010. Thirteenth session, agenda item 3. Mission to Brazil, 12–18 October 2009. <https://digitallibrary.un.org/record/677702>.
- Emelin, D., Le Bras, R., Hwang, J. D., Forbes, M., and Choi, Y. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 698–718. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.54.
- Enguehard, J., Ermengem, M. V., Atkinson, K., Cha, S., Chowdhury, A. G., Ramaswamy, P. K., Roghair, J., Marlowe, H. R., Negreanu, C. S., Boxall, K., and Mincu, D. Lemaj (legal llm-as-a-judge): Bridging legal reasoning and llm evaluation, 2025. URL <https://arxiv.org/abs/2510.07243>.
- Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., and Fernandez-Llorca, D. 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*, volume 8, pp. 850–864, 2025. doi: 10.1609/aies.v8i1.36595.
- Fan, Y., Ni, J., Merane, J., Tian, Y., Hermstrüwer, Y., Huang, Y., Akhtar, M., Salimbeni, E., Geering, F., Dreyer, O., Brunner, D., Leippold, M., Sachan, M., Stremitzer, A., Engel, C., Ash, E., and Niklaus, J. LEXam: Benchmarking legal reasoning on 340 law exams, 2025. Accepted to ICLR 2026.
- Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., and Ge, J. LawBench: Benchmarking legal knowledge of large language models, 2023.
- Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D., Zambrano, D., et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.
- Haupt, A., MacLennan, M., Ramli, U., Moreno Jiménez, R., Pentland, A., and Koyejo, S. un-bench: Closing the simulacra gap in development data. NeurIPS 2026 Competition Track, 2026. Stanford Trustworthy AI Research, Stanford HAI, UN Behavioural Science Group, UNICEF, and UNHCR. <https://humanitarianevals.org/>.
- Henderson, P., Krass, M. S., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. E. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*, 2022. arXiv:2207.00220.
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning AI with shared human values. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a. arXiv:2008.02275.
- Hendrycks, D., Burns, C., Chen, A., and Ball, S. CUAD: An expert-annotated NLP dataset for legal contract review. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Datasets and Benchmarks Track*, 2021b. arXiv:2103.06268.
- Holzenberger, N. and Van Durme, B. Factoring statutory reasoning as language understanding challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2021. arXiv:2105.07903.
- Holzenberger, N., Blair-Stanek, A., and Van Durme, B. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the Natural Legal Language Processing Workshop 2020*, 2020. arXiv:2005.05257.
- Jiao, J., Afroogh, S., et al. LLM ethics benchmark: A three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-18489-7. arXiv:2505.00853.
- Koskeniemi, M. *The Gentle Civilizer of Nations: The Rise and Fall of International Law 1870–1960*. Cambridge University Press, Cambridge, UK, 2001. ISBN 9780521623117.
- Langford, M. (ed.). *Social Rights Jurisprudence: Emerging Trends in International and Comparative Law*. Cambridge University Press, 2009.

- Li, H., Su, Y., Cai, D., Wang, Y., and Liu, L. LexEval: A comprehensive Chinese legal benchmark for evaluating large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024.
- Li, Y. and Wu, G. Legaleval-q: a benchmark for quality evaluation of llm-generated chinese legal text: Legaleval-q: a benchmark for quality evaluation... *Knowl. Inf. Syst.*, 68(1), February 2026. ISSN 0219-1377. doi: 10.1007/s10115-026-02703-7. URL <https://doi.org/10.1007/s10115-026-02703-7>.
- Lie, R. H. and Langford, M. The computational turn in international law. *Nordic Journal of International Law*, 93(1):38–67, 2024. ISSN 0902-7351. doi: 10.1163/15718107-bja10081. URL <https://doi.org/10.1163/15718107-bja10081>.
- Liu, Y., Cao, J., Liu, C., Ding, K., and Jin, L. Datasets for large language models: A comprehensive survey, 2024. URL <https://arxiv.org/abs/2402.18041>.
- Lourie, N., Le Bras, R., and Choi, Y. SCRUPLES: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479, 2021. doi: 10.1609/aaai.v35i15.17589.
- Medvedeva, M., Vols, M., and Wieling, M. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266, June 2020. ISSN 0924-8463. doi: 10.1007/s10506-019-09255-y. URL <https://doi.org/10.1007/s10506-019-09255-y>.
- Moyn, S. *The Last Utopia: Human Rights in History*. Belknap Press of Harvard University Press, Cambridge, MA, 2010. ISBN 9780674048720.
- Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., and Chalkidis, I. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023. arXiv:2301.13126.
- Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., and Ho, D. E. MultiLegalPile: A 689GB multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024. arXiv:2306.02069.
- Office of the United Nations High Commissioner for Human Rights. Frequently asked questions on a human rights-based approach to development cooperation. UN Doc. HR/PUB/06/8, United Nations, New York and Geneva, 2006.
- Office of the United Nations High Commissioner for Human Rights. Manual on human rights monitoring. Technical Report 7/Rev.1, United Nations, Geneva, Switzerland, 2011.
- Office of the United Nations High Commissioner for Human Rights. Human rights indicators: A guide to measurement and implementation. UN Doc. HR/PUB/12/5, United Nations, New York and Geneva, 2012.
- Office of the United Nations High Commissioner for Human Rights. The core international human rights instruments and their monitoring bodies. <https://www.ohchr.org/en/core-international-human-rights-instruments-and-their-monitoring-bodies>, 2025. Accessed: 2026-05-22.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Pedersen, S. *The Guardians: The League of Nations and the Crisis of Empire*. Oxford University Press, Oxford, 2015. doi: 10.1093/acprof:oso/9780199570485.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780199570485.001.0001>.
- Pesch, P. J. Potentials and challenges of large language models (llms) in the context of administrative decision-making. *European Journal of Risk Regulation*, 16(1): 76–95, January 2025. ISSN 2190-8249. doi: 10.1017/err.2024.99. URL <https://doi.org/10.1017/err.2024.99>.
- Rasiah, V., Stern, R., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D. E., and Niklaus, J. SCALE: Scaling up the complexity for advanced language model evaluation, 2023.
- Rovira, M. G. *The Project of Positivism in International Law*. Oxford University Press, Oxford, United Kingdom, 2013.
- Ruggie, J. G. *Just Business: Multinational Corporations and Human Rights*. W. W. Norton & Company, New York, 2013. ISBN 978-0-393-06288-5.
- Samway, K., Takagi, M. N., Mihalcea, R., Schölkopf, B., Chalkidis, I., Hershcovich, D., and Jin, Z. When do language models endorse limitations on human rights principles? In Demberg, V., Inui, K., and Marquez, L. (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6597–6623, Rabat, Morocco, March 2026. Association for Computational Linguistics.

- ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.347. URL <https://aclanthology.org/2026.findings-eacl.347/>.
- Schneiders, E., Seabrooke, T., Krook, J., Hyde, R., Leesakul, N., Clos, J., and Fischer, J. E. Objection overruled! lay people can distinguish large language models from lawyers, but still favour advice from an llm. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–14, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713470. URL <https://doi.org/10.1145/3706598.3713470>.
- Schwartz, R., Fiscus, J., Greene, K., Waters, G., Chowdhury, R., Jensen, T., Greenberg, C., Godil, A., Amironesei, R., Hall, P., and Jain, S. Assessing risks and impacts of ai (aria): Pilot evaluation plan. Technical report, National Institute of Standards and Technology (NIST), August 2024. URL https://ai-challenges.nist.gov/aria/docs/evaluation_plan.pdf. Accessed: 2026-05-21.
- Schwartz, R., Chowdhury, R., Kundu, A., Frase, H., Fadaee, M., David, T., Waters, G., Taik, A., Briggs, M., Hall, P., Jain, S., Yee, K., Thomas, S., Bhandari, S., Duncan, P., Thompson, A., Carlyle, M., Lu, Q., Holmes, M., and Skeadas, T. Reality check: A new evaluation ecosystem is necessary to understand ai’s real world effects, 2025. URL <https://arxiv.org/abs/2505.18893>.
- Shi, Y., Liu, H., Hu, Y., Song, G., Xu, X., Ma, Y., Tang, T., Zhang, L., Chen, Q., Feng, D., Lv, W., Wu, W., Yang, K., Yang, S., Wang, W., Shi, R., Qiu, Y., Qi, Y., Zhang, J., Sui, X., Chen, Y., Zhang, Y., Yang, A., Yu, B., Liu, D., Lin, J., Shen, W., Zhao, B., Clarke, C. L. A., and Wei, H. Plawbench: A rubric-based benchmark for evaluating llms in real-world legal practice, 2026. URL <https://arxiv.org/abs/2601.16669>.
- Sluga, G. *Internationalism in the Age of Nationalism*. Pennsylvania Studies in Human Rights. University of Pennsylvania Press, Philadelphia, 2013.
- Subramanian, R., Shiromani, T. S., Chaudhry, A., Li, R., Sharma, V., Zhu, K., and Dev, S. ProMoral-Bench: Evaluating prompting strategies for moral reasoning and safety in LLMs, 2026.
- Tuggener, D., von Däniken, P., Peetz, T., and Cieliebak, M. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1235–1241. European Language Resources Association, 2020.
- UN Committee on Economic, Social and Cultural Rights (CESCR). General comment no. 15: The right to water (arts. 11 and 12 of the covenant). Technical Report E/C.12/2002/11, United Nations, Geneva, 2002. URL <https://digitallibrary.un.org/record/486454>. Adopted at the Twenty-ninth Session of the Committee on Economic, Social and Cultural Rights.
- UN Human Rights Office of the High Commissioner. Advancing responsible development and deployment of generative AI: The value proposition of the UN guiding principles on business and human rights. B-tech project foundational paper, Office of the United Nations High Commissioner for Human Rights, Geneva, 2024. <https://www.ohchr.org/en/documents/tools-and-resources/advancing-responsible-development-and-deployment-generative-ai>.
- UN Human Rights Office of the High Commissioner (OHCHR). Business and human rights in technology project (b-tech): Scoping paper. Technical report, United Nations, Geneva, Switzerland, 2019. URL https://www.ohchr.org/Documents/Issues/Business/B-Tech/B-Tech_Scoping_paper.pdf.
- United Nations. Convention on the rights of the child, November 1989. Adopted 20 November 1989, Entered into force 2 September 1990.
- United Nations General Assembly. Universal declaration of human rights. Resolution 217 A (III), dec 1948. URL <https://un.org>.
- United Nations General Assembly. *International Convention on the Elimination of All Forms of Racial Discrimination*. United Nations, New York, 1965. URL https://treaties.un.org/pages/viewdetails.aspx?src=treaty&mtdsg_no=iv-2&chapter=4&clang=_en. United Nations, Treaty Series, vol. 660, p. 195.
- United Nations General Assembly. International Covenant on Economic, Social and Cultural Rights. United Nations, *Treaty Series*, vol. 993, p. 3, dec 1966a. URL <https://treaties.un.org/pages/showDetails.aspx?objid=080000028002b6ed>. Adopted 16 Dec. 1966, entered into force 3 Jan. 1976.
- United Nations General Assembly. International covenant on civil and political rights. Technical Report 2200A (XXI), December 1966b. URL <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>. Adopted 16 Dec. 1966, entered into force 23 Mar. 1976.

- United Nations General Assembly. Convention on the elimination of all forms of discrimination against women. Technical Report 34/180, United Nations, New York, NY, dec 1979. URL https://treaties.un.org/pages/viewdetails.aspx?src=treaty&mtds_g_no=iv-8&chapter=4&clang=_en. Adopted 18 December 1979, entered into force 3 September 1981.
- United Nations General Assembly. Convention on the rights of persons with disabilities. Resolution A/RES/61/106, United Nations, dec 2006. URL <https://www.refworld.org/legal/agreements/unga/2006/90142>.
- United Nations Office of the High Commissioner for Human Rights (OHCHR). Guiding principles on business and human rights: Implementing the united nations "protect, respect and remedy" framework. Technical report, New York and Geneva, 2011. URL <https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights>.
- Wang, S. H. et al. ACORD: An expert-annotated retrieval dataset for legal contract drafting, 2025. Full author list requires verification.
- Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., and Xu, J. CAIL2018: A large-scale legal dataset for judgment prediction, 2018.
- Xu, S., Staufer, L., T.Y.S.S., S., Ichim, O., Heri, C., and Grabmair, M. VECHR: A dataset for explainable and robust classification of vulnerability type in the European Court of Human Rights. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11738–11752. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.718.
- Yu, W. et al. Benchmarking multi-step legal reasoning and analyzing chain-of-thought effects in large language models, 2025. Full author list requires verification.
- Zhang, L., Grabmair, M., Gray, M., and Ashley, K. Thinking longer, not always smarter: Evaluating LLM capabilities in hierarchical legal reasoning, 2025.
- Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAAIL '21)*, pp. 159–168. Association for Computing Machinery, 2021. doi: 10.1145/3462757.3466088.
- Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. JEC-QA: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9701–9708, 2020. doi: 10.1609/aaai.v34i05.6519.
- Ziems, C., Yu, J. A., Wang, Y.-C., Halevy, A., and Yang, D. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3755–3773. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.261.

Appendix A: Expanded Review of the State of the Art in Legal Benchmarking

Summary table of sources of the human rights legal regime

Table 6. Sources of international human rights law, sorted by binding force.

Table 7. Two-tier categorization of international human rights instruments: legally binding treaties versus authoritative but non-binding soft-law instruments, with representative examples for each.

Category	Instrument Type	Examples
Binding	Core UN human rights treaties (obligations on ratifying states)	ICCPR, G.A. Res. 2200A (XXI) (1966); ICESCR, G.A. Res. 2200A (XXI) (1966); CERD (1965); CEDAW (1979); CRC (1989); CRPD (2006)
Authoritative but Non-Binding	Treaty body General Comments / General Recommendations (interpretive guidance from UN monitoring committees)	CESCR, General Comment No. 15: The Right to Water, U.N. Doc. E/C.12/2002/11 (2002); CCPR, General Comment No. 34: Article 19, U.N. Doc. CCPR/C/GC/34 (2011); CEDAW, General Recommendation No. 35 on Gender-Based Violence Against Women, U.N. Doc. CEDAW/C/GC/35 (2017); CRC, General Comment No. 25 on Children’s Rights in Relation to the Digital Environment, U.N. Doc. CRC/C/GC/25 (2021)
	Special Procedures (thematic and country-specific reports by independent experts appointed by the Human Rights Council)	Reports of the Special Rapporteur on Extreme Poverty and Human Rights ((Alston, 2015; de Schutter, 2010)); Special Rapporteur on the Right to Food; Working Group on Business and Human Rights; Special Rapporteur on the Situation of Human Rights Defenders
	Universal Periodic Review (UPR) recommendations (peer-review outcomes from the HRC’s state-to-state review cycle)	UPR Working Group reports issued under HRC Res. 5/1 (2007), covering all UN member states on a 4.5-year cycle

Literature Review

General Legal Reasoning Benchmarks. The most comprehensive English-language legal benchmark is LegalBench, which assembles 162 tasks across six categories of legal reasoning, including issue spotting, rule recall, and rule application (Guha et al., 2023). LEXam extends this paradigm to long-form legal argumentation, drawing 7,537 questions from 340 law school exams in English and German (Fan et al., 2025). Recent Chinese-language work has pushed the field toward structured reasoning evaluation. Like LegalBench, MSLR grounds its tasks in the IRAC framework (a common legal reasoning benchmark decomposing that practice into Issue Identification, Rule Recall, Rule Application, and Conclusion, see section XX.XX below) using real judicial decisions (Yu et al., 2025). LAiW organizes evaluation around the legal syllogism in three difficulty tiers (Dai et al., 2025), LawBench tests 21 LLMs across 20 tasks spanning memorization, understanding, and application (Fei et al., 2023), and LexEval offers the largest Chinese legal benchmark to date, structured by a cognitive ability taxonomy (Li et al., 2024). Across these benchmarks, model performance degrades sharply when tasks require multi-step reasoning rather than retrieval, indicating that current LLMs handle legal knowledge more effectively than legal inference (Fan et al., 2025; Yu et al., 2025). Counter-intuitively, LLMs occasionally perform much worse than LLMs on certain legal reasoning tasks, despite “thinking longer” (Zhang et al., 2025).

Legal Knowledge and Language Understanding. LexGLUE is the foundational benchmark for legal NLU, modeled on GLUE and covering seven classification and QA tasks across the European Court of Human Rights (ECtHR), the US Supreme Court, EU legislation, and commercial contracts (Chalkidis et al., 2022). CaseHOLD evaluates the identification of holdings in US case law through multiple-choice prompts derived from the Harvard Caselaw Access Project (Zheng et al., 2021), while LEDGAR tests multi-label classification of contract provisions from SEC filings (Tuggenier et al., 2020). LegalLAMA and LexFiles probe the legal knowledge stored in pretrained models across six common-law jurisdictions (Chalkidis et al., 2023). Pile of Law provides the underlying 256GB corpus used by many derivative resources (Henderson et al., 2022). Performance on these benchmarks has saturated relative to specialized legal language models such as LegalBERT and LegalXLM-R, suggesting that pure language understanding is no longer the binding constraint on legal AI performance (Niklaus et al., 2023).

Contracts, Statutes, and Case Law. Contract review is the most commercially mature subfield. CUAD contains 13,000 expert annotations across 510 commercial contracts and 41 clause categories, framed as a span-selection task (Hendrycks et al., 2021b). ACORD extends this to retrieval of precedent clauses for contract drafting (Wang et al., 2025). For statutory reasoning, SARA tests entailment and question-answering over the US Internal Revenue Code (Holzenberger et al., 2020), with a follow-up that decomposes statutory reasoning into discrete language understanding challenges (Holzenberger & Van Durme, 2021). Case-law prediction is dominated by Chinese benchmarks such as CAIL2018 (Xiao et al., 2018) and JEC-QA from the Chinese National Judicial Examination (Zhong et al., 2020). These domain-specific benchmarks consistently show that LLMs perform well on extraction and classification but struggle with tasks requiring the integration of statutory text, case facts, and doctrinal context (Holzenberger et al., 2020; Hendrycks et al., 2021b).

Multilingual and Cross-Jurisdictional Benchmarks. As in other fields, the English-only orientation of early legal NLP pretraining corpora has been a persistent limitation (Liu et al., 2024). MultiLegalPile addresses this with a 689GB pretraining corpus across 24 languages and 17 jurisdictions, paired with a family of LegalXLM-R models (Niklaus et al., 2024). LEXTREME provides the corresponding multilingual evaluation suite, covering classification tasks across European jurisdictions (Niklaus et al., 2023). SCALE tests long-document processing across five languages in the Swiss federal system, with documents extending to 50,000 tokens (Rasiah et al., 2023). These resources demonstrate that monolingual fine-tuning consistently outperforms multilingual pretraining for jurisdiction-specific tasks, raising open questions about the transferability of legal reasoning across legal systems (Niklaus et al., 2024).

Human Rights and International Law. Human rights NLP began with (Aletras et al., 2016), who showed that simple classifiers could predict ECtHR Article violations with 79 percent accuracy from case facts alone—work substantially advanced by (Medvedeva et al., 2020)’s critical engagement. (Chalkidis et al., 2019; 2022) subsequently formalized this as the ECtHR-A and ECtHR-B tasks within LexGLUE. VECHR introduces a more demanding objective: classifying vulnerability types in ECtHR decisions and VECHR evaluating model explanations against expert rationales (Xu et al., 2023). Most recently, the UDHR Trade-off Benchmark presents 1152 LLM-generated scenarios implicating 24 Universal Declaration articles to test how LLMs reason about trade-offs between rights and competing interests such as public safety or economic stability when using one of eight different evaluation languages (Samway et al., 2026). Coverage of substantive international human rights reasoning remains thin relative to domestic legal benchmarks, although this is beginning to change. UNICEF and UNHCR have teamed up to propose a NeurIPS 2026 competition to validate AI representations of hard-to-reach populations using agency microdata (Haupt et al., 2026). More directly, Kennedy and Heath propose a new threat model methodology for evaluating model vulnerabilities to producing deepfake media of POWs violative of several POW protections established in the Geneva Conventions (Kennedy and Heath, forthcoming 2026).

Normative, Moral, and Ethical Reasoning. Although not legal benchmarks strictly speaking, these kinds of evaluations heavily implicate “precursor” capabilities of interest to legal evaluations and often use legal sources to build evaluation datasets or scoring rubrics. The ETHICS benchmark introduced systematic evaluation of model alignment with datafied representations of justice, deontology, virtue ethics, utilitarianism, and commonsense morality constructs (Hendrycks et al., 2021a). SCRUPLES extended this to 625,000 ethical judgments over 32,000 real-life anecdotes (Lourie et al., 2021), while Social Chemistry 101 cataloged 292,000 rules-of-thumb governing everyday social norms (Forbes et al. 2020). Moral Stories tests norm-consistent and norm-violating action selection in branching narratives (Emelin et al., 2021), and the Moral Integrity Corpus annotates 38,000 dialogue turns with underlying rules of thumb (Ziems et al., 2022). More recent work has shifted toward multi-dimensional and adversarial evaluation: MoralBench offers metadata-rich diagnostic structure, ProMoral-Bench unifies evaluation across ETHICS, Scruples, and WildJailbreak under a single moral-safety score (Subramanian et al., 2026), and the three-dimensional LLM Ethics Benchmark assesses foundational principles, reasoning robustness, and value consistency (Jiao et al., 2025). A persistent finding across this literature is that models align more closely with individualistic Western moral frameworks than with collectivist or non-Western ones, indicating substantial cultural bias in normative training data (Jiao et al., 2025).

Evaluation Methodology. A parallel methodological literature has developed around how to score legal and normative outputs. LeMAJ decomposes legal answers into atomic “Legal Data Points” and uses LLM-as-judge scoring validated against LegalBench (Enguehard et al., 2025). PLAWBENCH applies rubric-based evaluation specifically to LLM legal agents (Shi et al., 2026). LegalEval-Q uses regression-based quality assessment across 49 models to evaluate generated legal text (Li & Wu, 2026). These approaches respond to a recurring concern that simple accuracy metrics inadequately capture the structured, justification-dependent character of legal reasoning (Guha et al., 2023; Fan et al., 2025).

Appendix B: Gaps

Legal valence. Human rights norms carry distinctive legal weight that does not obtain in other legal domains. Obligations to respect, protect, and fulfill are legally differentiated, and misidentifying which is at stake produces not merely an inaccurate answer but a legally consequential one. Current safety benchmarks and content policy evaluations are not designed to test this granularity. Existing guardrails—safety classifiers, RLHF alignment, model cards—address human rights concerns only incidentally, through vague values-based framing rather than grounding in the actual normative architecture of international law.

Grounding in practice. Evaluating models for human rights competencies requires a knowledge of the law and attending normative reasoning practices, but it also requires a scholastic understanding of the application of legal standards to situated facts. That is, it requires an understanding of norms as well as the developed practice of assessing whether certain actions contribute or degrade the progressive realization of those norms. Human rights practice therefore requires accounting for unnamed but specific and heavily implicated duty-bearers, rights-holders, and institutional remedies in concrete scenarios. Instance-level prompting strategies that attempt to elicit human rights reasoning on an ad hoc basis cannot substitute for systematic benchmark-level measurement. Synthetic generation of scenario data instead of expert-creation presents consequential risks to construct validity here (Bean et al., 2025; Eriksson et al., 2025; Schwartz et al., 2024; 2025).

Highly developed institutional space. International human rights law has a mature, documented interpretive ecosystem, including treaty bodies, General Comments, Special Procedures, regional courts, and an active body of jurisprudence. This institutional situatedness supplies ample signal towards which to optimize evaluation datasets as well as a demanding standard against which model outputs can be meaningfully assessed. But this requires benchmark specificity—general legal reasoning or legal knowledge benchmarks will not adequately cover this institutional space or its peculiar dynamics.

Difference from municipal law. Unlike domestic legal benchmarks, international human rights law is not the law of any single jurisdiction. It operates through treaty ratification, state practice, and authoritative interpretation, without a single apex court (IHL notwithstanding). Reasoning quality cannot be assessed against one jurisdiction’s doctrine; it must be evaluated against principles that apply universally while remaining sensitive to implementation variance and the jurisdiction-based legal tests specific judges operating within certain municipal traditions may reach for. This distinguishes it from many existing legal benchmarks, which are grounded in domestic statutory and case law within a single jurisdiction (Guha et al., 2023).

Unique cross-jurisdictional variance. The progressive realization standard that underpins economic, social and cultural rights in the international human rights regime acknowledges that states implement these rights differently depending on available resources and not only context (which is of course important for all human rights). A benchmark that encodes a single jurisdiction’s approach to a certain case as ground truth without nominating such a context will systematically misrepresent model performance elsewhere. Indeed, this is an instance of a broader problem in hierarchical benchmark design: imposing a uniform taxonomy risks obscuring the context-specific dynamics that determine whether a norm has been violated in practice. Other rubric-based approaches to benchmarking in different fields that apply layered taxonomies representing specific, intermediary, and universal criteria have produced better construct validity (Arora et al., 2025). Exceedingly few have been attempted in legal benchmarking.

Appendix C: Scenarios

Scenario: "In the sprawling informal settlement of "Aqualess Heights" in the State of Hydronia, thousands of residents, predominantly classified as low-income families, face a daily struggle to access clean and sufficient water. The main water supply consists of a few communal standpipes, often dry or providing water only for limited hours, and privately owned boreholes that charge exorbitant rates equivalent to multiple days of average wages in the area for unsafe water. The high cost of purchasing water from private vendors adds a severe economic strain on already struggling households, residents claim."

Sub-scenario: "This scarcity forces residents, particularly women and children, to walk for hours to distant, often contaminated, sources, exposing them to health risks like cholera and imposing a significant burden on their time and dignity."

Figure 2. example scenario with sub-scenario

Appendix C: Questions

There are 5 types of questions in our list.

I1: "Which one of these obligations was most clearly violated by the state? Respond with the appropriate answer letter and a short sentence explaining your answer."

I2: "Which of the following failure modes is most present in this scenario? Respond with the appropriate answer letter and a short sentence explaining your answer."

RR: "If the above scenario occurred as described, which of the following rules best applies? State the letter of the appropriate rule and provide one sentence of explanation."

RA: "Consider the following list of rules and rank them in order of relevance and applicability to the described scenario. Provide a sentence for each rule describing your reasoning for ranking it in that order."

PR: "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. "

Figure 3. Five questions types in our dataset.

Appendix D: Annotator ratings of IRAP questions by scenario

Table 8. Mean IRAP question ratings by scenario.

Scenario	I	R	A	P
Scenario 1	0.875	.75	0.50	1.00
Scenario 2	0.50	1.00	1.00	.50
Scenario 3	1.00	1.00	0.50	1.00
Scenario 4	1.00	1.00	0.33	0.67
Scenario 5	0.33	1.00	0.67	1.00
Scenario 9	0.67	0.73	0.80	0.93
Mean	0.729	0.913	0.633	0.850

Note: No ratings were available for Scenarios 6, 7, and 8.