LegalBench2: Towards Measuring Subjective Quality in Legal Work

Neel Guha

Most work on legal AI evaluation treats quality as binary (responses are either correct or incorrect) and objective (experts agree on what makes a response correct), relying on multiple-choice questions, classification tasks, and instance-level rubrics. Popular benchmarks in this tradition include LegalBench, LEXam, and Harvey's LAB. But much legal work resists this framing: quality is continuous—responses are better or worse, not right or wrong—and contested—experts may reasonably disagree about which response is better. Lawyers call this capacity *judgement*; in AI discussions, it is often called *taste*. We introduce LegalBench2, the first benchmark aimed at measuring judgement in the legal domain. LegalBench2 pairs tasks spanning transactional work, litigation, legal education, civil legal aid, arbitration, and other domains with responses ranked via pairwise preferences elicited from multiple expert attorneys. We describe the benchmark's construction and the research problems it opens up: building and validating faithful LLM judges, constructing personalized per-attorney leaderboards, and integrating validated judges into post-training pipelines.