Forthcoming in: Philipp Hacker, Andreas Engel, Sarah Hammer, Brent Mittelstadt (eds), The Oxford Handbook on the Foundations and Regulation of Generative AI (OUP, 2024). This is a draft version, and may be updated.

Building GenAI Benchmarks: A Case Study in Legal Applications

Neel Guha[1], Julian Nyarko, Daniel E. Ho, Christopher Ré
Stanford University

Abstract

GenAI's potential for use in highly technical fields like law and medicine has produced a corresponding need for domain-specialized benchmarks. This chapter provides an overview of domain-specialized benchmarking, using the legal field as a case study. First, it provides an overview of what benchmarking is, and what benchmarks consist of. Second, it draws on recent work in developing legal benchmarks to discuss the challenges with constructing domain-specialized benchmarks. Finally, it describes how despite these challenges, benchmark creation offers unique opportunities for interdisciplinary collaboration.

1.1 Introduction

Excitement about Generative AI (GenAI) and general purpose foundation models (FMs) centers on their potential to work across a wide range of fields (e.g., finance, education, journalism) [2]. Yet, questions emerge as to the trustworthiness and reliability of these tools. For every study praising GenAIs potential is another highlighting potential failures. FMs can hallucinate, producing knowledge that is unsubstantiated, misleading, or blatantly false [22]. They can replicate social biases, yielding poorer performance or even behaving differently when operating on data from minority demographic groups [16]. Already, for instance, have these failure modes led to both publicity crises and—in the case of lawyers—sanctions by courts [26].

Traditionally, researchers and stakeholders develop trust in AI systems through a process known as "benchmarking" [29]. Benchmarking consists of crafting datasets which allow stakeholders to observe the behavior of models under various conditions. They enable researchers to measure, for instance, how accurate a models outputs are for a task, or whether those outputs reflect social biases. Just as clinical trials allow drug developers to observe the effects of pharmaceuticals and crash tests allow automotive engineers to assess the crashworthiness of different vehicle designs, benchmarks allow AI developers to study a models safety and performance features.

[1] Corresponding author. Please reach out to nguha@stanford.edu.

Issues of trustworthiness and reliability are perhaps most stark when FMs are applied to specialized domains, like law, finance or medicine. Here, an important obstacle is the lack of benchmarks. In stark contrast to older, more traditional applications for AI—such as language classification, data extraction, or image recognition—benchmarks for specialized domains are far and few between. Thus, GenAIs arrival has triggered significant interest and effort into both building—and understanding how to build—rigorous benchmarks for specialized domains.

The lack of benchmarks for these domains is made all the more significant by the fact that FMs appear uniquely suited for specialized applications. Disciplines like law and medicine suffer supply shortages—there are far more individuals with legal and medical problems than there are lawyers and doctors to help them [28]. Technological solutions capable of improving the distribution of medical and legal services could thus theoretically alleviate long-standing social inequities. Second, GenAI models appear to be capable of performing tasks requiring significant knowledge and reasoning skills [2, 19, 15]. To the extent that such abilities are necessary for domains like law and medicine, GenAI systems appear more suitable than any technology yet developed.

This chapter provides an overview of domain-specialized benchmarking for GenAI, using the legal domain as a case study. It describes the considerations that incorporate into each part of the benchmarking process, from constructing datasets to analyzing results. In particular, it focuses on the unique technical challenges posed by benchmarking efforts in specialized domains.

Legal applications offer a useful prism for analyzing GenAI benchmarking for several reasons. The first is that the legal domain implicates specialized expertise. Lawyers typically undergo years of training, and the tasks they perform require sophisticated reading, writing, and reasoning skills. The second is that legal applications can be “high risk.” Legal mistakes can incur significant financial losses for clients and even lead to deprivations of liberty [30, 7]. Finally, the legal domain has been an area of particular excitement for GenAI. Commentators have identified it as an area particularly ripe for automation, with over $700 million in startup funding since 2023 and an influx of companies marketing AI tools to lawyers, in-house counsel, government and other legal providers [23].

The remainder of this chapter is structured as follows. Section II provides an overview of what benchmarks are, how theyre constructed, and how theyre used. Section III then uses the legal domain as a case study to describe the challenges with GenAI benchmarking. Finally, Section IV closes the chapter with an optimistic vision: though constructing benchmarks for specialized domains may be challenging, benchmark construction projects offer a unique opportunity to conduct interdisciplinary AI work.

1.2 An overview of benchmarks

Benchmarks are designed with the purpose of allowing engineers to evaluate a models performance on one or more tasks. For each task, the benchmark will contain a dataset consisting of pairs of “model inputs” and “desired outputs.” Model inputs correspond to the data that is fed into the model, and desired outputs correspond to what the model should—in an ideal world—produce.

The precise form of the model inputs and desired outputs depends on the task. For example, the IMDB benchmark measures sentiment classification—the task of determining the sentiment of a given text [25]. The model inputs for this benchmark correspond to the text movie reviews, and the outputs correspond to whether the review is “positive” or “negative” about the movie (i.e., the reviews sentiment).

In contrast, the CIFAR-10 dataset measures image recognition for ten different classes of objects [20]. The model inputs correspond to images (represented as pixel matrices), and the desired output corresponds to the object type (out of the ten possibilities) most prominently featured in the image. The specific object types captured in the benchmark are “airplane, automobile (but not truck or pickup truck), bird, cat, deer, dog, frog, horse, ship, and truck (but not pickup truck).”

When using a benchmark to evaluate a model, engineers will feed each of the inputs into the model, and collect the models output (i.e., a “prediction”). Across the entire dataset, they will then compare the predictions to the desired outputs in order to assess the performance of the model. When doing this, engineers typically rely on metrics which summarize performance into a single numerical measure. The choice of metric can depend on the task, the domain, or the structure of the models outputs. For instance, when tasks require models to predict categories (e.g., like CIFAR-10 or IMDB), it is common to measure performance using accuracy, which captures the proportion of the models predictions which matched the desired output. Benchmarks which require a model to predict a number, in contrast, often rely on metrics like mean-squared error.

Beyond computing metrics, engineers may also manually inspect model outputs and compare them to the desired outputs. Manual inspectionaka “error analysis”enables engineers to develop richer intuition regarding model behavior. Examining different inputs for which the models prediction was incorrect can surface properties of the model input which more frequently correlated with mispredictions. This in turn can lead engineers to explore which aspect of a models training or architecture may be responsible, and what corrective steps might be.

Manual inspection is also often necessary for more complex tasks, where models produce outputs not conducive to automated metrics [9]. This frequently arises in the context of generative AI, when engineers wish to evaluate models that produce unstructured text or images. Because metrics for measuring generated text quality can be inaccurate, researchers may instead have external human annotators score or rank model outputs in a blinded fashion. While this form of evaluation can provide a more accurate assessment of model performance, it is also significantly more expensive.

Benchmarks (and performance on them) are used by stakeholders to answer a variety of different questions. First, at the most basic level, benchmarks enable stakeholders to determine how often a models prediction is “correct,” or “good.” This is often useful when attempting to compare a models performance to human baselines. Medical AI developers, for instance, can use benchmarks to assess whether a diagnostic model performs at levels comparable to human physicians. In this manner benchmarks also serve an important governance purpose, as they allow stakeholders to predict how often a model will be wrong and what harms might result. They can additionally—as in the case of facial recognition benchmarks—highlight significant performance disparities across data subgroups [3]. Thus, a models performance on a benchmark can help stakeholders determine how the models benefits weigh against any risks of error. Second, benchmarks allow engineers and researchers to directly compare different models or algorithmic approaches. For example, it is now common for AI companies to promote the advantages of their products by reporting performance on popular benchmarks.

Importantly, the reliability of the inferences an engineer draws from a benchmark rests on several assumptions [27]. The first requirement is that benchmark data is “unseen” by the model. Significant research has demonstrated that modern machine learning models memorize data seen during model training [13]. When benchmark data “leaks” into a models training data, performance on the benchmark is no longer indicative of a models capacity to perform the task, and is often artificially inflated. In recent years, a number of studies have identified instances where high model performance could in fact be attributed to training-set data leakage [10]. When researchers removed the leaked samples from the benchmark, performance dropped significantly. This can also occur when practitioners select hyperparameters based on test set performance.

Benchmark	**What does it evaluate?**	**What form do model inputs and desired outputs take?**
IMDB [25]	Sentiment detection.	Inputs are movie reviews. Outputs denote the sentiment (positive or negative) of the movie review.
CIFAR-10 [20]	Object recognition in images.	Inputs are images containing different natural and every day objects. Outputs are the identities of the object in the image.
Massive Multitask Language Understanding [17]	Factual knowledge acquired by models during training.	Inputs are multiple multiple choice questions corresponding to different subjects. Outputs are the correct answer choice for the question.
ImageNet [8]	Object recognition in images.	Inputs are images containing different natural and every day objects. Outputs are the identities of the object in the image.
NIST FRVT [14]	Facial recognition from images	Inputs are images containing a face. Outputs are the identity of the individual in the image.
BoolQ [5]	Reading comprehension.	Inputs are short text passages and yes/no questions. Outputs are either “yes” or “no.”
GSM8K [6]	Grade-school mathematical reasoning.	Inputs are mathematical word problems. Outputs are the numeric answer to the word problem.

Table 1.1: Notable Benchmarks in Machine Learning.

The second assumption is that benchmark data actually allows engineers to accurately measure the desired objective. One failure mode is when models inadvertently learn to game a benchmark by relying on shortcuts that do not easily extend beyond the benchmark itself. Artifacts in the data sometimes mean that models can perform well by relying on spurious correlations in the benchmark, without actually learning the underlying task. In image object recognition tasks for example, researchers have noticed that models will sometimes predict object type based on the image background, as certain types of backgrounds appear more often with certain object types (e.g., green pastures with cows) [1]. When object backgrounds are exchanged, model predictive performance drops.

Another failure mode is when the behavior researchers wish to measure in a model is misaligned with the benchmark task. This is far more likely to occur when researchers wish to evaluate more abstract qualities of a model, like its ability to perform a certain kind of "reasoning." The concern here is that high performance on the task used to evaluate the model is insufficient to establish the presence of these abstract qualities.

1.3 Challenges with Benchmarking: Lessons from Law

This section uses the legal fieldand recent work on developing legal GenAI benchmarksto illustrate the unique challenges that arise when designing, constructing, and distributing domain-specific GenAI benchmarks. Though this section focuses on law, many of the discussed challenges also arise in other specialized domains, such as medicine and finance.

Evaluating Unstructured Text

A significant fraction of legal work involves producing text [2]. In transactional work, lawyers may draft contracts or other business documents. In litigation, lawyers file written complaints, motions, and briefs with the court. In rulemakings, attorneys may draft dense comment letters on proposed rules. And in adjudication, judges and presiding officials will write judgements or summaries of proceedings. GenAI methods like large language models (LLMs) are exciting because of their potential to assist, enhance, and perhaps even automate such forms of text production. At their best, LLMs could enable legal professionals to work more efficiently and accurately, with substantial implications for the delivery of legal services.

Unfortunately, benchmarking the text generation capabilities of LLMs in law is both technically challenging and expensive. As an example, consider an application in which users query an LLM with questions about everyday legal issues (e.g., Can I be fired for telling my coworker my salary?), and the LLM responds with answers. To develop a benchmark for this task, researchers might start by accumulating a large set of representative questions , along with ground-truth answers for these questions. The difficulty arises when assessing LLM responses to each question. Simply checking whether the LLMs response is identical to the ground-truth answer is inadequate: the LLM response could contain functionally equivalent information, but differ in word-choice or word-ordering. Researchers thus need to measure whether the LLMs response is semantically equivalent to the ground-truth answer.

The most rigorous way to perform this evaluation is for subject-matter experts to manually compare each LLM response to each ground-truth answer. The biggest issue with this approach, however, is that it can become extremely expensive. Manual review for general domain tasks typically makes use of crowdworker platforms like MechanicalTurk, where an hour of a single annotators time can cost $15-30. But for legal tasks like the one above, annotators must be legal subject-matter experts, who can charge upwards of a $1000 dollars per hour for legal representation.

The cost of legal expert time has important ramifications for how the evaluation is conducted. Researchers can only afford evaluation on smaller datasets, which have lower statistical power [4]. They may not be able to acquire multiple annotations per sample (i.e., from different lawyers), and are thus less likely to detect annotation mistakes. Finally, it means that evaluation can only be performed a small number of times. But because language models are frequently updated, evaluations performed on olderand ostensibly worseversions of the model lose relevance.

The difficulty of manual evaluation has motivated the development of techniques for automated evaluation, which require little or no human oversight. Historically, a number of metricslike Rouge, BLEU, BERT-score, and othershave been developed for this purpose, and applied to tasks like summarization and data-to-text generation. Broadly these metrics compute the similarity between a candidate generation and the ground-truth answer, using either embedding-distance or n-gram overlap. While cheap to compute however, these metrics are poor indicators of quality in practice and often uncorrelated with human judgements [12].

More recently, researchers have begun to explore using other LLMs to judge the quality of candidate-generations [9]. In this paradigm, researchers provide the question, ground truth response, and candidate response to a second LLM, which produces a score or rating for the candidate generation. While LLM judgements are generally more correlated with human judgment than metrics like Rouge, researchers have observed that LLM-judges suffer from several biases. For instance, LLM-judges tend to prefer generations from the same base model (e.g., GPT-4 scores GPT-4 generations more highly), and generations which are longer. How to effectively counteract these biases is an open and in-progress area of work.

Finally, it is unclear whether ostensibly generalist LLMs are capable of judging generations in contexts where generation quality depends on specialized knowledge. If the most advanced LLMs today struggle themselves to produce accurate generations to legal questions, is it reasonable to trust their judgements of other generations?

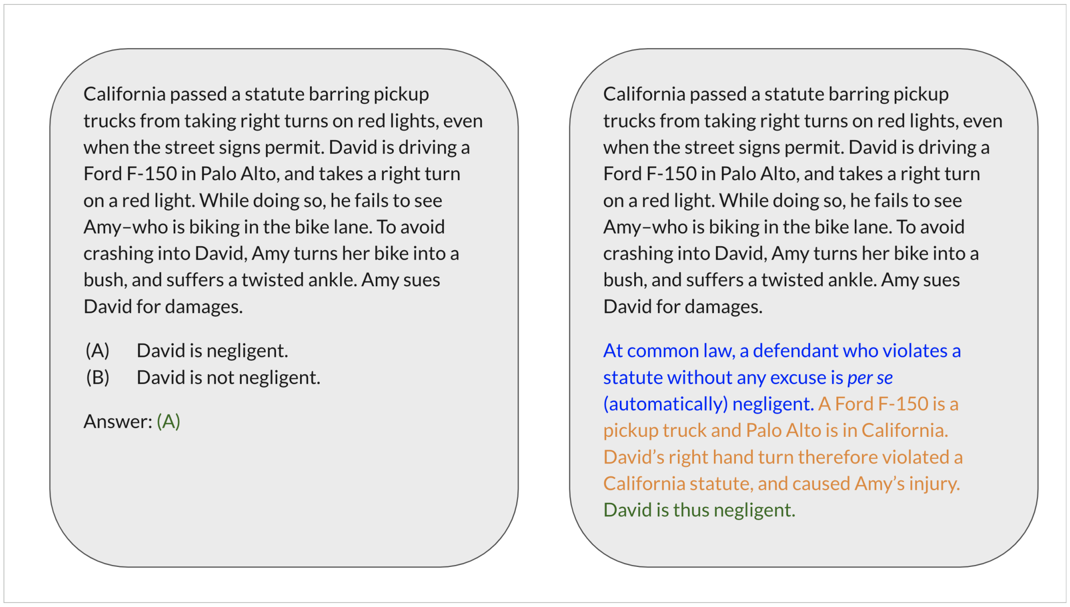
The difficulty in evaluating unstructured text means that most legal benchmarks are cast as multiple choice or classification tasks. While this makes evaluation cheaper and simpler, it also produces task formulations that fail to capture essential aspects of legal reasoningthe identification of a relevant legal principle, and the application of that principle to a set of facts.

Figure 1.1 below offers an example. On the left is an illustration of how traditional legal benchmarks evaluate legal reasoning using a multiple-choice format. Here, the model is correct if it generates (A), and incorrect otherwise. The right-hand side illustrates the type of answer a human lawyer would be expected to generate. Note that this answer communicates far more information about the reasoning by which the answer (in green text) was arrived at. It states the applicable legal principle (in blue text), and illustrates the application of that principle to the provided facts (in orange text).

The difficulty gap between both responses poses an important challenge for legal benchmarking. Producing the answer on the right requires greater capacity with regards to knowledge recall (i.e., remembering the legal principle) and reasoning (applying the principle to the facts). But if our benchmarks primarily consist of multiple choice questions, we have limited insight into whether modern LLMs can reliably perform these skills. Simplifying tasks—in order to make evaluation practical—can alter how we perceive the competency of these models.

Train-test leakage

A second challenge is the tension between the benefits of benchmark distribution and the risk of train-test leakage with modern large language models. Conventionally, researchers distribute benchmarks by making them publicly accessible online, often through platforms



California passed a statute barring pickup trucks from taking right turns on red lights, even when the street signs permit. David is driving a Ford F-150 in Palo Alto, and takes a right turn on a red light. While doing so, he fails to see Amy–who is biking in the bike lane. To avoid crashing into David, Amy turns her bike into a bush, and suffers a twisted ankle. Amy sues David for damages.

(A) David is negligent.
(B) David is not negligent.

Answer: (A)

California passed a statute barring pickup trucks from taking right turns on red lights, even when the street signs permit. David is driving a Ford F-150 in Palo Alto, and takes a right turn on a red light. While doing so, he fails to see Amy–who is biking in the bike lane. To avoid crashing into David, Amy turns her bike into a bush, and suffers a twisted ankle. Amy sues David for damages.

At common law, a defendant who violates a statute without any excuse is *per se* (automatically) negligent. A Ford F-150 is a pickup truck and Palo Alto is in California. David's right hand turn therefore violated a California statute, and caused Amy's injury. David is thus negligent.

Figure 1.1: **(L)** The standard multiple-choice construction of a legal task, found frequently in existing benchmarks, with a correct model response. **(R)** The same legal question, with the response a human lawyer would be expected to generate. The different colors denote different essential elements of the answer.

like Huggingface or Github.[2] Making benchmark data easy to access and download provides important benefits. First, it improves transparencyanyone can inspect a dataset, critique its composition, and identify errors. Over time, this enables benchmarks to improve in quality, and allows researchers to develop a more nuanced understanding of how the benchmark should be used. Second, public benchmarks allow for benchmark reuse and modification. Researchers can build on top of the benchmark, by adding more samples, more annotations, or other useful metadata. Finally, public benchmarks allow researchers and engineers to develop a better understanding of model performance. Benchmark access allows researchers to manually inspect model errors and identify how algorithmic or architectural changes alter the distribution of those errors. Without access to the benchmark samples, such analysis would not be possible.

These benefits are particularly stark in the context of legal applications. Developing legal benchmarks is expensive and hard. Making benchmarks publicly accessible thus has an important democratic access implication. Without public benchmarks, the capacity to evaluate and inspect models might otherwise be limited to solely those entities with financial resources to do so. Public benchmarks enable a broader swath of stakeholders to take an active role in evaluating machine learning systems.

The challenge raised by liberal public distribution of benchmarks is that the risk of train-test leakage for commercial LLMs greatly increases. Though the precise details regarding

[2] Public distribution can additionally be required by research funding or conference publication requirements. For instance, the Neural Information Processing Systems Track for Benchmarks and Datasets requires that accepted papers make all data available (barring a narrow set of exceptions).

how modern commercial LLMs are trained are not publicly known, there is broad consensus that these models are trained on snapshots of the entire Web, and all text contained therein. Thus, it is highly likely that many benchmarks available for download online have already been trained on by commercial LLMsand that any benchmark distributed online in the future will be included in the training set of next-generation LLMs. Researchers have found for instance, significant performance disparities between benchmarks released before and after web cut-off dates [21].

Researchers must hence choose between transparency and benchmark informativeness. Distributing a benchmark online enables greater community engagement and pushes scientific understanding forward. But it also risks training leakage with respect to commercial LLMs, and means that the performance of those LLMs on the benchmark should be taken with a grain of salt. Public benchmarks in the age of LLMs thus appear to come with expiration dates, and offer only a narrow window of utility before the risk of contamination grows too high.

Dataset costs

A third challenge is that developing high quality legal benchmarks can be extremely expensive. Constructing robust datasets requires adherence to several principles. First, the dataset should be large, in order to allow for fine-grained comparisons between models. Second, dataset labels should be generated by processes that incorporate multiple annotators, in order to minimize the effect of annotator bias, subjectivity, or error. Consulting multiple annotators also ensures that samples for which the choice of label is a subjectively close call are appropriately identified and/or excluded from the benchmark. Third, insofar as it is possible, benchmark data should be representative, and capture different manifestations of the task.

Collectively, these principles mean that constructing benchmarks require significant annotation effort. For general domain machine learning taskslike identifying the topic of a document, the object in an image, or the relevant passage to a question in a Wikipedia articlebenchmark designers can utilize online crowdworkers to perform annotation. The general knowledge nature of these tasks mean that online crowdworkers already possess the skills to accurately perform annotation or at the very least, can quickly learn the necessary task.

In domains like law however, generating annotations often requires specialized legal expertise that reflects deep subject-matter knowledge and familiarity. Crowdworkers are unlikely to possess such expertise, and teaching them the requisite knowledge or skills is impractical. Utilizing actual lawyers to annotate datawhile desirableis almost always economically prohibitive. And because much data generated naturally by lawyers in legal practice is subject to attorney-client privilege, simply calling on law firms to release collected data isnt feasible.

An illustration of the significant expenses implicated is illustrated by the Atticus Projects efforts to develop a benchmark for evaluating contract understanding (CUAD) [18]. CUAD encompasses 13,000 expert annotations across 500 contracts, classifying different clauses within these contracts by category. To develop CUAD, the Atticus Project relied on volunteer labor from lawyers, law students, and machine learning researchers. In order to develop expertise in the task, law student annotators underwent between 70-100 hours of training. Each annotation underwent multiple rounds of review, by both annotators and supervising attorneys. The Atticus Project conservatively estimates that this datasetonly moderately sized by general domain standardswould have otherwise cost over $2 million to construct.

An important implication of the cost of constructing legal datasets is that developed benchmarks tend to reflect the priorities and interests of legal institutions with the financial capacity to construct datasets [11]. In turn, this raises an important emerging concern regarding legal AI: though AIs potential to expand access-to-justice is often touted, the financial barriers to developing benchmarks may actually stymie this goal.

Recent work has explored two strategies to overcome the significant costs of manual annotation. In the first strategy, researchers attempt to derive labels from naturally occurring/already existing annotations in publicly available data. CaseHOLD, for instance, relies on the fact that legal citations in judicial opinions often contain a summary of the cited case that is relevant to the preceding text [31]. By writing regex functions to extract these summaries from the opinions, the developers of CaseHOLD were able to create a multiple-choice benchmark in which models had to select the summary which best supported a given passage of judicial reasoning.

Yet, a downside to crafting benchmarks from already-existing annotations is that observable data in the legal system is often subject to influential selection biases. For instance, not all legal complaints culminate in a judicial decision, and not all judicial decisions produce published opinions. Cases which have an obvious conclusion on the law are less likely to result in opinions than cases implicating a closer call. As a result, benchmarks consisting solely of observed datawithout accounting for the processes that produced that datamay not generalize to realistic applications and fail to capture the full spectrum of relevant tasks.

The second strategy is illustrated by efforts like the LegalBench project [15]. Rather than collecting a large number of samples for a single task, LegalBench sought to collect smaller sized datasets for many different tasks. Because individual tasks were smalland required relatively less manual effort to constructthe LegalBench researchers were able to crowd-source the construction of these tasks by soliciting contributions from different legal expert individuals and organizations (e.g., law professors, lawyers, researchers). Combining these tasks with existing refactored datasets yielded a benchmark of over 160 tasks, spanning a diverse range of legal areas.

Benchmark subjectivity

A final challenge for legal benchmarking is the prevalence of tasks which involve subjective judgements. Most general domain machine learning benchmarks in AI correspond to tasks where there is often an objectively correct answer, and during benchmark construction researchers sometimes intentionally remove samples for which there is subjective disagreement. The rationale is straightforward: if humans reasonably disagree on a samples label, then there is no firm ground-truth against which to assess the models prediction.

The challenge for law, however, is that a substantial number of lawyerly tasks necessarily involve subjective analyses [24]. Consider, for example, the task of determining whether a particular legal claim is likely to be meritorious. At the outer boundarieswhen the claim is obviously frivolous or obviously likely to succeedthe correct answer is clear. But in the murky middle between such extremes, it is more difficult to distinguish right from wrong. Equally competent and thoughtful lawyers may disagree on the expected outcome, for reasonable reasons.

This nuance complicates legal benchmarks. If the most complex and interesting instances have no obvious answerand subject matter experts reasonably disagreeits difficult to design benchmarks which distinguish good models from bad models. In other words: predictive performance on these tasks conveys no information about the reasoning abilities or legal knowledge of evaluated models.

One approach is to instead reorient these types of tasks around explanations, instead of predictions. The observation here is that while prediction correctness may not be indicative of legal reasoning ability, the strength of an explanation for a prediction can be. When asked about the merits of a claim, what distinguishes better and worse lawyers is their ability to identify and engage with the strengths and the weaknesses of each sides case, and the different factors a court might address. In short, it is their ability to persuasively support their prediction. Thus, rather than evaluating models on their ability to predict whether the plaintiffs will prevail on the claim, we might instead evaluate models on their ability to identify the best arguments for either side. Of course, while focusing on task explanations instead of predictions circumvents problems with subjectivity, it creates new challenges with regards to evaluating free-form responses.

1.4 Benchmarks as a Locus for Interdisciplinary Collaborations

We close on a note of optimism. Accompanying the rush of algorithmic and methodological work on GenAI has been an explosion in benchmarking work. Across a wide range of domains, academic researchers, industry, and civil society groups have collaborated to develop and release new benchmark datasets, practices, and approaches. While the above challenges remain significant, these efforts suggest that they can be overcome, through both technical novelty and focused effort. Our own work has shown us, moreover, that domain-specialized benchmarks are artifacts which can help develop symbiotic relationships between practitioners of different disciplines.

Generalist AI researchers can derive value from specialized domain benchmarks because they provide an opportunity to study existing models and methods with new types of tasks. Specialized domains are often more complex, and thus present harder problems for AI systems. For instance, legal applications require models to draw on copious domain specific knowledge, perform multi-step reasoning, and parse lexically complex and extremely long documents. Studying traditional AI approaches in these settings often inspires new technical developments. As an illustrative example, LegalBench has already helped researchers study new questions around prompting, domain-specialized finetuning, and efficient evaluation.

Domain-specialized benchmarks can also be useful for domain experts who are trying to understand the benefit and risk tradeoffs of using AI. At core, benchmarks provide domain experts with information regarding how well different types of AI models perform on different types of domain tasks. From a governance and ethics perspective, this allows domain experts to appropriately calibrate their understanding of risk. And from a resource allocation perspective, this allows domain experts to focus on those tasks for which AI is most fruitful. The LegalBench project showed, for instance, that LLMs are particularly adept for annotation tasks relevant to legal researchers. Building from LegalBench, a number of recent works have explored new legal empirical questions by incorporating LLMs.

That domain benchmarks engage both domain experts and AI researchers can produce mutually reinforcing cycles of benefits. AI researchers desire datasets upon which they can study models and establish the need for novel methodological approaches. Subject-matter practitioners can fulfill this need by helping create benchmarks incorporating tasks from their domain. Similarly, these subject-matter practitioners benefit from understanding whether machine learning models can perform essential tasks, but lack the expertise to both evaluate models and explore new approaches. AI researchers—through their use of benchmarks—fulfill this informational need.

The broad appeal of domain specialized benchmarks thus makes them ideal vehicles for bringing together experts from different fields. The process of creating a benchmark puts these experts into conversation with each other. It encourages them to develop a shared vocabulary and understanding, enabling AI researchers to learn about the specialized domain, and domain researchers to learn about AI. As AI is increasingly applied to specialized domains, benchmarking projects offer unique opportunities to guide AI development towards more socially useful and impactful applications.

Bibliography

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595*, 2020.

[5] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[7] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. generalization: Quantifying data leakage in nlp performance evaluation. *arXiv preprint arXiv:2102.01818*, 2021.

[11] David Freeman Engstrom and Jonah B Gelbach. Legal tech, civil procedure, and the future of adversarialism. *U. Pa. L. Rev.*, 169:1001, 2020.

[12] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[13] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[14] Patrick J Grother, Patrick J Grother, Mei Ngan, and K Hanaoka. *Face recognition vendor test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology, 2014.

[15] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Amit Haim, Alejandro Salinas, and Julian Nyarko. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*, 2024.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[18] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

[19] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480, 2024.

[22] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[23] Megan Ma, Aparna Sinha, Ankit Tandon, and Jennifer Richards. Generative ai legal landscape 2024, 2024.

[24] Megan Ma, Brandon Waldon, and Julian Nyarko. Conceptual questions in developing expert-annotated data. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 427–431, 2023.

[25] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[26] Sara Merken. New york lawyers sanctioned for using fake chatgpt cases in legal brief. *Reuters*, 26 June 2022. Available at: `https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/`.

[27] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11), 2021.

[28] Lisa R Pruitt and Zach Newman. California's attorney deserts: Access to justice implications of the rural lawyer shortage. *UC Davis Legal Studies Research Paper, Forthcoming*, 2019.

[29] Vijay Reddi. Benchmarking ai. `https://harvard-edge.github.io/cs249r_book/contents/benchmarking/benchmarking.html`.

[30] John Roberts. 2023 year-end report on the federal judiciary. `https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf`.

[31] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168, 2021.