LEXCHEX: Evaluating LLM Knowledge of Local Law

Denis Peskoff[1,2,*] **Joe Barrow**[3,*] **Teresa Choi**[1] **Faith Wu**[1] **Diag Davenport**[1,2]
[1] UC Berkeley [2]School of Information [3]Independent Researcher [*]Equal contribution
{dpeskoff, diag}@berkeley.edu

Abstract

What does it mean for a large language model (LLM) to "know" the law? We study this question in the context of local laws—an extensive, legally binding body of public law that is fragmented across jurisdictions and often distributed through private platforms with restrictive access and reuse constraints. This setting provides a natural test of access-sensitive knowledge: information that is public and important, yet unevenly represented in model training data and variably accessible at inference time. We introduce LEXCHEX, a large-scale evaluation of local-law question answering consisting of 15,600 questions spanning counties in California and Florida. Questions span major regulatory domains and are inspired by specific statutory provisions. We evaluate leading models from OpenAI, Google, and Anthropic under three access regimes: model-only prompting, open-web search, and corpus-grounded legal retrieval. This design lets us distinguish, operationally, between knowledge stored in model weights and answer quality made possible by inference-time access. Our evaluation focuses on three complementary dimensions: *correctness* (whether answers match governing law), *decisiveness* (whether models provide clear, actionable answers rather than retreating into ambiguity), and *agreement* (the extent to which different models converge on the same answer). We reach four conclusions. First, web search impacts correctness modestly and unevenly across models. Second, agreement is not the same as knowledge: models often converge even when they are not reliably correct, suggesting that shared answers may reflect shared priors rather than recovery of the governing ordinance. Third, corpus-grounded RAG reveals a second bottleneck: retrieval alone does not reliably produce grounded legal answers. Fourth, failures are patterned rather than random, varying by legal domain and county prominence. Taken together, these results show that performance on local law is best understood not as a single latent capability, but as the joint product of parametric storage, inference-time access, and successful grounding against fragmented legal sources. More broadly, we position local law as a tractable and policy-relevant testbed for studying when public information becomes usable machine knowledge.

1 Introducing Local Laws to the NLP Community

Large language models (LLMs) are increasingly evaluated on whether they can reason about law. Existing work has shown strong performance on contracts, merger agreements, bar-style exams, legal briefs, and broad legal benchmark suites (Hendrycks et al., 2021; Wang et al., 2023; Katz et al., 2024; Guha et al., 2023; Fei et al., 2024; Woo et al., 2025). But these settings leave open a more basic question: *can models answer ordinary questions about the local rules people actually live under?* This question matters practically, because many real interactions with law do not occur in appellate opinions or standardized legal corpora. They occur when a person asks whether they may park on a street overnight, keep an animal on a property, rent a unit, dispose of waste in a particular way, operate a home business, or modify land use. Federal and state law matter profoundly, but often episodically. Local law is different: it governs repeated features of ordinary life. Over a life course, direct encounters with local law are therefore not edge cases; they are close to universal.

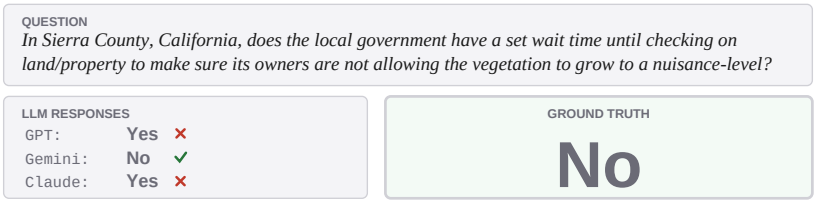

Figure 1: A question from our dataset with the respective answers.

From a machine-learning perspective, local law is more than a socially important domain. It is a revealing testbed for *usable knowledge*. The central premise of this paper is that public law is not the same thing as accessible machine knowledge. County-level ordinances are formally public and legally binding, yet digitally they are fragmented, unevenly updated, and often published through non-uniform interfaces operated by private vendors. Copyright and terms of use further complicate whether these laws are legally collectible at scale for pretraining, benchmarking, or retrieval. As a result, answering a question about local law may require more than generic legal reasoning. It may depend on whether the relevant rule was ever internalized during pretraining, whether the current interface can retrieve it at inference time, and whether the model can verify that the retrieved text is applicable.

At first glance, local law might seem unusually favorable for evaluation. Unlike many open-ended tasks, many local-law questions are grounded in a specific ordinance or a small set of provisions, so there is a relatively determinate source of truth. One might therefore expect that giving models better access to the same governing text would act like a common signal and increase convergence across models. Our results suggest a more subtle picture. In the language of Blackwell informativeness, richer access can make a given decision-maker better informed on a fixed task without making heterogeneous decision-makers more similar, because the effective signals they receive need not be identical. Different systems may store different fragments of local law in their weights, issue different queries, retrieve different webpages or statutory sections, receive different surrounding context, and apply different thresholds for when evidence is sufficient to justify an answer. Better access need not monotonically increase agreement even when it improves average correctness.

This distinction connects local law to several broader questions at the heart of current language-model research. Prior work on parametric knowledge asks how much factual information can be stored reliably in model weights, especially for long-tail or obscure facts (Roberts et al., 2020; Kandpal et al., 2023). A parallel line of work on retrieval-augmented generation shows that adding external access can improve factual performance, but also shifts the problem toward search, grounding, and verification (Lewis et al., 2020). Existing legal NLP benchmarks have substantially advanced the evaluation of legal reasoning and domain knowledge (Guha et al., 2023; Fei et al., 2024; Zheng et al., 2021), yet they primarily operate over comparatively centralized or already-curated legal corpora. Local law exposes a different failure mode: knowledge may be public, important, and binding, but still remain difficult for models to internalize, retrieve, or verify. Empirically, we find that correctness and agreement move differently across access conditions: web search changes answers at different levels which causes cross-model agreement to notably fall, and a corpus-restricted legal RAG setting does not fully restore consensus. We interpret these patterns cautiously as consistent with a shift from shared priors toward model-contingent retrieval and evidence-conditioned interpretation. That makes this setting useful not only for legal AI, but for studying model stability over obscure facts and for evaluating retrieval-augmented systems in access-constrained, effectively zero-trust environments.

We study this problem through LEXCHEX, a benchmark for county-level local-law question answering in the United States. We focus first on California and Florida. This choice is not arbitrary convenience sampling, but a strategic contrast set. California and Florida are large, policy-salient, and institutionally distinct state ecosystems, with substantial heterogeneity across counties and markedly different political and regulatory contexts. Together they

provide a meaningful initial comparison before scaling nationally. Our goal is not to treat ideology as an estimand. Rather, it is to test whether failures of legal knowledge persist across two substantively different local-law environments. Methodologically, we design the benchmark to make access conditions part of the experiment rather than an unexamined assumption. Because local ordinances are not uniformly available as open, machine-ready corpora, we manually collect the relevant laws, convert them into machine-processable text, and construct FEVER-style, evidence-grounded questions inspired by actual local laws (Thorne et al., 2018). Each human-written question is systematically prefixed with each county name in our study. This design preserves legal specificity while avoiding the need to redistribute the underlying corpus as a standardized benchmark artifact. It also mirrors the way end users actually interact with legal information: through concrete questions whose answers must be anchored in particular provisions.

Our evaluation protocol separates what models appear to know from what they can access. We therefore compare leading LLMs under three access regimes: a model-only setting that better isolates behavior attributable to parametric memory, a web-search setting that allows models to access outside information at inference time, and a corpus-grounded legal RAG setting that restricts retrieval to the county laws we collect. We do not claim to observe training data directly. Instead, the comparison provides an operational way to study how answer behavior changes when inference-time access changes. We evaluate each system along three complementary axes. *Correctness* measures whether an answer matches the governing law. *Decisiveness* measures whether the model gives a clear, unambiguous answer rather than retreating to vague language such as "it depends" or "unclear." This dimension is important because in legal settings, a system can fail not only by being wrong, but also by being too indeterminate to be useful. Finally, *agreement* measures the extent to which different models converge on the same answer. Agreement is diagnostically useful precisely because it need not move in lockstep with correctness when access changes.

This paper makes three contributions. **(1)** We identify county-level local law as a new testbed for *access-sensitive machine knowledge*: information that is public and binding, yet difficult to internalize and unevenly accessible at inference time. **(2)** We introduce LEXCHEX, a benchmark that operationalizes this setting through grounded local-law question answering across two large and heterogeneous state ecosystems. **(3)** We present an evaluation framework that compares parametric, web-search, and corpus-grounded performance, and measures not only correctness, but also decisiveness and cross-model agreement in a partially trusted retrieval setting. Taken together, our aim is not simply to ask whether LLMs "know the law." It is to study when public information becomes usable model knowledge, when it does not, and how that boundary should be evaluated when the same law is neither uniformly represented in model weights nor frictionlessly accessible at inference time. Section 2 motivates the importance of studying local laws. Section 3 discusses the complications that arise from copyright and terms of use in studying the law. Section 4 introduces LEXCHEX, a dataset designed to question LLMs about their knowledge of local laws. Section 5 discusses the LLMs we choose to investigate and our methodology. Section 6 presents our results. Finally, Section 7 discusses what these findings imply for evaluating LLMs on local law and other fragmented, legally consequential public corpora.

2 Why Local Law is a Revealing Testbed

Local laws have been understudied relative to federal laws (Chalkidis et al., 2020; Henderson et al., 2022). Local laws are *predominantly* set at the county level. For this bellwether analysis, we prioritize the states where home rule creates strong counties.[1] Maryland serves as a mid-Atlantic foil, which we use for priming, to a heavily liberal California and a heavily conservative Florida in our study (Appendix B). Counties are themselves required to publish the laws on limited resources. This causes them to use third-party technical providers, which leads to copyright and terms of use issues, which are discussed in Section 3.

[1] Per Cornell Law's Wex "is a provision of the state constitution . . . granting a local municipality a certain amount of autonomy to allocate powers between the state and the local government."

2.1 Which Themes are Important in Local Laws

We focus our attention on six categories (or "themes") of local law. These legal themes are a particularly useful lens for studying LLM "knowledge" because they concentrate precisely the kind of information that is public, binding, consequential, and yet difficult to treat as stable machine-accessible fact. Public space, noise and nuisance, land use, housing, business licensing, and building and safety govern recurring features of ordinary life: where people may gather, what uses of property are permitted, when conduct becomes sanctionable, and what conditions attach to operating a dwelling or business. They therefore capture law as it is actually encountered by most people—not primarily through appellate opinions or elite legal disputes, but through the local rules that structure everyday action. At the same time, these domains are typically encoded in fragmented county ordinances rather than centralized legal corpora, making them an especially revealing setting in which to distinguish between information that is formally public and information that is actually available to a model in usable form. Furthermore, they induce variation in the kind of competence an LLM must exhibit. Some questions in these domains have crisp, operational answers; others are conditional, exception-laden, or highly jurisdiction-specific. Many also invite strong commonsense or stereotype-based priors that may sound plausible while still being legally wrong in a particular county. As a result, performance on these categories is informative not only about correctness, but about the source of apparent knowledge itself: whether a model has internalized the relevant rule in its parameters, can retrieve it reliably at inference time, and can distinguish governing text from generic legal expectation. In that sense, these categories do more than organize the benchmark substantively. They help make visible a broader point about knowledge: knowing is not merely producing a plausible answer, but being able to access, ground, and verify the right answer under the institutional conditions that determine whether the relevant information is actually reachable.

3 Access Constraints in Local Law

One would expect the law to be publicly accessible, and an excellent domain for research. However, copyright and Terms of Use requirements paired with a consolidation of tech companies has made this nonviable. The laws for 57 out of 58 of the California and 20 out of 20 Florida counties we study are maintained (and controlled) by a single company.

Copyright. Copyright can stifle innovation if applied broadly. If there was any sort of technical service provided to the counties that manipulated the law or the data, this would be a valid protection of a technology company's intellectual property. We found no evidence of textual or thematic standardization, other than of the URL. As a glaring example, California's counties primarily have single column documents while Florida's are two column. Sections do not have common headers and the comparison of laws across counties requires substantial work. The copyright conditions of the county laws state:

> *Content may not be copied, reproduced, modified, published, uploaded, posted, transmitted, performed, or distributed in any way, and you agree not to modify, rent, lease, loan, sell, distribute, transmit, broadcast, or create derivative works based on the Content, the Site, or any Solution, in whole or in part, by any means.*

Terms of Use. Furthermore, the terms of use for these county laws require that automated systems do not request more messages than a human can reasonably produce. Full text is provided in Appendix A, alongside a discussion of robots.txt and LexisNexis.

4 LEXCHEX: A Benchmark for Jurisdiction-Specific Legal QA

Given the copyright and terms of use restrictions, using the data directly is a major barrier to assessing LLMs. This both motivates our research direction, and motivates the creation of this dataset as a way to provide researchers with preliminary insight into the content of American local laws.

Example yes/no questions.
Does the seller near a military installation notify the buyer of high noise levels pre-sale?
If graffiti appears on my property that I didn't create, must I remove it?
Are regulations established for keeping hens on owner-occupied property?
Example Short Form questions.
What must property owners do if rodent infestations occur on premises?
What is the minimum distance for commercial/industrial buildings from residential-zoned property?
What permit is required to sell goods in public parks?

Table 1: These are questions across different themes, representing both yes/no and short-form style questions. We intentionally encourage unique styles in writing the questions.

Question Answering is a popular way to test the knowledge of models (Iyyer et al., 2014; Rajpurkar et al., 2016; Kwiatkowski et al., 2019). Our largest dataset decision choice is between setting up literal questions like in FEVER (Thorne et al., 2018) or creating a claim verification system like in SciFact (Wadden et al., 2022). Ultimately, we choose to align our task with the expected use case: citizens subject to laws are more likely to ask questions—both direct and open-ended—about the law, rather than engage in counterfactual studies of the law. We encourage alternative methods of evaluation in future work.

4.1 Collecting and Extracting the Laws

All of the laws referenced in our analysis can be found publicly. We *manually* download the laws to comply with any terms of use issues regarding scraping. These downloaded files can range thousands of pages in length and contain the ordinances we are interested in, as well as multi-page tables, lengthy procedural descriptions, and other information less salient to a county resident. We obtain all the laws as PDFs, typically between one thousand and three thousand pages. In order to consistently extract and structure all of laws from these PDFs, we develop a single "OCR then post-process" pipeline. We simultaneously OCR and convert the documents to markdown using LightOnOCR-2-1b (Taghadouini et al., 2026), a 1 billion parameter vision language model built atop Qwen-3 (Bai et al., 2025; Yang et al., 2025). We post-process the markdown to remove page artifacts — such as page headers, footers, and page numbers — and split the documents at likely headings. The entire OCR-then-post-process pipeline is available as part of the code and data release.

4.2 Writing the Questions

We create a dataset broadly following FEVER (Thorne et al., 2018) that includes: (a) yes/no questions, and (b) short-form questions.

Priming Question Writers. How local laws are written cannot be assumed to be common knowledge. To provide perspective on both the style and the content to question writers, we collect every county ordinance in Maryland, a highly county-oriented less polarized state for grounding our analysis. From our processed text data, we annotate every paragraph using OpenAI GPT 5.4 and identify the relevant laws, and classify them according to theme. A substantial amount are characterized as "other", for follow-up research. Our prompt is provided in Appendix C. We sample 50 from each of the 6 themes. The themes are not equally distributed with 100 laws pertaining to Noise Disturbance but 1,600 related to Land Use, so a stratified sample is important to avoid creating a question set centering around land and buildings. A user assigned to business licensing received the sample in Table 2.

County	Function	Text
Allegany	Rules	No person shall sell or solicit orders for the sale of any merchandise of any description, including books or periodicals, from door to door of private residences without…
Howard	Enforce	(a) *Dispatch Records*. The Department of Police and the Bureau of Communications shall record, for each alarm signal: (1) The date and time of receipt of the dispatch request;…
Anne Arundel	Rules	A person engaging in the business of plumbing under this subtitle may not undertake to do any plumbing work within the County unless the person carries general liability insurance in the amount of $300,000 …

Table 2: Three of 50 laws provided as a Business Licensing sample. Priming the writer encourages detailed questions by providing perspective on the content and format of laws.

Who Writes the Questions? Our pool of users is a dozen undergraduate and masters students at UC Berkeley engaged in legal NLP research. Students are asked to write 50 questions, half yes/no and half short form and are given 10 days. Quality over speed is encouraged and use of LLMs for question writing is strictly prohibited.

4.3 Quality Assurance

Our final dataset contains 15,600 questions; half are written to be yes/no and half are written to short form, both on the same topic. These are derived by prefacing a geographic qualification (e.g., "In Orange County, Florida") to each of the 200 unique questions. We release all questions, along with LLM answers, for both reproducibility and further study.

Paper authors, who have written over 25 questions themselves, perform a second level review of questions on newer question writers. Questions that are highly-similar, ambiguous, or compound in nature are flagged and sent for revisions. Relative to a LLM, we believe having a dozen unique question writers better reflects the diverse ways in which people inquire about the law. After encouraging preliminary questions with minimal guidance outside of question requirements and the samples, we provide further feedback of examples that are good and bad questions. After the final version, we aggressively cull over half the questions that are non-atomic, leading, or vague.

5 Experimental Design: Comparing Access Regimes

Parametric Condition. If the laws are this hard for researchers to access, are they equally hard for LLMs to use? We are curious whether the hosting infrastructure of local laws affects how those laws are represented in LLMs. We investigate the leading three proprietary LLMs: GPT 5.4, Gemini 3.1 Pro, and Claude Opus 4.6.

Open-web Search Condition. In contrast, appending a search tool—and crucially informing the LLM that it has a search tool— allows the model to search for information outside of its training data. OpenAI, Google, and Anthropic provide `web_search` tools, which enable models to search the web before returning a response. Given our sample size, we allow web-searches free rein—Claude uses 150 million tokens over 6,000 searches—to see the full extent of current capabilities.

Corpus-grounded RAG Condition Our second search condition gives the models access to an index that contains just the county laws in order to answer the questions. We encode each law with a multivector retrieval model, namely MixedBread's 32 million parameter ColBERT model (Clavié et al., 2025; Takehi et al., 2025), and index them for retrieval with PyLate (Chaffin & Sourty, 2024). We supply a simple agentic harness with two tools: search,

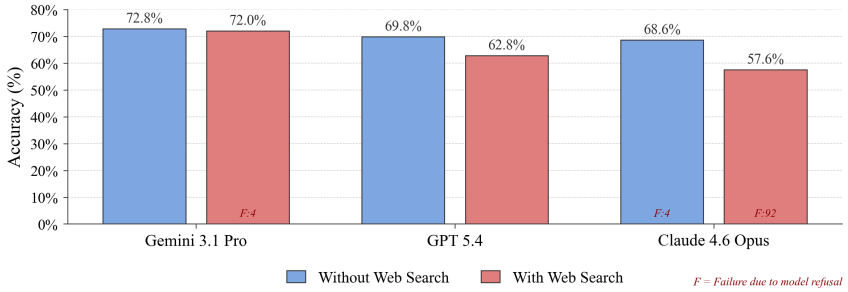

Figure 2: Accuracy of three frontier LLMs on ternary county-ordinance questions. Each model is shown with and without web-search.

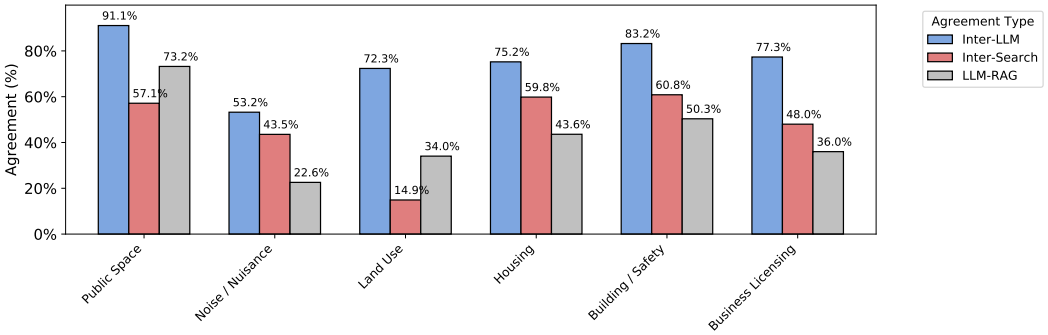

Figure 3: LLMs have strong preconceived biases about the law. Searching, especially about land use or business licensing, causes them to change their preconceived notions.

which takes a query string, and answer, which returns a structured answer and a list of sections provided as evidence. Each model may take up to 8 search steps.

Evaluation Given the cost—in time for annotation and money for particularly search computation—we create a 500 row stratified sample that covers every county and every possible question. For approximating correctness, we compare results to a human annotator, a paper author with past question writing experience motivated to be meticulous in evaluation. For decisiveness, we see what ratio of answers are Yes/No rather than Unknown. For agreement, we care about the answer being the same. For short-form questions we focus on decisiveness and agreement given the complications of short form evaluation.

6 Results

We present four empirical patterns from our experiments that, taken together, support the central claim of the paper: whether an LLM appears to "know" local law depends not only on general reasoning capacity, but on the interaction between parametric memory, inference-time access, and the ability to ground answers in fragmented legal sources.

Search changes outputs, but modestly and unevenly. Figure 2 shows that web search changes accuracy for every model we evaluate and increases error. Even with search enabled, accuracy remains well below a level that would justify high confidence in model outputs for real legal use. More substantively, the figure suggests that performance in this setting is not determined solely by general reasoning ability: it also depends on whether the model can locate, interpret, and ground the relevant local rule under fragmented access conditions. Search therefore appears to function less as a substitute for legal knowledge than as a noisy complement to it, slightly changing outputs while still leaving substantial room for retrieval failure, misinterpretation, and incomplete verification.

Agreement is not the same as knowledge. Figures 3 and 2 jointly show that convergence across models should not be mistaken for genuine legal knowledge. In several domains, models agree with one another at relatively high rates even when overall accuracy remains modest. Without search, inter-LLM agreement is especially high—reaching 91% in Public Space and 72% in Land Use—despite the fact that model-only accuracy in Figure 2 remains far from perfect. This pattern suggests that shared answers can arise from shared priors, similar heuristics, or common training regularities rather than from correct recovery of the governing ordinance. Put differently, models can be confidently aligned with one another while still being jointly wrong.

The contrast becomes even sharper once search is introduced. Web search increases accuracy on average, but it substantially reduces inter-model agreement in most categories, with especially large declines in Land Use and Business Licensing. At the same time, agreement with RAG-grounded answers remains only moderate, again particularly low in Land Use and Business Licensing. Taken together, these results imply that consensus is an ambiguous signal in access-constrained legal settings. High agreement without search may reflect the stability of a shared stereotype rather than true knowledge, while lower agreement with search may reflect models encountering the underlying legal heterogeneity more directly. The broader lesson is that knowledge in this setting cannot be inferred from model consensus alone; it must be evaluated against grounded legal evidence, because what looks like agreement may simply be coordinated plausibility rather than correct understanding.

RAG reveals a second bottleneck: retrieval is not enough. If the central problem were merely lack of access, then retrieval over the relevant county ordinances should largely close the gap. Our results suggest otherwise. Even when models are given access to a corpus restricted to the relevant legal materials, agreement with RAG-grounded answers remains only moderate and varies substantially across regulatory themes. In Figure 3, agreement with RAG is 73% in Public Space, but falls to 50% in Building / Safety, and just 23% in Noise / Nuisance. These are not patterns one would expect if retrieval alone were sufficient to produce reliable legal answers.

This matters because the RAG setting is in some sense favorable to the models. Unlike open web search, it narrows the search space to the relevant county-law corpus and reduces the problem of discovering where the law might live online. The remaining failures therefore point to a second bottleneck downstream of access itself: models must still translate a legal question into an effective search strategy, identify the relevant provision, determine whether the retrieved text is complete and controlling, and map that text onto a correct answer. In legal domains such as land use, housing, and business licensing—where rules are often conditional, cross-referenced, and exception-laden—these interpretive and grounding demands appear to remain substantial even after retrieval succeeds mechanically. The broader implication is that access and retrieval should not be conflated with knowledge. Retrieval can expose the relevant text without yielding correct legal application, just as parametric fluency can produce plausible answers without grounding. In this sense, RAG helps reveal that local-law question answering is constrained by at least two separable bottlenecks: whether the relevant law can be reached at all, and whether the model can correctly use what it has reached. For evaluating legal AI systems, this distinction is important. A model that fails without access exhibits one kind of limitation; a model that still fails after access exhibits a deeper one.

Short form questions provide knowledge on the Unknown. Repeating LLM searches on short form questions reveals an interesting model design difference. Gemini and GPT return nearly half the answers as unknown, while Claude answers over 80% of them. Hence, only 14% of the questions are unanimously Unknown.

Failures are patterned by theme and county prominence. Errors are not randomly distributed across questions or jurisdictions. Instead, they exhibit systematic structure both across regulatory themes and across counties of differing prominence. As shown in Figure 3, domains such as Land Use and Business Licensing exhibit lower agreement and greater instability, consistent with their higher legal complexity, conditionality, and reliance on

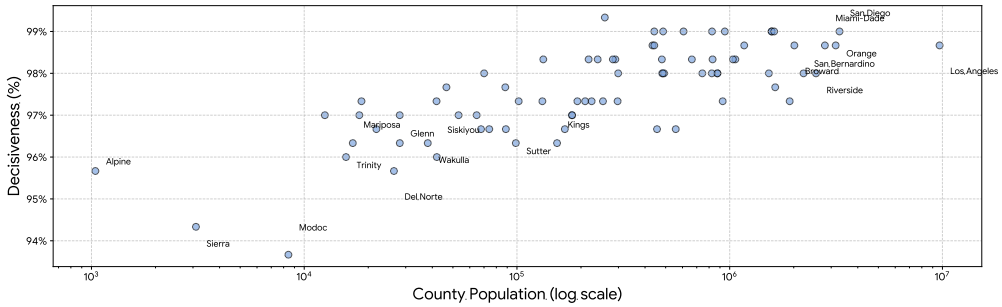

Figure 4: LLMs are more certain about larger and wealthier counties

cross-referenced provisions. At the same time, Figure 4 shows that models are more decisive for larger counties, which are also more likely to be represented in training data and more accessible through search.[2] This creates an asymmetry: models are most confident precisely where the law is more visible, and less stable where it is more fragmented or obscure. Taken together, these patterns suggest that failure is shaped jointly by the structure of the legal domain and the visibility of the jurisdiction. Local-law knowledge, in this sense, is not missing at random—it is systematically weaker in places where access is harder and rules are more complex, and stronger where both exposure and retrieval are more favorable.

7 When Public Law Becomes Usable Machine Knowledge

So what does it mean for an LLM to "know" the law? Our results show that, at least for local law, the answer cannot be reduced to a single capability. County ordinances are public, binding, and consequential, yet they are fragmented across jurisdictions, unevenly represented online, and often difficult to access in machine-usable form. Here, model performance is best understood not as a direct readout of abstract legal reasoning ability, but as the product of several interacting conditions: whether the relevant rule was internalized during training, whether it can be reached at inference time, and whether the model can correctly ground and verify what it finds. Across our experiments, accuracy is imperfect, decisiveness is inconsistent, and agreement is incomplete. Search can change predictions in some cases, but does not eliminate error; agreement across models can remain high even when correctness is limited; and retrieval over the legal corpus itself does not fully resolve failure. Taken together, these patterns are consistent with a simple but important conclusion: public information does not become usable machine knowledge merely by being online.

Much current discussion of model "knowledge" treats knowing as if it were equivalent to producing a plausible answer. In access-constrained domains, usable knowledge depends on whether a system can reliably reach the governing source and apply it correctly to a concrete question. What looks like knowledge in one interface condition may disappear in another; what looks like consensus may reflect shared priors rather than grounded understanding; and what looks like successful retrieval may still fail at the point of legal interpretation. Local law is therefore valuable not only because it is socially important, but because it makes these distinctions visible in unusually sharp form.

We do not claim that local law exhausts the problem of legal knowledge, nor that our results generalize mechanically across all legal systems or domains. But they do identify a tractable and policy-relevant testbed for studying when public information becomes machine-usable and when it does not. If the broader goal is to build AI systems that can answer real questions about the rules people actually live under, then the challenge is not simply to make models more fluent, or even more accurate in the aggregate. It is to understand the conditions under which legally operative text can be transformed into reliable, verifiable, and appropriately bounded machine knowledge. Local law gives us a concrete place to study that boundary. Our results suggest that today's systems have not crossed it.

[2]Based on official population from US Census and GDP from Bureau of Economic Analysis.

Ethics Statement

Studying the law is at its core an ethical issue. Navigating data access as researchers in computational social science and natural language processing has become increasingly challenging. We create additional data to encourage further study, which we also hope will draw attention to this issue.

Our paper is not originated nor written by LLMs. LLMs are used for code assistance, figure beautification, and obviously for evaluation. Data is verified, such as for the population of counties.

Access to Large Language Model research currently requires non-trivial funds: searching 15,000 times with Tavily, a leading LLM search tool, costs $100, our Anthropic and Gemini search experiments cost $500. Furthermore, hidden rate limits on these models make large scale experimentation complicated: Anthropic web-search was limited to a semaphore of 3 with an entire minute back-off when we exceeded the token rates per minute. We hope to provide some knowledge about challenges like these to both technical and general communities.

We investigate local laws in the United States, which are written in English which is a limitation of this work. Follow-up work should investigate local laws written in the minority of states where the city predominantly dictates the laws, and local laws outside of the United States. The legal systems, and the hierarchy thereof, varies across countries so our results are not directly transferrable to the local laws of another country.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Antoine Chaffin and Raphaël Sourty. Pylate: Flexible training and retrieval for late interaction models, 2024. URL https://github.com/lightonai/pylate.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904. Association for Computational Linguistics, November 2020. URL https://aclanthology.org/2020.findings-emnlp.261.

Benjamin Clavié, Sean Lee, Rikiya Takehi, Aamir Shakir, and Makoto P. Kato. Simple projection variants improve colbert performance, 2025. URL https://arxiv.org/abs/2510.12327.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7933–7962. Association for Computational Linguistics, 2024.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Peter Henderson, Mark S Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel E Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 29217–29234, 2022.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 13328–13340, 2021.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 633–644. Association for Computational Linguistics, 2014. URL https://aclanthology.org/D14-1070.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. URL https://aclanthology.org/Q19-1026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. URL https://aclanthology.org/D16-1264.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 5418–5426, 2020.

Said Taghadouini, Adrien Cavaillès, and Baptiste Aubertin. Lightonocr: A 1b end-to-end multilingual vision-language model for state-of-the-art ocr, 2026. URL https://arxiv.org/abs/2601.14251.

Rikiya Takehi, Benjamin Clavié, Sean Lee, and Aamir Shakir. Fantastic (small) retrievers and how to train them: mxbai-edge-colbert-v0 tech report, 2025. URL https://arxiv.org/abs/2510.14880.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734. Association for Computational Linguistics, December 2022. URL https://aclanthology.org/2022.findings-emnlp.347.

Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. In *Proceedings of*

the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 16369–16382. Association for Computational Linguistics, 2023.

Jesse Woo, Fateme Hashemi Chaleshtori, Ana Marasovic, and Kenneth Marino. BriefMe: A legal NLP benchmark for assisting with legal briefs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13139–13190. Association for Computational Linguistics, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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 (ICAIL)*, pp. 159–168. ACM, 2021.

A Appendix

Robots.txt. Municode's *robots.txt* suggests that it uses Drupal; it also requests a high crawl-delay of 15. AI scraping has pushed most websites away from specifying any crawl-delay; https://www.reddit.com/robots.txt stwhen we exceeded dates "*Reddit believes in an open internet, but not the misuse of public content.*" and has a "disallow". For a government comparison, https://www.nasa.gov/robots.txt is set to 1.

Full Copyright: CivicPlus owns and retains all copyrights in its proprietary Content. Except as specifically permitted on the Site or any Solution as to certain CivicPlus Content, the CivicPlus Content **may not be copied, reproduced, modified, published, uploaded, posted, transmitted, performed, or distributed in any way, and you agree not to modify, rent, lease, loan, sell, distribute, transmit, broadcast, or create derivative works based on the Content, the Site, or any Solution, in whole or in part, by any means.** Customer does not receive any right or license to use the foregoing. CivicPlus may use and incorporate into the Site or the CivicPlus Service any suggestions or other feedback you provide, without payment or condition.

Full Terms of Use:

Use or launch any automated system, including without limitation, "robots," "spiders," or "offline readers," that accesses the Site or any Solution in a manner that sends more request messages to the CivicPlus servers in a given period of time than a human can reasonably produce in the same period by using a conventional on-line web browser, unless specifically permitted by CivicPlus and in accordance with any specific rate or use limitations.

LexisNexis. LexisNexis is the largest owner of legal data in the United States, and poses another avenue to accessing the data. A university representative approximated our data request to cost $10,000, which would have been prohibitive for our research. Furthermore, this would be subject to the same terms of use limitations and not be reproducible.

B State Selection

Maryland has the strongest county law. While it consistently votes Democrat since the 1990s, it does not have a strong association in the US public the way California or Florida

does. Additionally, it has a population of 6 million and $500 billion GDP. The population is close to average for the United States and the GPD is above average. These characteristics make it a solid choice for grounding our question writers.

California and Florida represent both an idea of America outside of America, and represent a significant share of internet search traffic within America. Despite their similarities in relative name-recognition, population, and Gross Domestic Product, they have dramatically different political views, which is likely to extend into their law-making. By the numbers, **California** is liberal and has not voted Republican since 1988 in a presidential election. California has 58 counties and an approximately 4 trillion GDP and 39 million population. **Florida** is conservative and while its presidential voting history is more stratified, the 2024 election represented the largest margin for a Republican candidate since the same 1988 election. Florida has 67 counties but only 20 are home rule counties. However, these 20 counties house the majority of the population. The state has approximately a 1.8 trillion GDP and a 20 million population.

C LLM Prompt for Legal Text Classification

The following system prompt and user instructions are utilized to categorize municipal and county codes. The language model is instructed to analyze the provided text and output the results as a strictly formatted JSON object.

```
# System prompt forcing structured JSON output
SYSTEM_PROMPT = """You are a legal text classifier specializing in
    municipal and county codes. You must output a JSON object with four
    keys: "is_substantive" (integer), "primary_function" (string), "sub-
    category" (string), "logic" (string)."""

USER_INSTRUCTIONS = """
Task: Categorize the provided text based on its Primary Legal Function.
    Choose the most appropriate category from the five below:

    Context: Defines terms, establishes scope, or provides introductory
    intent. (e.g., "For the purposes of this chapter, 'Pet Shop' means...
    ")

    Rules: Imposes permissions, obligations, or prohibitions on
    individuals or entities. (e.g., "No person shall operate a pet shop
    without a license.")

    Process: Describes administrative procedures, assigns authority, or
    dictates government structure. (e.g., "The Board shall appoint
    members for three-year terms.")

    Enforcement: Specifies penalties, identifies violations, or outlines
    exceptions and appeals. (e.g., "Violation of this section is a Class
    C misdemeanor.")

    Structural: Non-substantive text including Tables of Contents, HTML
    tags, History/Source notes, Page Numbers, or Section Headers. (e.g.,
    "[HISTORY: Adopted...]", "ARTICLE 1", "<td>...</td>")

Output Requirements (Return as JSON):

    "is_substantive": [1 if the text is a "Rule" or "Enforcement" action;
     0 if it is Context, Process, or Structural.]

    "primary_function": [Context / Rules / Process / Enforcement /
    Structural]

    "sub_category": (Only if is_substantive is 1) Choose from one of the
    6 categories: Land use, Noise/Nuisance, Housing, Business licensing,
```

```
    Public space, Building/Safety. If it does not fit in any then record
    it as Other. If is_substantive is 1, you must assign one of these
    options, verbatim. If is_substantive is 0, leave this field empty or
    null.

    "logic": A brief one-sentence explanation.

Text to Classify:
[INSERT TEXT HERE]
"""
```

D Example Document

Alameda – 1.04.210

Section Title: 1.04.210 – Person.
Page Numbers: [9]
State: California
County: Alameda
Index: 38

"Person" includes a natural person, firm, partnership, co-partnership, association, organization, company or corporation.

(Prior gen. code § 1-2.9)