Selective Overconfidence: How Large Language Models Underserve Social Welfare Law Across Jurisdictions

Shera Potka[1], **Alex Thomo**[1], **Paula Helm**[2], **Wulf Loh**[3]
[1]University of Victoria, Canada
[2]Goethe University Frankfurt, Germany
[3]Universität Tübingen, Germany
spotka@uvic.ca, thomo@uvic.ca, helm@c3s.uni-frankfurt.de, wulf.loh@izew.uni-tuebingen.de

Abstract

Large language models are increasingly used in legal information systems, yet their knowledge is not uniformly distributed across legal domains. We introduce a knowledge-probe pipeline that measures whether an LLM exhibits differential confidence across areas of law. Applying the pipeline to 1,200 court decisions across three jurisdictions (Germany, France, Canada), we find that Llama 3.3 70B systematically under-hedges on social welfare law relative to tax law in non-English jurisdictions: 49.8% of tax probes elicit admitted uncertainty vs. only 6.8% for social welfare in Germany ($p < 0.0001$), with a smaller but significant gap in France ($p = 0.003$) and no gap in Canada (English). The pattern is consistent with the representation of each language in training corpora and replicates on Qwen3 32B. The model achieves comparable accuracy in both domains, yet signals false confidence precisely in the legal areas that serve socioeconomically vulnerable populations. A retrieval-augmented intervention, providing related legal text at query time, closes 87% of the gap, suggesting the overconfidence stems from a knowledge deficit. We frame this as a form of compounding injustice: already marginalized users receive confidently stated but unreliable information in legal areas that are historically underserved and underrepresented in NLP, with no signal that the system's knowledge is thinner in their area of need than in other areas of law, which are much better hedged.

Introduction

When a person seeking disability benefits asks a large language model (LLM) about their legal rights, the answer they receive has proven to be far from 100 per cent reliable. The same unreliability applies to a corporate tax lawyer prompting to receive information about deductions. Despite these comparable reliability rates, our research reveals a troubling asymmetry: the model is far less likely to warn the disability claimant when it is unsure than it is to warn the tax lawyer. Both answers may be equally unreliable, but only the tax answer comes with the necessary caveat.

This paper documents and explains this phenomenon, which we term *selective overconfidence*: a pattern in which an LLM's willingness to admit uncertainty varies across legal domains even when its accuracy is comparable. We find that the models we tested hedge their answers substantially more often in tax law than in social welfare law, despite performing at similar levels of accuracy in both. The people who rely on social welfare law (disability claimants, social assistance recipients, the unemployed, single parents) are often those least likely to have access to professional legal counsel, and therefore more likely to depend on the outputs received through automated legal information (Chien and Kim 2025; Ogunde 2024; Legal Services Corporation 2022). The selective overconfidence we document compounds this disadvantage: users with the least access to professional legal help receive the worst epistemic service from the systems they may be tempted to fall back on because of socioeconomic necessity.

The asymmetry is not random but reflects linguistic asymmetries well known to be persistent in NLP resources (Joshi et al. 2020). It follows a *language-mediated gradient*: large in German, moderate in French, and absent in English. In Germany, the model admits uncertainty on 49.8% of tax law probes but only 6.8% of social welfare probes ($p < 0.0001$). In France, the gap is smaller but significant (11.2% vs. 6.3%, $p = 0.003$). In Canada, where both domains are in English, no gap exists ($p = 0.17$). This gradient is consistent with the availability of legal text in each language: social welfare law, though it generates far more court cases than tax law, produces less commercially published commentary and fewer digitally accessible decisions, particularly in non-English languages.

These findings emerge from a *knowledge-probe pipeline* that we apply to 1,200 court decisions (200 per domain per country). The pipeline extracts legal knowledge from each decision, generates questions that test general domain expertise, obtains blind answers from the model, and judges those answers against the extracted ground truth. All steps use the same model, so any systematic tendencies cancel in the cross-domain comparison.

The finding is not an artifact of a single model. We replicate the experiment on three model families: Llama 3.3 70B (Meta), Qwen3 32B (Alibaba), and gpt-oss 120B (OpenAI). Llama and Qwen, despite being trained independently by different organizations on different data, both hedge significantly more on tax law than on social welfare law. The gap is larger for Llama (43 percentage points) than for Qwen (4.5 points), but both are statistically significant and in the same direction. The third model, gpt-oss, reveals a different prob-

lem: it almost never hedges in either domain despite lower accuracy, a form of *uniform* overconfidence that harms all users equally. Together, these results suggest that confidence miscalibration in legal AI is widespread, but takes different forms across architectures.

Crucially, the selective overconfidence can be reduced. A simple retrieval-augmented intervention, in which related legal text from the underrepresented domain is provided to the model at query time, closes 87% of the confidence gap. The model does not become more accurate; rather, exposure to the complexity of the domain restores its ability to recognize what it does not know. This is consistent with the overconfidence being driven by a knowledge deficit: without exposure to social welfare legal text, the model lacks the context to assess its own ignorance.

We interpret these findings through the lens of *compounding injustice* (Hellman 2018, 2023). The problem is not only that LLMs produce erroneous or misleading legal information in the domain of social welfare law. It is that they do so with a level of confidence that can make such information difficult for users to evaluate, especially when users lack access to alternative sources of legal expertise. In this context, overconfidence becomes socially consequential as it may prevent individuals from accurately assessing their legal status, recognizing possible entitlements, or pursuing claims to which they may be entitled.

Moreover, the problem of overconfidence is unevenly distributed. Individuals and communities who rely on readily available sources of legal information, including LLMs, often do so under conditions of constrained access due to limited financial resources, limited time, language barriers, distrust of institutions, administrative complexity, or lack of affordable legal support. These conditions are themselves socially patterned. Members of protected or historically disadvantaged groups, including along lines of race, gender, age, disability, migration status, and class, may therefore be more likely to depend on low-threshold digital information systems for economic and institutional reasons. In this sense, the harms of LLM overconfidence may produce a disparate impact (Barocas and Selbst 2016) where erroneous outputs impose heavier burdens on those already less able to verify, challenge, or correct it.

The injustice is compounded because this disparate impact does not arise on socially neutral grounds. The very groups most likely to depend on LLM-based legal information may already be disadvantaged by prior injustices that shape their access to legal knowledge, representation, administrative support, and institutional trust. When an overconfident model misstates eligibility, misrepresents legal standards, or presents uncertain information as authoritative, it does not merely create a new informational error. It adds an additional layer of disadvantage to existing social and epistemic inequalities. In Hellman's terms, the harm *compounds* prior injustice (Hellman 2018, 2023). Unequal access to reliable legal support makes certain groups more dependent on LLMs, while LLM overconfidence then further undermines their capacity to understand, claim, and contest their rights.

In what follows, we first situate the study in relation to three strands of work: LLM uncertainty and calibration, LLM evaluation in legal domains, and bias and fairness in legal AI. We then present our knowledge-probe pipeline and report the results. The results are organized around five findings: confidence asymmetry, accuracy scores, selected qualitative examples, language gradients, and cross-model comparison. In the discussion, we show how these empirical findings can be interpreted as an instance of compounding injustice, with particular attention to the role of training data in producing and amplifying this mechanism. We conclude by developing five implications for the deployment of AI in legal contexts.

Related Work

LLM uncertainty and reliability

A growing literature measures LLM calibration and how models express uncertainty about their own answers. Kadavath et al. (2022) show that larger models can be reasonably well-calibrated when predicting whether they will answer multiple-choice questions correctly, and Yin et al. (2023) find that models can identify some unanswerable questions but remain below human proficiency. Lin, Hilton, and Evans (2022) show that a GPT-3 model can be fine-tuned to produce calibrated verbalized uncertainty ("I am 70% confident"), an approach that requires fine-tuning access not available for most deployed LLMs. What matters for users, however, is not whether LLMs can be calibrated in principle, but whether they spontaneously hedge in the answers they actually receive.

Xiong et al. (2024) evaluate confidence-elicitation strategies across five model families and find that LLMs tend to verbalize overconfidence on specialized tasks. They report that targeted prompting and consistency-aggregation strategies reduce this tendency, but no single method consistently excels on tasks requiring professional knowledge. Geng et al. (2024) survey the broader literature, charting techniques across white-box and black-box settings and noting that calibration evaluation across heterogeneous tasks and domains remains an open challenge.

Uncertainty connects to other dimensions of LLM reliability. Sharma et al. (2024) document that state-of-the-art AI assistants consistently exhibit sycophancy, the tendency to match user beliefs even at the expense of accuracy, across diverse free-form generation tasks. Loh and Seyfert (2026) argue that this behavior may itself be linked to uncertainty, hypothesizing that less confident models are more inclined to follow user input even when that input contains biases or falsehoods. This connects calibration to deliberative reliability. A system that cannot reliably signal what it does not know is more easily steered by what its user wants it to say, and is therefore harder to deploy as a tool for public reason. Our measurement of spontaneous hedging contributes to this broader project of assessing LLM reliability in high-stakes settings where users may lack alternative reference points.

Prior work asks whether LLMs are calibrated and whether they remain truthful under user pressure. We ask instead whether LLMs' ability to hedge varies systematically across legal domains, and how that variation tracks language and

training-data representation across three jurisdictions.

LLM evaluation in legal domains

Recent work has benchmarked LLM performance on legal tasks including bar exams (Bommarito and Katz 2022), contract review (Hendrycks et al. 2021), and legal reasoning (Guha et al. 2024). Subsequent studies have evaluated LLM use in legal education (Choi et al. 2024) and the construction of domain-specialized legal LLMs (Cui et al. 2023). A parallel line of work has profiled how often LLMs hallucinate legal content. Dahl et al. (2024) report error rates between 58% and 88% on questions about US federal court cases with foundation models, and Magesh et al. (2025) audit deployed commercial legal AI tools and find hallucination rates between 17% and 33% despite vendor claims of being "hallucination-free." These results document what LLMs get right and wrong about law. They do not tell us whether the model warns users when it might be wrong.

A persistent limitation of this literature is its reliance on benchmark tasks that test isolated capabilities such as bar-exam multiple choice or contract-clause extraction, and on aggregate accuracy metrics that average over heterogeneous legal domains. Such evaluations may obscure how performance varies across the substantive areas of law that different user populations actually encounter, particularly as legal information increasingly reaches users through online and AI-mediated channels (Susskind 2023). Moreover, most evaluations focus on English-language, common-law settings. While multi-jurisdictional legal NLP resources have begun to emerge (Chalkidis et al. 2023), they remain English-language and accuracy-focused. To our knowledge, no prior work compares LLM performance across legal domains in non-English jurisdictions, nor measures *self-reported confidence* rather than accuracy. We do both, across three jurisdictions.

Bias and fairness in legal AI

Algorithmic bias in legal contexts has been studied extensively, particularly in relation to recidivism risk assessment (Angwin et al. 2016) and criminal sentencing (Ryberg and Roberts 2022; Lippert-Rasmussen 2022; Madden et al. 2017). Scholars have also shown how algorithmic decision-making can automate inequality, especially in the domain of social welfare, where it contributes to what Eubanks (2018) calls the "digital poorhouse". With the rise of generative AI, however, the locus of concern shifts from algorithmic decision-making to the legal information that consumer-facing AI systems produce. Bender et al. (2021) argue that large language models can amplify harms through scale, opacity, and the absence of grounded reference. Similarly, Hofmann et al. (2024) show that language models can encode covert racial bias triggered by dialect features, with downstream effects on judgments in employment, criminal sentencing, and related domains.

These concerns resonate with broader critiques of technical approaches to fairness (Barocas, Hardt, and Narayanan 2023). More specifically for the legal domain, Selbst et al. (2019) cautions that purely technical fairness interventions often fail to account for the sociotechnical contexts in which legal decisions are made. This critique is consistent with our finding that confidence behavior cannot be assessed independently of the legal domains and language ecosystems in which it occurs. Existing benchmark regimes have likewise been criticized as too crude to provide sufficient accountability (Raji et al. 2021), prompting the development of alternative audit and accountability frameworks, including end-to-end internal auditing (Raji et al. 2020) and operational tools for measuring fairness in deployed systems (Bird et al. 2020).

Less attention, however, has been paid to *informational bias*: the possibility that AI-based legal information systems provide systematically less reliable service for some areas of law than for others. A related question, which researchers are only beginning to investigate, concerns the social effects of *language modeling bias*: the structural tendency of AI systems to represent and process some languages with greater precision, reliability, and nuance than others due to inequalities in linguistic resources, data availability, and training materials. Such bias may affect social groups differently, thereby both reflecting and compounding existing social and informational inequalities (Bella et al. 2024). In this paper, we measure one such inequality: the differential expression of uncertainty across legal domains. We further provide an interpretation of its social origins and normative significance.

Multilingual LLM performance

Joshi et al. (2020) systematically map the linguistic-diversity gap in NLP by showing that only a small fraction of the world's languages are meaningfully represented in language technology research. Their analysis links this exclusion to the uneven distribution of linguistic resources and to the way research attention in NLP follows already well-resourced languages. This finding resonates with Bender (2011), who criticizes the field's frequent claim to language-independence and argues that truly language-independent NLP would require substantially more linguistic and typological sophistication than is usually built into system design and evaluation.

Subsequent work shows that this imbalance persists in contemporary large language models. Ahuja et al. (2023) evaluate generative AI models across a broad multilingual benchmark and find that performance remains uneven across languages and tasks, with persistent challenges for low-resource languages. Lai et al. (2023) similarly show that ChatGPT's multilingual performance varies substantially across 37 languages and multiple NLP tasks, with high-, medium-, low-, and extremely low-resource languages not benefiting equally from the model's general capabilities. These studies suggest that multilingual capacity in LLMs should not be inferred from fluent performance in English or other high-resource languages.

A related strand of work shifts the focus from aggregate performance to the normative and sociotechnical meaning of linguistic bias. Blodgett et al. (2020) argue that work on "bias" in NLP often lacks conceptual clarity and insufficiently specifies what kinds of system behavior are harmful, to whom, and why. This is important for our purposes

because linguistic bias in legal information systems cannot be understood only as a technical deviation from benchmark accuracy; it must also be assessed in relation to the social conditions under which users depend on language technologies. Kathunia et al. (2024), for example, show in the domain of sentiment analysis that model performance varies across multilingual settings and translation-to-English pipelines, illustrating how English can function not only as a dominant training language but also as an evaluative intermediary through which other languages are processed.

Finally, recent work has begun to theorize the epistemic and sociocultural consequences of such asymmetries. Helm et al. (2024) argue that language modeling bias can lead to epistemic injustice when concepts, meanings, or worldviews from less dominant language communities become less expressible or less intelligible. Olojo et al. (2025), in turn, show how machine translation systems in Nigeria can flatten sociocultural meaning, mishandle linguistic nuance such as diacritics, and reproduce hierarchies between majority and minority languages. Taken together, these studies highlight how language-related model failure concerns the social organization of linguistic visibility, the loss of contextual meaning, and the unequal distribution of epistemic authority.

Our contribution builds on this literature by showing that degradation does not manifest only as lower accuracy, which evaluators can in principle verify against a ground truth. It also appears as a failure of uncertainty expression. In the cases we examine, models do not merely provide less reliable answers in some legal domains and language settings, but fail to signal that their answers may be unreliable. This matters because end users encounter the generated reply without necessarily having the legal knowledge, linguistic resources, or external reference points required to recognize that a hedge, caveat, or uncertainty marker is missing. This finding is especially consequential as it disproportionally regards legal domains such as social welfare law, where affected users are often already positioned in conditions of heightened institutional dependency and constrained access to professional legal support.

Method

Knowledge-probe pipeline

Our pipeline has four stages, all executed by the same model. Using a single model for all stages means that any systematic tendencies (e.g., a general reluctance to admit uncertainty) affect both domains equally within each jurisdiction, and the cross-domain comparison remains valid. Our primary analysis uses Llama 3.3 70B; we replicate on two additional models in Section *Cross-model comparison*.

1. **Ground-truth extraction.** Given a court decision, the model extracts structured legal knowledge: the core legal issue, the legal test applied by the court, and key statutes with section numbers. This structured output serves as the reference against which blind answers are later evaluated. The prompt is in the language of the decision (German for Germany, French for France, English for Canada).

Table 1: Question types used in the analysis.

Type	Tests	Example
Legal test	How courts reason	"What is the legal test for deducting gift-contract payments as operating expenses?"
Statute	What the law says	"What does § 9 Abs. 1 EStG provide?"
Procedural	How the system works	"What is the standard of review for a BFH appeal?"

2. **Question generation.** From the extracted ground truth, the pipeline generates three questions per case, each targeting a different dimension of legal knowledge (Table 1). A legal-test question asks how courts reason about a particular issue (e.g., "What is the legal test for deducting gift-contract payments as operating expenses?"). A statute question asks what a specific provision says (e.g., "What does § 9 Abs. 1 EStG provide in German law?"). A procedural question asks how the judicial system works (e.g., "What standard of review applies when a BFH decision is appealed?"). All three types test general domain expertise that does not require access to the specific court decision from which the question was derived.
3. **Blind answering.** In a separate call with no case context, the model answers each question using only its parametric knowledge. The prompt explicitly instructs the model: "If unsure, say so." This instruction is identical across domains, so any baseline tendency to hedge or not hedge applies equally to tax and social welfare probes.
4. **Judging.** In a further separate call, the model evaluates the blind answer against the ground truth. The primary output is a binary classification: did the blind answer *admit uncertainty*? The judge also assigns accuracy scores (1–5) on whether the answer identifies the right legal standard, cites the correct statutory provision, and is specific rather than vague. These scores serve as a secondary check, allowing us to verify that any confidence asymmetry is not simply explained by one domain being easier than the other.

Measuring admitted uncertainty. The binary "admitted uncertainty" indicator is our primary outcome variable. The judge classifies a blind answer as admitting uncertainty when the answer contains explicit hedging language, such as "I'm not entirely certain," "I am unsure about the specific details," or "I cannot confirm." An answer that states a legal proposition without qualification, even if wrong, is classified as not admitting uncertainty. We measure the *rate* of admitted uncertainty per domain (the proportion of probes in which the model hedges) and compare this rate between tax and social welfare law within each jurisdiction.

Judge validation. To validate the binary admitted-uncertainty classifier, we drew a stratified subsample of 50 probes per jurisdiction (25 tax / 25 social welfare, balanced on the judge's labels) and had two annotators in-

Table 2: Data sources by jurisdiction and domain.

Country	Domain	Court / Source	n
Germany	Tax	Finanzgerichte / BFH[a]	200
Germany	Social	Bundessozialgericht[a]	200
France	Tax	Conseil d'État, fiscal[b]	200
France	Social	Conseil d'État, social[b]	200
Canada	Tax	Tax Court of Canada[c]	200
Canada	Social	Social Security Tribunal[c]	200

[a]openlegaldata.io [b]Legifrance Open Data [c]A2AJ

dependently code each probe under two explicit rubrics: a *strict* rubric counting only literal uncertainty statements, and a *liberal* rubric also counting conditionality and warnings a non-expert would notice. Inter-rater agreement between annotators was substantial-to-perfect (Cohen's $\kappa = 0.83$–1.00); judge–human agreement was substantial-to-near-perfect ($\kappa = 0.73$–0.96) across all six cells (three jurisdictions $\times$ two rubrics). Where judge and humans disagreed, the error was predominantly judge over-counting hedging on boilerplate phrases such as generic recommendations to consult a professional. Full numbers in Table 8.

Defining overconfidence. We use "overconfidence" to mean a mismatch between the model's hedging behavior and its actual accuracy. A model that admits uncertainty at similar rates across domains of similar accuracy is well-calibrated in this sense. A model that hedges less in one domain without being more accurate there is selectively overconfident: it signals reliability it has not earned.

Data

We apply the pipeline to court decisions from three jurisdictions, using 200 tax law cases and 200 social welfare cases per country (Table 2).

Germany. Tax cases are drawn from Finanzgerichte (tax courts) and the Bundesfinanzhof (Federal Fiscal Court). Social welfare cases are from the Bundessozialgericht (Federal Social Court), covering disputes over disability benefits, unemployment insurance, and social assistance under the Sozialgesetzbuch (SGB). All decisions are in German.

France. Both domains come from the Conseil d'État, the supreme administrative court. Tax cases involve disputes under the Code général des impôts. Social cases involve aide sociale, RSA (revenu de solidarité active), disability benefits (AAH), and social security. Decisions are downloaded as XML from the open-data portal of the French administrative judiciary. All decisions are in French.

Canada. Tax cases are from the Tax Court of Canada. Social welfare cases are from the Social Security Tribunal, covering CPP disability, employment insurance, and old age security appeals. All decisions are in English.

Retrieval-augmented mitigation

If the confidence gap is caused by missing knowledge in the model's training data, then providing relevant legal text at query time should reduce it. We test this using retrieval-augmented generation (RAG): instead of answering from memory alone, the model first receives excerpts from related court decisions, then answers the question.

In our setup, this works as follows. We assemble a separate *retrieval corpus* of 100 tax and 100 social welfare decisions from the same German courts, with no overlap with the 400 decisions used for probing. We split each decision into passages of approximately 500 words. When the model receives a probe question, we search these passages for the three most relevant ones using BM25, a keyword-matching algorithm that ranks passages by term overlap with the question. These passages are prepended to the prompt as contextual information before the model generates its answer.

Crucially, the retrieval corpus never contains the specific case from which the question was derived. The model receives related legal material from the same domain (relevant statutes, legal standards, terminology) but not the answer itself. This tests whether domain-level context is sufficient to restore appropriate hedging behavior. We run this experiment on Germany only, where the confidence gap is largest, and compare the same probes with and without retrieval.

Table 3: Admitted uncertainty rates by jurisdiction and domain ($n = 600$ probes per cell, i.e., 200 cases $\times$ 3 questions). Two-proportion z-test comparing tax vs. social welfare within each jurisdiction.

Country	Tax unc.	Social unc.	z	p
Germany	49.8%	6.8%	16.53	<0.0001
France	11.2%	6.3%	2.96	0.003
Canada	24.5%	28.0%	−1.38	0.17

Statistical analysis

For each jurisdiction, we compare the proportion of probes in which the model admits uncertainty between the two domains using a two-proportion z-test. We also report mean accuracy scores across the three judging criteria to verify that any confidence asymmetry is not simply explained by differential accuracy.

Results

Confidence asymmetry

Table 3 presents the main results. In Germany and France, the model admits uncertainty at significantly higher rates on tax law probes than on social welfare probes. In Canada, no significant difference exists.

The magnitude of the asymmetry appears to track the uneven representation of particular language-domain combinations in the model's training ecology. German legal text in the area of social welfare law is likely far less visible in web-scale corpora than German tax law, which overlaps more substantially with commercial, financial, and administrative text. French legal text seems better represented than German in this respect, but still less extensively than English. In English, both legal domains are sufficiently represented for the asymmetry to disappear. This pattern offers a concrete case of language modeling bias where the issue is not simply that one language is "supported" while another

Table 4: Mean accuracy scores (1–5) by jurisdiction and domain. Test: whether the answer identifies the right legal standard. Stat.: whether the correct statutory provision is cited. Spec.: whether the answer is specific rather than vague.

		Score (1–5)		
Country	Domain	Test	Stat.	Spec.
Germany	Tax	3.03	3.83	3.79
	Social	2.65	3.09	3.42
France	Tax	2.27	2.47	3.81
	Social	2.42	2.59	3.68
Canada	Tax	3.01	3.79	4.18
	Social	2.39	3.15	3.81

is not, but that models represent and process some linguistically and institutionally specific forms of knowledge with greater precision than others. In this sense, the observed domain-specific asymmetry does not merely indicate uneven model performance but exemplifies how language modeling bias can produce epistemic injustice by differentially shaping which forms of legal knowledge are responsibly articulated and represented through AI-mediated systems (Helm et al. 2024; Loh 2024).

Accuracy scores

Table 4 shows mean accuracy scores across the three judging criteria. Within each jurisdiction, scores are comparable across domains, confirming that the confidence asymmetry is not driven by one domain being objectively easier.

In Germany, social welfare scores are slightly lower than tax scores (e.g., statute: 3.09 vs. 3.83), yet the model hedges *less* on social welfare, the opposite of what calibrated uncertainty would predict. In France, scores are essentially equal across domains, yet a significant confidence gap persists. In Canada, social welfare scores are somewhat lower, and the model hedges at comparable rates. The key finding is that accuracy differences do not explain the hedging pattern: the model is *not* hedging more where it performs worse.

Qualitative examples

To illustrate what selective overconfidence looks like in practice, we present two German probes with comparable accuracy scores but different hedging behavior.

Tax law (model hedges, accuracy 1.0/5). When asked "What does § 11 Abs. 2 VergnStGBR provide in German law?" (VergnStGBR[1] is the Brandenburg state amusement tax law, now repealed) the model responds: "VergnStGBR is not a well-known or widely cited statute in German law. I couldn't find any information on this specific statute." It then adds: "*I'm unsure about the exact content and application of § 11 Abs. 2 VergnStGBR.*" The answer scores 1 on all accuracy criteria, but the hedging language gives the user a clear signal to verify elsewhere.

[1] Vergnügungsteuergesetz für das Land Brandenburg.

Table 5: Confidence asymmetry across model families (Germany, 200+200 cases). All models use the same probe questions.

Model	Tax unc.	Soc. unc.	z	p
Llama 3.3 70B (Meta)	49.8%	6.8%	16.53	<0.0001
Qwen3 32B (Alibaba)	18.8%	14.3%	2.10	0.036
gpt-oss 120B (OpenAI)	0.5%	0.5%	0.00	1.0

Social welfare law (model does not hedge, accuracy 1.3/5). When asked "Was regelt § 160 Abs. 2 SGG?" (What does § 160(2) of the Social Courts Act provide?), the model responds confidently: "§ 160 Abs. 2 SGG regelt die Befugnis der Sozialgerichte, Beweise zu erheben" (it governs the authority of social courts to take evidence). This is incorrect; § 160(2) SGG actually governs the grounds for appeal to the Federal Social Court. Yet the answer contains no hedging language. On the contrary, it concludes: "*Ich bin mir in dieser Antwort sicher*" (I am certain of this answer). The prompting person without legal training is deceived into trusting the answer given.

Both answers are comparably wrong, but only the tax answer signals its unreliability. The social welfare user receives a confidently stated falsehood.

The language gradient

The confidence gap (tax uncertainty rate minus social uncertainty rate) decreases monotonically: 43.0 percentage points for German, 4.9 for French, and -3.5 for English (i.e., no gap). This gradient is consistent with the availability of legal text in each language. English accounts for approximately 41% of Common Crawl content, while German and French each account for roughly 5–6%.[2] Within these smaller shares, the imbalance between commercially published tax law and less widely published social welfare law is amplified, as we discuss under *Training data as structural cause* below.

Cross-model comparison

To test whether the confidence asymmetry is specific to Llama or generalizes across model families, we repeat the German experiment (200 tax + 200 social welfare cases, same questions) on two additional models: Qwen3 32B (Alibaba) and gpt-oss 120B (OpenAI). Table 5 presents the results.

Llama and Qwen exhibit the same pattern: both hedge significantly more on tax law than on social welfare law, with Llama showing a larger gap (43.0 points) than Qwen (4.5 points). The finding thus generalizes across two independently trained model families from different organizations (Meta and Alibaba), with different architectures and parameter counts.

OpenAI's gpt-oss 120B presents a different picture. It admits uncertainty on fewer than 1% of probes in either domain, despite scoring lower on accuracy than both Llama

[2] Common Crawl statistics, https://commoncrawl.github.io/cc-crawl-statistics/plots/languages, accessed April 2026.

Table 6: Mean accuracy scores by model (Germany, all domains pooled).

Model	Test	Stat.	Spec.
Llama 3.3 70B	2.84	3.46	3.60
Qwen3 32B	2.58	3.13	3.65
gpt-oss 120B	2.02	2.39	2.83

Table 7: Effect of retrieval-augmented generation on the German confidence gap.

Condition	Tax unc.	Social unc.	z	p
No retrieval	49.8%	6.8%	16.53	<0.0001
With retrieval	43.5%	37.8%	2.00	0.046

and Qwen (Table 6). This model does not exhibit selective overconfidence. It exhibits *uniform* overconfidence: confidently wrong across the board. The social welfare users we are concerned about are still poorly served, but through a different mechanism. Rather than receiving selectively worse epistemic treatment than tax law users, they receive the same poor treatment as everyone else.

These results suggest two distinct failure modes in legal AI systems. *Selective overconfidence* (Llama, Qwen) creates a differential in epistemic service: one group of users receives substantially more hedging than another. *Uniform overconfidence* (gpt-oss) harms all users equally but may be harder to detect, since cross-domain comparisons reveal no asymmetry. Our pipeline is designed to detect the first failure mode; detecting the second requires comparing absolute accuracy against hedging rates, not just the gap between domains.

Retrieval-augmented mitigation

Table 7 shows the effect of providing retrieved legal context on the German confidence gap. Without retrieval, the model admits uncertainty on 49.8% of tax probes but only 6.8% of social welfare probes, a 43-point gap. With retrieval, the social welfare uncertainty rate rises to 37.8%, closing 87% of the gap.

Tax uncertainty decreased moderately (49.8% to 43.5%), while social welfare uncertainty rose sharply (6.8% to 37.8%). The asymmetry in this response is itself informative. In tax law, where the model already has substantial knowledge, retrieved context acts as confirmation, slightly increasing confidence. In social welfare law, where the model has less knowledge, retrieved context exposes the complexity of the domain, prompting the model to recognize the limits of what it knows. A residual gap of 5.7 percentage points remains marginally significant ($p = 0.046$), suggesting that simple keyword-based retrieval does not fully close the gap. More sophisticated retrieval methods or larger retrieval corpora may reduce it further, but even minimal retrieval achieves the bulk of the correction.

One might expect retrieval to make the model *more* confident, not less: more information should mean better answers. For tax law, this is what happens. The model already has substantial knowledge of the domain, retrieval provides confirmatory context, and hedging decreases slightly. For social welfare law, the opposite occurs. Without retrieval, the model has so little domain knowledge that it does not even recognize the complexity of the area. It answers from vague associations and sounds confident. When given actual social welfare legal text, it encounters detailed statutory provisions and legal standards that reveal the gap between what it knows and what the domain requires, and it begins to hedge. The effect is analogous to a student who thinks an exam was easy until flipping through the textbook and realizing how much the subject contains. This asymmetric response to retrieval is itself evidence that the confidence gap originates in a knowledge deficit: the model's overconfidence in social welfare law reflects not a tendency toward assertiveness, but an inability to recognize what it does not know.

From a practical standpoint, retrieval thus serves a dual purpose: it may improve the accuracy of the model's answers, but even when it does not, it restores appropriate uncertainty signaling. Legal AI systems operating in underrepresented domains could retrieve domain-specific legal text before generating responses, even when the retrieved text does not directly answer the user's question.

Discussion

Selective overconfidence as an instance of "compounding injustice"

Most work on bias in AI focuses on disparities in accuracy: a system that predicts recidivism less accurately for one racial group, or a hiring tool that scores women lower than men (Allhutter et al. 2020). These disparities are serious, but they are at least *visible*. An audit can measure them. What we document here is different. The model's accuracy is comparable across domains. The disparity is not in what the model gets right, but in what it *tells users* about what it might be getting wrong. This selective overconfidence is harder to detect than accuracy bias, and potentially more dangerous, because the misinformed users have no signal that anything is amiss.

To be sure, the model itself does not discriminate against individuals; it reflects a form of language modeling bias (Helm et al. 2024), in which commercially viable legal domains leave a larger digital footprint in training corpora than domains serving vulnerable populations. A disability claimant in Germany who asks an LLM about § 160 of the Social Courts Act receives a confident, detailed, and completely wrong answer, complete with the assurance "I am certain of this answer." A taxpayer asking about an obscure provision of the Brandenburg amusement tax law receives a frank admission: "I'm unsure about the exact content and application of this statute." Both answers are wrong, but only one warns the user.

The access-to-justice literature makes clear why this matters. Low-income individuals overwhelmingly lack adequate legal assistance (Chien and Kim 2025). Self-represented litigants, who increasingly turn to LLMs for legal guidance, are precisely those "who lack any form of legal training or knowledge" and for whom "inherent deficiencies in LLMs

may mean that LLMs do more harm than good" (Ogunde 2024). When the model sounds confident, these users have neither the legal literacy to spot the error nor the resources to seek a second opinion. The selective overconfidence thus falls hardest on those least equipped to detect it. Recent litigation illustrates the stakes. A US disability claimant relied on ChatGPT to challenge a settled benefits claim, leading to *Nippon Life v. OpenAI* (Robert 2026); a self-represented UK taxpayer presented nine AI-fabricated decisions in *Harber v HMRC* [2023] UKFTT 1007 (First-tier Tribunal (Tax Chamber) 2023).

This constitutes a form of disparate impact in the sense of an unintentional statistical discrimination that disadvantages protected groups (Behrendt and Loh 2022). In our case, it is not an algorithmic discrimination strictly speaking (Barocas, Hardt, and Narayanan 2023), as the disparate impact does not directly depend on the statistical output of the LLMs, but rather on the disparate dependency on the LLM usage. However, since the LLM contributes to this disparate impact, we will refer to it as an *AI-related disparate impact*. Disparate impacts of this kind are recognized as legally relevant statistical discrimination in many legal orders. Rather than relying on anti-discrimination law, however, we argue that this disparate impact is morally wrong because it "compounds" (Hellman 2018, 2023) prior injustices that make members of protected categories more likely to lack the educational and financial means to afford adequate legal assistance. Even though the LLMs treat all users alike, the prior injustices result in a disparate impact for members of such protected groups due to historical and structural discriminations, thereby involuntarily compounding the "history of mistreatment" (Hellman 2018, p. 120) that those groups have suffered in the past.

This mistreatment can be perpetuated or exacerbated by LLMs "either by making the harm caused by the prior injustice worse or by causing it to lead to a new harm in another sphere of life" (Hellman 2018, p. 113). For this to be the case, the lower uncertainty signaling in the realm of social welfare law would have to exacerbate historical discriminations that now make members of certain protected groups more vulnerable to this unequal uncertainty signaling. In other words, they would have to be more likely to rely on LLMs for legal information with regard to social welfare law, because they are statistically worse off than the average population *as a result* of these historical discriminations. They are therefore a) more likely to have to apply for some kind of social welfare, and/or b) do not have the means to pay for professional legal information. For many protected groups, it is well researched that they in the past were victims of such a prior direct discrimination that leaves them on average more destitute than the rest of the population, e.g. through practices of racial segregation under Jim Crow, or sexist employment policies of the past (Anderson 2010; Goldin 2021; Pateman 1989; Powers and Faden 2019; Rothstein 2017). As a result, it is feasible to assume that they are subjected to both disadvantages a) and b), thereby exacerbating the mentioned historical discriminations.

In addition, while dependency on social welfare may not constitute a "new harm in another sphere of life", this may very well be true for LLMs and dependency on LLMs for legal information. The "new harm" in this case is not the being underserved with respect to legal information per se, but being underserved by a tool that promises to equalize the disadvantages caused by disparate access to professional legal information. After all, companies developing and deploying LLMs claim that they can help with all kinds of personal and professional issues. If members of historically disadvantaged groups believe these promises, they are being disadvantaged in another "sphere of life", namely their interactions with LLMs.

In order to compound a prior injustice - either by exacerbating it or by moving it to another sphere of life - the agents developing or employing the LLM do not have to be responsible for the prior injustice themselves. Instead, they violate a duty of justice to victims of prior injustice by compounding that prior injustice. As a result, developers and employers incur a moral responsibility for these disparate impacts. This includes a responsibility to due care not to compound prior injustices, and revise compounding outputs as soon as they receive notice of them.

Training data as a source for selective overconfidence

The language gradient in our results is consistent with training data composition as the dominant structural driver. LLM training corpora are dominated by English-language text, with progressively less representation for other languages (Joshi et al. 2020). Within any given language, commercially relevant domains (tax, finance, contracts) generate more publicly available text than social welfare domains (disability benefits, social assistance), which are often handled by under-resourced tribunals that publish fewer decisions and less legal commentary.

Concrete evidence supports this asymmetry, though not in the way one might expect. German social courts actually handle far more cases than tax courts, roughly six to ten times as many first-instance filings per year (Statistisches Bundesamt (Destatis) 2024). The asymmetry lies not in case volume but in *digital publication and commercial ecosystem*. The Bundesfinanzhof has published all decisions online since 2010, including those not designated for official publication (Bundesfinanzhof 2024); the Bundessozialgericht publishes only those decisions it has designated for publication. On openlegaldata.io, the largest open German legal database (Ostendorff, Blume, and Ostendorff 2020), tax courts contribute roughly twice as many decisions as the BSG (as of March 2026). More importantly, German tax law supports a large commercial publishing ecosystem: at least 15–20 specialized journals (DStR, NWB, IStR, and others), dedicated publishers (Otto Schmidt, Haufe, Stollfuss), and a professional audience of approximately 100,000 tax advisors (Bundessteuerberaterkammer 2024) who generate continuous demand for commentary, practice guides, and newsletters. Social welfare law has roughly 5–7 specialized journals and a much smaller professional readership. Web-crawled training corpora like Common Crawl favor well-linked domains with high centrality scores (Baack 2024). Commercial tax law publishers, with their extensive

web presence, are likely to score higher on such metrics than the smaller social welfare law community, amplifying the gap. The result is that LLMs trained on web data encounter disproportionately more German tax law text despite social courts handling more cases, a structural inequality that is invisible to users.

Non-English social welfare law is thus underrepresented at two levels. First, non-English languages receive less overall coverage in web-crawled corpora, which explains why no confidence gap appears in English (where both domains are well-covered). Second, within non-English languages, the commercial publishing asymmetry between tax and social welfare law means the model encounters far more of one domain than the other. The model, having seen less social welfare text, has less knowledge, but its confidence-signaling mechanisms do not proportionally adjust. The model does not know what it does not know. Our RAG experiment is consistent with this explanation: when the model is given access to social welfare legal text at query time, the confidence gap shrinks by 87%, suggesting that exposure to domain-specific text restores appropriate hedging behavior.

Implications for legal AI deployment

Our findings have practical implications for the deployment of LLMs in legal information systems that are direct results of the moral requirement of developers and deployers not to compound historical injustices:

1. **Domain-specific calibration is needed.** Current LLMs should not be deployed with uniform confidence across legal domains without domain-specific uncertainty calibration, particularly in non-English jurisdictions. LLM outputs should use uncertainty language that is appropriate to users of different demographics and educational and AI literacy backgrounds.
2. **Uncertainty audits should be standard.** Before deploying a legal AI tool, developers should audit not just accuracy but the model's uncertainty behavior across the domains it will serve.
3. **Retrieval mitigates overconfidence.** Our RAG experiment suggests that providing domain-specific legal text at query time reduces the confidence gap by 87%, even with a simple keyword-based retrieval method. Legal AI systems serving underrepresented domains should consider retrieval from authoritative legal sources as a mitigation strategy.
4. **Vulnerable populations require extra safeguards.** Legal AI tools serving social welfare domains should include explicit disclaimers or mandatory second-opinion recommendations, whether or not retrieval augmentation is in place.
5. **Prioritize public-interest governance of legal AI.** Systems that answer legal queries about citizens' rights should operate under public-benefit, non-commercial, or public-interest governance schemes with strict data protection requirements.

These implications follow from the fact that legal AI systems do not fail uniformly across domains, languages, and user groups. Domain-specific calibration is necessary because a model that appears reliable in one legal area may be overconfident in another, especially where legal texts are less available, less standardized, or less represented in training data. Uncertainty audits should therefore become part of legal AI evaluation where developers need to assess whether the model's hedging, caveats, and expressions of doubt correspond to the actual reliability of its answers. Our retrieval-augmented experiment suggests one concrete mitigation route: grounding model responses in authoritative, domain-specific legal sources can substantially reduce overconfidence, even when retrieval is implemented in a relatively simple way. However, technical mitigation alone is insufficient in domains such as social welfare law, where users may face financial constraints, language barriers, administrative complexity, or limited access to professional legal support. For this reason, systems serving vulnerable populations require additional safeguards, including clear warnings, escalation pathways, and recommendations to seek qualified information. More broadly, public-facing legal AI should not be treated as an ordinary commercial service. Because such systems may shape whether citizens understand, claim, or abandon legal rights, they should also be subject to licensing, public-interest governance, independent auditing, and strict data protection standards.

Limitations

While we test three model families, our cross-jurisdictional analysis uses only Llama 3.3 70B; future work should extend the multi-model comparison to France and Canada. The knowledge-probe pipeline uses the same model for all stages, which means hallucinations can occur at each step. In ground-truth extraction, the model may misidentify statutes or fabricate case citations from the source decision. In judging, it may misclassify whether the blind answer hedged or assign incorrect accuracy scores. These errors are likely to affect both domains similarly within each jurisdiction, since the same model processes tax and social welfare probes. If the model hallucinates at similar rates across domains, such errors add noise (reducing statistical power) but do not create directional bias. We cannot fully rule out domain-dependent hallucination rates, but the replication of the same pattern on Qwen3 32B, a separately trained model, makes it unlikely that the finding is an artifact of one model's failure mode. As an additional check on the judging step, we validate the LLM judge against two human annotators on a 50-probe stratified subsample per jurisdiction (Appendix); judge–human agreement is substantial-to-near-perfect (Cohen's $\kappa = 0.73$–0.96).

Our classification of French court decisions into tax and social welfare relies on keyword matching, which may introduce noise, though the requirement of at least two domain-specific keywords with no keywords from the other domain reduces misclassification.

We examine only tax vs. social welfare law. Other legal domains that serve vulnerable populations (immigration, criminal defense, housing) may exhibit similar patterns and warrant investigation.

Conclusion

We find that LLMs exhibit selective overconfidence in social welfare law across non-English jurisdictions: they know roughly the same amount about both domains but hedge significantly less on the one that serves vulnerable populations. The effect appears in two independently trained model families (Llama 3.3 70B and Qwen3 32B) and scales with training-data representation, creating a language-mediated gradient from German (large gap) through French (moderate gap) to English (no gap). A third model (gpt-oss 120B) reveals a complementary failure mode, uniform overconfidence across all domains, suggesting that the problem takes different forms across architectures. A simple retrieval-augmented intervention closes 87% of the gap in the most affected jurisdiction, suggesting that selective overconfidence stems largely from a knowledge deficit and pointing toward a practical remedy.

These findings raise a broader concern. The training data behind LLMs reflects not just what has been written, but what has been published, digitized, and made commercially viable online. Tax law benefits from a large advisory industry that produces vast amounts of web content. Social welfare law, which serves people who cannot afford professional representation, generates far less. When LLMs are trained on this corpus, they inherit its inequalities, and they pass them on to users in a particularly insidious form: not as visible errors, but as false confidence. The person asking about disability benefits receives an answer that sounds just as authoritative as the one about tax deductions. Only one of them comes with a warning.

Combining a knowledge-probe pipeline with the concept of compounding injustice, our findings provide empirical evidence for this extended mechanism and an interpretation of its social relevance. The model does not discriminate against users directly; it discriminates between *legal domains*. But those domains are not socially neutral. Social welfare law serves people in situations of economic vulnerability, people who are more likely to turn to automated legal tools and less able to independently assess whether an LLM's answer is reliable (Chien and Kim 2025). When the model hedges on tax law but not on social welfare law, it is not making a judgment about who deserves caution. It is reflecting the structure of its training data, in which commercially published tax commentary vastly outweighs the sparser digital footprint of social welfare jurisprudence. The result is that an asymmetry in web publishing becomes an asymmetry in epistemic service, denying the more vulnerable group the resource of knowing what the system does not know.

As LLMs become embedded in legal information infrastructure, the question is not only whether they get the law right, but whether they tell users when they might be getting it wrong. Our results suggest that, for the legal domains that matter most to vulnerable populations, they do not. Closing this gap is partly technical, partly institutional, and partly a question of who legal AI is built for.

Appendix: Validation of the admitted-uncertainty classifier

We validated the LLM judge against two human annotators on a stratified subsample of 50 probes per jurisdiction (25 tax / 25 social welfare, sampled from the production runs and balanced on the judge's labels). Each annotator independently coded every probe under two explicit rubrics: a *strict* rubric counting only literal uncertainty statements (e.g., "I am not sure"), and a *liberal* rubric also counting conditionality and warnings a non-expert would notice (e.g., "it depends on the specific facts," "consult a lawyer"). The rubrics were translated into the language of the annotated probes.

Table 8: Validation κ on 50-probe stratified samples per jurisdiction. Inter-rater = between the two human annotators; judge–human = average across the two annotators.

Jurisdiction	Rubric	Inter-rater κ	Judge–human κ
Germany	Strict	1.000	0.789
Germany	Liberal	0.880	0.820
France	Strict	0.960	0.940
France	Liberal	1.000	0.960
Canada (English)	Strict	0.940	0.768
Canada (English)	Liberal	0.834	0.917

Inter-rater κ between annotators is substantial-to-perfect across all cells, indicating the rubric is reliably teachable. Judge–human κ is substantial-to-near-perfect, with the lowest value (0.768, Canada strict) still in the substantial range. Where judge and humans disagreed, the error was predominantly one-sided: the judge labeled YES on borderline phrases (such as generic recommendations to consult a professional) that both annotators read as standard disclaimer rather than explicit uncertainty. Our reported domain hedging rates are therefore upper bounds on the true human-defined rates.

References

Ahuja, K.; Diddee, H.; Hada, R.; et al. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of EMNLP*.

Allhutter, D.; Cech, F.; Fischer, F.; Grill, G.; and Mager, A. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, 3: 5.

Anderson, E. 2010. *The imperative of integration*. Princeton: Princeton University Press. ISBN 9780691139814.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. *ProPublica*. 23 May 2016.

Baack, S. 2024. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of ACM FAccT*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and Opportunities*. Boston: MIT Press.

Barocas, S.; and Selbst, A. D. 2016. Big Datas Disparate Impact. *California Law Review*, 104: 671–732.

Behrendt, H.; and Loh, W. 2022. Informed Consent and Algorithmic Discrimination: Is Giving Away Your Data the New Vulnerable? *Review of Social Economy*, 80(1): 58–84.

Bella, G.; Helm, P.; Koch, G.; and Giunchiglia, F. 2024. Tackling Language Modelling Bias in Support of Linguistic Diversity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 562–572. Rio de Janeiro Brazil: ACM. ISBN 9798400704505.

Bender, E. M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

Bird, S.; Dudík, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. Technical report, Microsoft Research. Version: 15 May 2020.

Blodgett, S. L.; Barocas, S.; Daumé, H., III; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP.

Bommarito, M. J.; and Katz, D. M. 2022. GPT Takes the Bar Exam. *arXiv preprint arXiv:2212.14402*.

Bundesfinanzhof. 2024. Entscheidungen online. https://www.bundesfinanzhof.de/de/entscheidungen/entscheidungen-online/. All decisions (V- and NV-Entscheidungen) published since 2010.

Bundessteuerberaterkammer. 2024. Berufsstatistik der Bundessteuerberaterkammer. https://www.bstbk.de/downloads/bstbk/ebooks/Berufsstatistik-2024.pdf. 104,845 members as of January 1, 2025.

Chalkidis, I.; Garneau, N.; Goanţă, C.; Katz, D. M.; and Søgaard, A. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 15513–15535.

Chien, C. V.; and Kim, M. 2025. Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap. *Loyola of Los Angeles Law Review*, 57: 903.

Choi, J. H.; Hickman, K. E.; Monahan, A.; and Schwarcz, D. 2024. ChatGPT Goes to Law School. *Journal of Legal Education*, 71(3): 387–400.

Cui, J.; Li, Z.; Yan, Y.; Chen, B.; and Yuan, L. 2023. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *arXiv preprint arXiv:2306.16092*.

Dahl, M.; Magesh, V.; Suzgun, M.; and Ho, D. E. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1): 64–93.

Eubanks, V. 2018. *Eubanks, V: Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York, NY: St Martin's Press, illustrated edition edition. ISBN 978-1-250-07431-7.

First-tier Tribunal (Tax Chamber). 2023. *Harber v HMRC* [2023] UKFTT 1007 (TC). Judgment of Tribunal Judge Anne Redston and Ms Helen Myerscough, 4 December 2023.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2024. A Survey of Confidence Estimation and Calibration in Large Language Models. In *Proceedings of NAACL*, 6577–6595.

Goldin, C. 2021. *Career and Family: Women's Century-Long Journey Toward Equity*. Princeton University Press. ISBN 9780691226736.

Guha, N.; Nyarko, J.; Ho, D. E.; Ré, C.; et al. 2024. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *arXiv preprint arXiv:2308.11462*.

Hellman, D. 2018. Indirect Discrimination and the Duty to Avoid Compounding Injustice. In Collins, H.; and Khaitan, T., eds., *Foundations of indirect discrimination law*, 105–122. Portland OR: Hart Publishing an imprint of Bloomsbury Publishing Plc. ISBN 978-1-50991-254-4.

Hellman, D. 2023. Big Data and Compounding Injustice. *Journal of Moral Philosophy*, 21(1-2): 62–83.

Helm, P.; Bella, G.; Koch, G.; and Giunchiglia, F. 2024. Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics and Information Technology*, 26(1).

Hendrycks, D.; Burns, C.; Chen, A.; and Ball, S. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. In *Proceedings of NeurIPS*.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633: 147–150.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 6282–6293.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

Kathunia, A.; Kaif, M.; Arora, N.; and Narotam, N. 2024. Sentiment Analysis Across Languages: Evaluation Before and After Machine Translation to English. ArXiv:2405.02887 [cs].

Lai, V. D.; Ngo, N. T.; Veyseh, A. P. B.; et al. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *arXiv preprint arXiv:2304.05613*.

Legal Services Corporation. 2022. The Justice Gap: The Unmet Civil Legal Needs of Low-income Americans. https://justicegap.lsc.gov/.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*.

Lippert-Rasmussen, K. 2022. Algorithm-Based Sentencing and Discrimination. In Ryberg, J.; and Roberts, J. V., eds., *Sentencing and artificial intelligence*, Studies in penal theory and philosophy, 74–93. New York: Oxford University Press. ISBN 9780197539538.

Loh, W. 2024. Generative KI, digitale Teilhabe und epistemische Ungerechtigkeit. *Rechtsphilosophische Zeitschrift (RphZ)*, 215–233.

Loh, W.; and Seyfert, M. 2026. Generative AI and Public Reason: AI Sycophancy as Obstacle to AI Agents as Deliberative Facilitators. Cambridge University Press. Forthcoming.

Madden, M.; Gilman, M. E.; Levy, K.; and Marwick, A. E. 2017. Privacy, Poverty and Big Data: A Matrix of Vulnerabilities for Poor Americans.

Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C. D.; and Ho, D. E. 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*.

Ogunde, F. 2024. Generative AI and Access to Justice in Canada: The Case of Self-Represented Litigants. *Windsor Yearbook of Access to Justice*, 40: 211.

Olojo, S.; Zakrzewski, J.; Smart, A.; van Liemt, E.; Miceli, M.; Ebinama, A.; and Amugongo, L. M. 2025. Lost in Machine Translation: The Sociocultural Implications of Language Technologies in Nigeria. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 2249–2259. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.

Ostendorff, M.; Blume, T.; and Ostendorff, S. 2020. Towards an Open Platform for Legal Information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*.

Pateman, C. 1989. *The Disorder of women: Democracy, feminism and political theory*. Cambridge: Polity Press. ISBN 0745607896.

Powers, M.; and Faden, R. 2019. *Structural Injustice*. Oxford UK: Oxford Univ Press.

Raji, D.; Denton, E.; Bender, E. M.; Hanna, A.; and Paullada, A. 2021. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 33–44.

Robert, A. 2026. OpenAI sued for practicing law without a license. *ABA Journal*, 6 March 2026. Reporting on *Nippon Life Insurance Co. of America v. OpenAI Foundation*, N.D. Ill.

Rothstein, R. 2017. *The color of law: A forgotten history of how our government segregated America*. New York and London: Liveright Publishing Corporation a division of W. W. Norton & Company, first edition edition. ISBN 9781631492853.

Ryberg, J.; and Roberts, J. V., eds. 2022. *Sentencing and artificial intelligence*. Studies in penal theory and philosophy. New York: Oxford University Press. ISBN 9780197539538.

Selbst, A. D.; boyd, d.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 59–68.

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; et al. 2024. Towards Understanding Sycophancy in Language Models. In *Proceedings of ICLR*.

Statistisches Bundesamt (Destatis). 2024. Statistischer Bericht: Sozialgerichte 2023; Finanzgerichte 2023. https://www.destatis.de/DE/Themen/Staat/Justiz-Rechtspflege/. Published July 2024.

Susskind, R. 2023. *Tomorrow's Lawyers: An Introduction to Your Future*. Oxford: Oxford University Press, 3rd edition.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of ICLR*.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? In *Findings of the Association for Computational Linguistics: ACL*, 8653–8665.