

Law is vulnerable to AI influence; interface design can help

Aileen Nielsen, Chelse Swoopes, Elena Glassman¹

Abstract

As large language models (LLMs) enter judicial workflows, courts face mounting risks of uncritical reliance, conceptual brittleness, and procedural opacity in the unguided use of these tools. Jurists' early ventures have attracted both praise and scrutiny, yet they have unfolded without critical attention to the role of interface design. This Essay argues that interface design is not a neutral conduit but rather a critical variable in shaping how judges can and will interact with LLM-generated content. Using Judge Newsom's recent concurrences in *Snell* and *Deleon* as case studies, we show how more thoughtfully designed, AI-resilient interfaces could have mitigated problems of opacity, reproducibility, and conceptual brittleness identified in his explorative LLM-informed adjudication.

We offer a course correction on the legal community's uncritical acceptance of the chat interface for LLM-assisted work. Proprietary consumer-facing chat interfaces are deeply problematic when used for adjudication. Such interfaces obscure the underlying stochasticity of model outputs and fail to support critical engagement with such outputs. In contrast, we describe existing, open-source interfaces designed to support reproducible workflows, enhance user awareness of LLM limitations, and preserve interpretive agency. Such tools could encourage judges to scrutinize LLM outputs, in part by offering affordances for scaling, archiving, and visualizing LLM outputs that are lacking in proprietary chat interfaces. We particularly caution against the uncritical use of LLMs in "hard cases," where human uncertainty may perversely increase reliance on AI tools just when those tools may be more likely to fail.

Beyond critique, we chart a path forward by articulating a broader vision for AI-resilient law: a system of incorporating law that would support judicial transparency, improve efficiency without compromising legitimacy, and open new possibilities for LLM-augmented legal reading and writing. Interface design is essential to legal AI governance. By foregrounding the design of human-AI interactions, this work proposes to reorient the legal community toward a more principled and truly generative approach to integrating LLMs into legal practice.

¹ Aileen Nielsen, Visiting Assistant Professor, Harvard Law School, ainielsen@law.harvard.edu. Chelse Swoopes, Ph.D. Candidate, Harvard University, cswoopes@g.harvard.edu. Elena Glassman, Assistant Professor of Computer Science, Harvard University, eglassman@g.harvard.edu. For helpful comments and conversation, the authors thank Oren Bar-Gill, Ken Holstein, Daniel Schwarcz, Alexander Stremitzer, and Cass Sunstein. The authors thank Darya Bershanskaya, Justin Curl, Lisa Rydin and Lauren Wood for excellent research assistance.

Table of Contents

<u>INTRODUCTION</u>	<u>3</u>
<u>THE IMPORTANCE OF AI-RESILIENCY</u>	<u>6</u>
<u>AI-RESILIENT INTERFACES ENCOURAGE ARCHIVING AND AUTOMATION</u>	<u>11</u>
<u>AI-RESILIENT INTERFACES ENHANCE TRANSPARENCY AND EFFICIENCY</u>	<u>18</u>
<u>WHY JUDGES TURN TO AI WHEN THEY SHOULDN'T: A PRIEST-KLEIN HYPOTHESIS FOR LLMS</u>	<u>25</u>
<u>NEW POSSIBILITIES.....</u>	<u>28</u>
GENERATIVE READING	28
GENERATIVE DRAFTING	30
<u>CONCLUSION</u>	<u>32</u>

Introduction

When OpenAI created a conversational interface wrapper around their large language model—the first iteration of ChatGPT—anyone could ask the model questions, or give it tasks for personal or even professional purposes. While the launch of ChatGPT has often been celebrated as an outstanding technical achievement, it was primarily an ingenious feat of human-computer interaction. As Meredith Ringel Morris, Director for Human-AI Interaction at Google DeepMind, has argued, the design of human-computer interaction is not secondary to AI development, but rather central to its trajectory.²

The recent explosion of public interest in large language models (LLMs) has been driven as much by interface design as by raw computational advances. The graphical presentation of LLM outputs: what is provided, what is not provided, what is made easy with a button, and what buttons are missing, are all interface decisions that strongly influence how people use (or misuse) LLMs. Interface design is not peripheral to legal AI governance but rather *essential* to ensuring transparency, reproducibility, and epistemic integrity.

Judges and lawyers have joined the throngs of people and organizations exploring how LLM-backed chat interfaces can be used in professional settings. Consider the recent 11th Circuit decision of *Snell v. United Specialty Insurance Company*. Federal Circuit Judge Kevin Newsom made national headlines³ penning a concurrence in which he said “the unsayable”:

Those ... who believe that “ordinary meaning” is the foundational rule for the evaluation of legal texts should consider — consider — whether and how AI-powered large language models ... might — might — inform the interpretive analysis.⁴

Judge Newsom specifically wrestled with the question of whether an in-ground trampoline was covered by an insurance provision about “landscaping”. The question was a difficult one, and, by his own telling, Judge Newsom contemplated the conundrum for “hours and hours (and hours)”.⁵ Finding guidance from dictionaries to be unhelpful,⁶ he turned to LLM chat interfaces to explore

² This was stated in Morris’s talk at UIST 2023. Meredith Ringel Morris, Director for Human-AI Interaction, Google DeepMind, Talk at the ACM Symposium on User Interface Software and Technology (Oct. 31, 2023), <https://programs.sigchi.org/uist/2023/program/session/127846> (while there is no record of that talk, Morris confirmed this statement in a personal communication to the authors on July 16, 2025).

³ See e.g. Nate Raymond, *U.S. Judge Runs Mini-Experiment with AI to Help Decide Case*, REUTERS (Sept. 6, 2024), <https://www.reuters.com/legal/transactional/us-judge-runs-mini-experiment-with-ai-help-decide-case-2024-09-06/>.

⁴ *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1221 (11th Cir. 2024).

⁵ *Id.* at 1221-22.

⁶ *Id.* at 1221-23. (“As occasionally happens, the dictionaries left a little something to be desired. From their definitions alone, it was tough to discern a single controlling criterion. Must an improvement be *natural* to count as

whether they might "shed some light" on the question.⁷ He expanded this experiment later that year in *U.S. v. Deleon*, a statutory interpretation case in which he studied multiple outputs from multiple LLMs in hopes of identifying a "common core".⁸ Judge Newsom almost certainly conducted interactions with these LLMs through standard chat interfaces.

Often, it seems, a convenient chat interface has coincided with overly casual or uncritical legal use of LLM outputs. A number of lawyers have been sanctioned for false citations (that is, hallucinations) in their submissions to courts, and it seems likely they have used a chat interface in all such cases.⁹ There is also some evidence that judicial use of LLMs accessed by chat interfaces may be problematic. An ethnographic study of judicial use of LLMs in a Chinese court showed that judges make decisions about outcomes and then prompt LLMs to justify the decision.¹⁰ But is a judge even deciding when she doesn't draft her own opinion? Traditionally, much emphasis has been placed on the analytical work done by a judge precisely when she justifies her decision to others in a written form that will be examined and contested by others.¹¹

'landscaping'? Maybe, but that would presumably exclude walkways and accent lights, both of which intuitively seemed (to me, anyway) to qualify. Perhaps "landscaping" work has to be done for *aesthetic* reasons? That, though, would rule out, for instance, a project to regrade a yard, say away from a house's foundation to prevent basement flooding. I once regraded my own yard, and while my efforts did nothing to improve my house's curb appeal, they served an important functional purpose—and for what it's worth, I definitely thought I was engaged in 'landscaping.'")

⁷ Judge Newsom presented both advantages and disadvantages of incorporating LLMs into textual analysis, and he helpfully presented his thoughts in a list. As to the advantages, he cited, "1. LLMs train on ordinary-language inputs...2. LLMs can "understand" context. 3. LLMs are accessible. 4. LLM research is relatively transparent...5. LLMs hold advantages over other empirical interpretive methods." *Snell* at 1226-30. As to the disadvantages, Newsom listed "1. LLMs can 'hallucinate' ...2. LLMs don't capture offline speech...3. Lawyers, judges, and would-be litigants might try to manipulate LLMs...4. Reliance on LLMs will lead us into dystopia." *Snell* at 1230-1232. In a subsequent district court case, *Ross v. U.S.*, in which a dissent included outputs from ChatGPT and cited Judge Newsom's concurrence, a concurring judge pointed out additional concerns not addressed by the *Snell* concurrence: "security, privacy, reliability, and bias, to ensure ethical use." *Ross v. U.S.*, No. 23-CM-1067 (D.C. Feb. 20, 2025) Howard concurring. Due to nuances in the applicable law and history of dealing between the parties, Judge Newsom's chats with LLMs would not have been dispositive in the case no matter how the LLM outputs came out, but Judge Newsom had nonetheless kicked off a new line of judicial conversation. *Snell* at 1222.

⁸ *U.S. v. Deleon*, 116 F.4th 1260 (11th Cir. 2024).

⁹ Sara Merken, *New York Lawyers Sanctioned for Using Fake ChatGPT Cases in Legal Brief*, REUTERS (June 26, 2023), <https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/>. Debra Cassens Weiss, *No. 42 Law Firm by Head Count Sanctioned over Fake Case Citations Generated by AI*, ABA J. (Feb. 10, 2025, 10:30 AM CST), <https://www.abajournal.com/news/article/no-42-law-firm-by-headcount-could-face-sanctions-over-fake-case-citations-generated-by-chatgpt>. Of course we do not have evidence to establish causation as between chat interfaces and misconduct; but future work should investigate that possibility.

¹⁰ John Zhuang Liu & Xueyao Li, *How Do Judges Use Large Language Models? Evidence from Shenzhen*, 16 J. LEGAL ANALYSIS 235 (2024), <https://doi.org/10.1093/jla/laec009>.

¹¹ James Grimmelmann et al., *Generative Misinterpretation*, 63 HARV. J. ON LEGIS. (forthcoming 2025) (Univ. of Wis. Legal Studies Rsch. Paper No. 1878, VANDERBILT L. SCH. LEGAL STUD. RSCH. PAPER No. 5309575) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5309575 (posted June 18, 2025) (describing the act of adjudication as a social fact not just a linguistic product: "The authority of an adjudication depends not just on formal resemblance to a reasoned judicial opinion, but on a set of social practices that imbue *particular* artifacts of legal reasoning with authority."). John Zhuang Liu & Xueyao Li, *How Do Judges Use Large Language Models? Evidence from Shenzhen*, 16 J. LEGAL ANALYSIS 235 (2024), <https://doi.org/10.1093/jla/laec009> ("As AI-generated arguments always support judges' initial decisions, the judges that reviewed AI-generated reasoning can easily

It may be the simplicity and apparent ease of LLM chat interfaces that make them so inviting but possibly also so prone to misuse. Chat interfaces deliver a single, final answer from a single model, hiding helpful information about the highly noisy and stochastic process that produces those outputs. The interface design choices reflected in a chat interface provide users with far less information than they need to fully assess an LLM output. A chat interface does not readily permit users to perform comparative analysis of multiple LLM outputs or otherwise critically evaluate LLM outputs. Responsible analysis of LLM outputs should likely involve, at the least, assessing the robustness, accuracy, and potential biases of LLM outputs. None of this readily comes to hand with a standard chat interface.

Our recent research in human-computer interaction is revealing guidelines and principles to protect us. We aim to design interfaces that inform users without biasing their judgments, encourage critical evaluation of AI-generated content rather than passive acceptance, and support broader adoption without requiring users to write their own code. We can continue to learn more about how to design AI-resilient LLM interfaces in such a way that humans can easily notice and recover from inaccurate or misaligned outputs.¹² We need to prioritize more interface design and testing as AI use grows substantially, including in the legal community. We can do much better as a society to present AI outputs in formats that are less likely to mislead us and more likely to empower us. By the way, such arguments should sound familiar to those who think about behavioral economics and policymaking. Interface decisions for the presentation of AI outputs are important for many of the same reasons as broader questions of choice architecture in legal or policy settings.

In the remainder of this work, we offer four observations. First, we demonstrate a heretofore unidentified brittleness in Judge Newsom's *Snell* concurrence, identifying how an approach that seemed measured and cautious on first inspection nonetheless was problematic and poorly documented even at the time the decision was published. We posit that this lack of robustness would likely have been identified sooner had Judge Newsom accessed the AI-resilient interface we demonstrate here. Second, we look to Judge Newsom's more ambitious, multiple sample analysis in *Deleon* and show how Judge Newsom could have performed the same analysis more quickly and transparently, again a win for AI-resilient interfaces. Contrary to Judge Newsom's contention that LLMs can enhance the transparency of judicial interpretation, it is appropriate tooling, not LLMs as such, that can enhance transparency.¹³ Third, we hypothesize a behavioral danger posed by judicial use of LLMs: the possibility that judges will be most likely to use LLMs in difficult rather than straightforward decisions. While most empiricists and theorists alike suggest that gains from AI use in law will most likely be realized in deploying AI for

overlook errors that align with their decisions. In other words, the hallucination problem can easily seep into judges' decisions through the workflow, even when judges are accountable for reviewing and scrutinizing the reasons generated by AI. This clearly poses a significant risk to judicial decision-making.”)

¹² "An AI-resilient text rendering technique for reading and skimming documents." Ziwei Gu et al., in CHI '24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., Honolulu, Haw., May 11–16, 2024, (2024).

¹³ In fact, Judge Newsom's use of LLMs has been widely criticized on the grounds of transparency because the consumer-facing LLMs Judge Newsom used are notoriously lacking in transparency and reproducibility for a host of reasons.

routine cases, it may be that judges will use LLMs to resolve difficult cases rather than simple, routine ones. This possibility could point to even greater risks that poorly designed LLM-informed adjudication may deform substantive legal outcomes. Finally, we close in highlighting near-future opportunities for AI-resilient tooling to enhance judicial efficiency and efficacy in ways that go beyond the problematic paradigm of generative interpretation. We demonstrate how existing, open-source tooling can solve problems faced by the judiciary but also by other legal actors.

The importance of AI-resiliency

There are many reasons to expect that, just like the rest of us, judges are vulnerable to AI influence. Judge Newsom’s concurrences show that judges are likely to query LLMs about difficult points of textual interpretation through the chat interface. But, LLMs can be contextually inappropriate or even objectively wrong with outputs that are statistically typical.¹⁴ Such mistakes could be difficult to notice. Indeed, the mere act of simply seeing any single answer, even a bad or random answer, can anchor one’s thinking, including in legal reasoning.¹⁵ This could happen with judicial use of LLMs.

Uncritical use of LLMs can deform legal practice, and this is particularly likely to be the case with poor interface design. Could it be that judges query ChatGPT about sentencing guidelines and then become unconsciously anchored to that number? Possibly, but if judges continue using consumer-facing LLM chat interfaces, this would be difficult to detect, due to a lack of documentation and transparency. Could it be that judges lean heavily into logic that appears to be consistent to an uncritical reader but that actually flows from a false premise? Possibly, but this too might be difficult to surface, especially when judges so often resolve disputes with unpublished opinions.¹⁶

Judges (and other legal actors) need guidance and tooling to promote the appropriate and responsible use of AI. So far judges have been left without any concrete proposals for how they can empower themselves to handle AI more effectively and resiliently. Despite an already lengthy lead-up in the history of law and AI,¹⁷ judges still lack systems to practice and promote

¹⁴ The professors Carl T. Bergstrom and Jevin D. West, who wrote a recent book on bullshit, i.e., what it is and why it’s tricky, call large language models *bullshit machines*, i.e., <https://thebullshitmachines.com/>. Bullshit sounds plausible because it is statistically typical.

¹⁵ See e.g. Holger Spamann et al., *Judges in the Lab: No Precedent Effects, No Common/Civil Law Differences*, 13 J. LEGAL ANALYSIS 110 (2021), <https://doi.org/10.1093/jla/laaa008> (showing that the effects of an arbitrary sentencing anchor were at least as substantial as effects of precedent in a lab experiment on judge-implemented and student-implemented sentencing).

¹⁶ Will Yeatman, *Ninth Circuit Review-Reviewed: A Surfeit of Unpublished Opinions?*, YALE J. ON REG.: NOTICE & COMMENT (Sept. 16, 2024), <https://www.yalejreg.com/nc/ninth-circuit-review-reviewed-a-surfeit-of-unpublished-opinions/> (“In 1987, unpublished opinions comprised 62% of decisions issued by circuit courts. By the mid-1990s, their share was about 75%. Today [in 2024], it is about 86%.”).

¹⁷ John G. Roberts, Jr., *2023 Year-End Report on the Federal Judiciary*, U.S. SUPREME COURT (Dec. 31, 2023), <https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf>. See also *Snell v. United Specialty Ins. Co.*, 102 F.4th 1208, 1234 (11th Cir. 2024) (referencing the Robert’s report’s implicit assumption that “AI is here to

AI-resilient law.¹⁸ AI-resilient law describes a scenario of legal practice and adjudication that would integrate AI to improve the efficiency, accuracy, and accessibility of legal functions where achievable, but without permitting unintended¹⁹ deformations of substantive outcomes.²⁰ For example, AI-resilient law might incorporate AI tools that help judges review case backgrounds more efficiently, so long as the tools don't lead judges to miss key information. AI-resilient law might also encourage judges to rely strongly on machine judgment, say of recidivism risk, when that judgment is superior to their own, but would also leave room for judges who estimate risk better than AI to be aware of that fact and incorporate their own information or judgment.²¹

AI-resilient interfaces support users in recognizing when a model makes an objective factual mistake or contextually inappropriate choice. Further, AI-resilient interfaces enable easy recovery from undesirable outcomes.²² For example, when summarizing lengthy legal documents, AI models with simple, default interfaces might omit contextually critical details or misrepresent information by presenting it without the context necessary to understand correctly. An AI-resilient interface could retain all original document content while modifying it visually, e.g., highlighting the most important aspects at the document or sentence level to enable legal professionals to quickly skim and identify key points but also leaving them in full possession of the underlying text.²³ In another scenario, when a judge asks the model to provide citations, an AI-resilient interface could display multiple responses side by side and highlight overlapping or

stay.”). Paul E. Burns & Mindy M. Morton, *Technology-Assisted Review: The Judicial Pioneers*, THE SEDONA CONF. INST. (Mar. 13–14, 2014), https://www.americanbar.org/content/dam/aba/publications/litigation_committees/commercial/materials/technology-assisted-review-the-judicial-pioneers.pdf.

¹⁸ With the phrase “AI-resilient law”, we extend to law ideas developed by Elena L. Glassman et al., *AI-Resilient Interfaces*, ARXIV (May 14, 2024), <https://arxiv.org/abs/2405.08447> (examining how human cognitive processes can benefit from AI without being inadvertently distorted by AI).

¹⁹ What is “intended” can be difficult to assess. For example, numerous studies of criminal recidivism scoring instruments have shown an influence on outcomes different from what was originally the announced policy in adopting the tools; nevertheless given the multiplicity of actors (by design) in criminal adjudication, it is less straightforward to assert that such outcomes are unintended by all the actors in the system.

²⁰ Cf Janet Freilich, *Law as a Lamp Post*, IOWA L. REV. (2024), https://scholarship.law.bu.edu/faculty_scholarship/3955 (suggesting ways in which the selective production of information by law and therefore the unrepresentative nature of legal outputs as training data for AI could result in legal AI systems trained unevenly or even arbitrarily on the resulting legal corpora).

²¹ See Victoria Angelova, et al., *Algorithmic Recommendations and Human Discretion*, NBER WORKING PAPER No. w31747 (Sept. 2023), <https://ssrn.com/abstract=4589709> (last revised Sept. 29, 2024) (finding that around 10% of judges outperformed a recidivism prediction instrument).

²² Elena L. Glassman et al., *AI-Resilient Interfaces*, ARXIV (May 14, 2024), <https://arxiv.org/abs/2405.08447>.

²³ See e.g. Ziwei Gu et al., *An AI-Resilient Text Rendering Technique for Reading and Skimming Documents*, in CHI '24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., Honolulu, Haw., May 11–16, 2024, (2024). Aileen Nielsen, et al., *Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency*, 21 J. EMPIRICAL LEGAL STUD. 979 (2024).; for non-legal work, see e.g. Albert Ziegler et al., *Measuring GitHub Copilot's Impact on Productivity*, COMM. ACM (Feb. 15 2024), <https://cacm.acm.org/research/measuring-github-copilots-impact-on-productivity/>.

differing citations.²⁴ While the reader should not assume correctness even when content is consistent across models or prompts, AI-resilient interfaces can assist judges to reflect on patterns and spot inconsistencies between LLM outputs. By enabling easier comparison, AI-resilient interfaces promote more effective critical evaluation. Some AI-resilient interfaces have already been tested and empirically shown not to deform legal judgment.²⁵ We need more such experiments to assess the impacts of interfaces in legal use. We urgently need to put AI-resilient interfaces into the hands of the judiciary as they pioneer the use of LLMs (and other forms of AI) to enhance judicial efficiency and efficacy.²⁶

To return to the impetus for our case study on AI-resilient interfaces, we welcome Judge Newsom's transparency. Examples of disclosed, transparent use of LLMs are difficult to come by, even though it is surprisingly easy to find hallucinated citations in court filings (a likely symptom of undisclosed LLM use).²⁷ We believe there may be substantial undisclosed use of LLMs both by litigators but also by judges themselves. Such individuals are very likely using the standard chat interface, not only because most judges and lawyers don't have computer coding or interface design backgrounds, but also because they may lack awareness of alternative, AI-resilient tools.²⁸ Initial forays by the judiciary into LLM-informed textual analysis threaten to

²⁴ Cf. Katy Ilonka Gero et al., *Supporting Sensemaking of Large Language Model Outputs at Scale*, in CHI '24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., at 1 (May 11, 2024), <https://doi.org/10.1145/3613904.3642139>.

²⁵ Aileen Nielsen et al., *Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency*, J. EMPIRICAL LEGAL STUD. 979 (2024).

²⁶ While this work focuses on examples of AI-resilient interfaces for LLMs, we emphasize that AI outputs from other models raise the same concerns about appropriate and effective presentation of information.

²⁷ LLM hallucinations in the context of this discussion refer to the phenomenon of LLMs generating references to legal citations that do not actually exist. We are aware of no systematic empirical work inferring the rate of undisclosed use of LLMs in legal work, but the rate of undisclosed LLM use is likely non-zero. For some indicia that attorneys are using LLMs to prepare court filings without explicit acknowledgement is the fact that a recent investigation turned up multiple discussions of hallucinated citations. See Eugene Volokh, *AI in Court: 11 Court Opinions in the Last 30 Days Mention AI-Hallucinated Material, and ... That's Likely Just the Tip of the Iceberg*, REASON (Mar. 25, 2025, 2:52 PM ET), <https://reason.com/volokh/2025/03/25/11-court-opinions-in-the-last-30-days-mention-ai-hallucinated-material/>. On the other hand, when we looked at a selection of recent filings in front of judges who require explicit disclosure of LLM use, we could find no such disclosures. This evidence is of course not dispositive, but taken together, these findings are consistent with higher rates of use than of disclosure. It is also worth noting that hallucinations in legal work products need not be definitively the result of LLM hallucinations; for example a recent experimental study carried out on law students turned up examples of hallucinations even in the work product of students who had been instructed not to use LLMs. Daniel Schwarcz et al., *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice*, MINN. LEGAL STUD. RSCH. PAPER No. 25-16 (Mar. 2, 2025), <https://ssrn.com/abstract=5162111>.

²⁸ Several judges across the United States now require human review of all LLM-generated content, and yet a search of recent filings turned up no evidence of parties acknowledging use of LLMs, even in matters before those judges. Law360 Pulse, *AI Tracker*, <https://www.law360.com/pulse/ai-tracker> (last visited Mar. 19, 2025). We looked at a sample of 20 recent filings in front of some California judges with such requirements, but we could not identify any explicit admission of LLM use. Disclosure requirements need not be interpreted as condemnation; some judges have even taken up Judge Newsom's example or endorsed his reasoning. *Ross v. U.S.*, No. 23-CM-1067 (D.C. Feb. 20, 2025) (dissenting opinion). Nonetheless, it seems that some attorneys and judges likely avoid disclosure, perhaps fearing that they will be assumed to be using LLMs carelessly, as appears to have been the case in numerous instances of attorneys submitting filings with hallucinated case references, and leading to professional embarrassment and judicial condemnation. Thomas Claburn, *LLM aka Large Legal Mess: Judge Wants Lawyer*

normalize brittle and opaque uses of LLMs; AI-resilient interfaces can mitigate these present dangers. Good interfaces combined with simple tooling would substantially improve upon what is already happening and, we believe, very likely to continue.²⁹ Here we focus our discussion on how Judge Newsom or others could use LLMs more efficiently, transparently, and effectively if provided with more AI-resilient interfaces.

AI-resilient tooling could offer a substantial improvement upon current judicial reliance on the proprietary chat-interfaces of consumer-facing LLMs.³⁰ We make an initial demonstration of the use of AI-resilient tooling to promote AI-resilient law. We take up the same examples Judge Newsom addressed, and we show how a more AI-resilient interface could have rendered this work more robust, efficient, and transparent. We will show here an open-source AI-resilient interface that enables users to generate, record, and analyze many variations of the outputs of LLMs in an archivable and automated manner.

Of course, some readers may fairly question whether we really want judges using LLMs to inform their textual interpretation. We share such concerns. There are good reasons to be cautious about incorporating LLMs into adjudication, and especially into the process of textual interpretation. A host of critiques of the project of LLM-informed textual interpretation have already surfaced in recent scholarship. Brandon Waldon and coauthors, as well as Thomas Lee and Jesse Egbert have raised questions about the premise that LLMs provide legitimate and accurate information about how ordinary people use and understand language.³¹ Jonathan Choi has examined the instability of LLM-based legal analyses, while James Grimmelman and coauthors have similarly provided demonstrations that question whether LLMs plausibly meet even minimal standards of interpretive consistency and reliability.³² Justin Curl and coauthors

Fined \$15K for Using AI Slop in Filing, THE REGISTER (Feb. 25, 2025), https://www.theregister.com/2025/02/25/fine_sought_ai_filing_mistakes/. Sara Merken, *Lawyer Who Used Flawed AI Case Citations Says Sanctions Unwarranted in Whistleblower Case*, REUTERS (Aug. 27, 2024), <https://www.reuters.com/legal/transactional/lawyer-who-used-flawed-ai-case-citations-says-sanctions-unwarranted-2024-08-27/>. Mass. Law. Wkly. Staff, *Lawyer Sanctioned over False AI-Generated Content*, MASS. LAW. WKLY. (Feb. 13, 2024), <https://masslawyersweekly.com/2024/02/13/lawyers-sanctioned-over-false-ai-generated-content/>.

²⁹ See Daniel Wiff-Townsend et al., *Generated Legal Texts*, SSRN (May 6, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5243382 (working paper) (describing the rise of many forms of using LLMs for legal work by the executive and judicial branches of federal and state government).

³⁰ Judge Newsom as well as all the judges and clerks who used ChatGPT in *Ross* appear to have used the default consumer-facing chat interface. We do not see evidence that any of the judicial users made use of computer code and the direct APIs for these products, nor do we see evidence that any of the judicial users made use of third-party tooling such as what we present here.

³¹ Thomas R. Lee & Jesse Egbert, *Artificial Meaning?*, BYU L. RSCH. PAPER (Oct. 1, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4973483 (posted Nov. 8, 2024) (last revised Apr. 18, 2025). And, Brandon Waldon et al., *Large Language Models for Legal Interpretation? Don't Take Their Word for It*, GEORGETOWN L. J. 114 (Feb. 3, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5123124 (posted Apr. 2, 2025) (last revised Feb. 19, 2025).

³² Jonathan H. Choi, *Off-the-Shelf Large Language Models Are Unreliable Judges* (Feb. 28, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5188865. And, James Grimmelman et al., *Generative Misinterpretation*, 63 HARV. J. ON LEGIS. (forthcoming 2025) (UNIV. OF WIS. LEGAL STUDIES RSCH. PAPER No. 1878, VANDERBILT L. SCH. LEGAL STUD. RSCH. PAPER No. 5309575) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5309575 (posted June 18, 2025).

have pointed out the potential for deliberate manipulation of LLM content to influence consequential litigation.³³ We share these concerns, but our critique is orthogonal.

Our interface-based critique applies regardless of the underlying quality of the LLM model or even the underlying form of AI invoked for legal work. AI-resilient interfaces are valuable regardless of model quality. No AI model is so good that it eliminates the need for thoughtful, well-designed, human-centered computer-interaction. Even as LLMs improve, interfaces that empower users to detect, interpret, correct, and analyze AI outputs will be necessary. But such interfaces will likely not be the default interface presented by a proprietary firm. Firms will likely often have incentives at odds with the critical analysis of their model's outputs, and so it is prudent to look elsewhere for best practices for AI interfaces when possible.

Most people do not have a background in researching interfaces, and they may not appreciate that interface itself provides a host of opportunities to improve the use of LLMs. For example, in their best practice recommendations, Waldon et al. suggested that judges might interrogate LLMs (in the chat interface) to provide the strongest arguments for both sides, a process they described as dialectical legal AI.³⁴ In providing such a practice recommendation, Waldon et al. accept the chat interface as a given. But making space to imagine something beyond what LLM firms themselves give us in the chat interface will open up many more possibilities. We rather fear that dialectical uses of AI may themselves lack AI resiliency and open users up to a host of worrying effects and misbehaviors.³⁵ Dialectical legal AI practiced in a chat interface still raises concerns about robustness, scalability, transparency, and anchoring.

Our discussion of interface design is not just a critique of existing uses but a gesture to a promising future. Under the right conditions, LLMs could support judicial work. When properly designed and deployed, some AI tools enhance objective and perceived legal efficiency³⁶ and efficacy.³⁷ And, while not yet confirmed with legal work evaluations, recent AI-powered tools

³³ Justin Curl et al., *Judges Shouldn't Rely on AI for the Ordinary Meaning of Text*, *LAWFARE* (May 22, 2025), <https://www.lawfaremedia.org/article/judges-shouldn-t-rely-on-ai-for-the-ordinary-meaning-of-text>.

³⁴ Waldon et al. at 53.

³⁵ For intriguing suggestions of cognitive differences in people using LLMs to write versus others, see Nataliya Kosmyna et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*, ARXIV (June 10, 2025), <https://arxiv.org/abs/2506.08872>.

³⁶ See e.g. Aileen Nielsen, et al., *Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency*, 21 J. EMPIRICAL LEGAL STUD. 979 (2024).; Colleen V. Chien & Miriam Kim, *Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap*, 57 LOY. L.A. L. REV. 903 (2025) (describing self-reported measures that attorneys using LLM in the field believed the tool enhanced their efficiency). But see e.g. Matthias Stadler et al., *Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry*, 160 COMPUTS. HUM. BEHAV. 108386 (2024), <https://doi.org/10.1016/j.chb.2024.108386> (documenting reduced cognitive load with the use of LLMs but also noting negative effects coinciding with reduced cognitive load).

³⁷ See Daniel Schwarcz et al., *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice*, MINN. LEGAL STUD. RSCH. PAPER No. 25-16, U. MICH. PUB. L. RSCH. PAPER No. 24-058 (May 27, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111 (showing performance and efficiency gains for students using LLMs with RAGS as compared to students not using LLMs). More generally, see

with AI-resilient interfaces have been found to enhance both objective and subjective generic reading and writing experiences. For example, CorpusStudio helps writers explore relevant corpora while writing, without risking plagiarism from generated content.³⁸ Grammar-Preserving Text Saliency Modulation (GP-TSM) reifies internal sentence structure, with its grammatical core and contextualizing clauses, to help readers answer comprehension questions more accurately in less time without hiding any text.³⁹ A reasonable near-future goal for judicial use of LLMs would be to build on these early gains without compromising the procedural or substantive integrity of the legal process, undermining intellectual and professional development, or diminishing the distinctive, AI-exceeding capabilities that judges bring to their work.⁴⁰ Gaining from the strengths of LLMs without losing the advantages of human involvement will require use of LLM workflows and interfaces that promote desired behaviors and outcomes while mitigating or preventing undesirable or unacceptable ones.⁴¹ AI-resilient interface design will be a critical tool in bringing about AI-resilient law.

AI-resilient interfaces encourage archiving and automation

We here present a replication effort of Judge Newsom’s *Snell* concurrence. Despite investigating this LLM use only a year after publication, it’s been difficult to determine exactly what Judge Newsom did in producing his *Snell* concurrence, and when he did it. Part of the lack of information is explained by Judge Newsom’s use of the chat interface. That interface does not lend itself to archiving, reproducing, or scaling analyses conducted with LLMs. Of course, Judge Newsom made efforts in furtherance of transparency and reproducibility: he provided the text of his chat-based interactions in an appendix. But Judge Newsom’s efforts gave us little information about when he ran his prompts through the LLM, which turns out to be important. Further, Judge

Michelle Vaccaro et al., *When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis*, 8 NAT. HUM. BEHAV. 2293 (2024).

³⁸ Hai Dang et al., *CorpusStudio: Surfacing Emergent Patterns in a Corpus of Prior Work While Writing*, in CHI ’25: Proceedings of the 2025 CHI Conf. on Hum. Factors in Computing Sys. (2025).

³⁹ Ziwei Gu et al., *An AI-Resilient Text Rendering Technique for Reading and Skimming Documents*, in CHI ’24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., Honolulu, Haw., May 11–16, 2024, (2024).

⁴⁰ Cf Krzysztof Z. Gajos and Lena Mamykina, *Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning*, in Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI ’22) Association for Computing Machinery, New York, N.Y., USA, 794–806 (2022), <https://doi.org/10.1145/3490099.3511138> (showing how some workflows with AI assistance can impair the kind of learning that happens when we receive feedback during or after engaging in a task). Victoria Angelova et al., *Algorithmic Recommendations and Human Discretion*, NBER WORKING PAPER No. w31747 (Sept. 2023), <https://ssrn.com/abstract=4589709>.

⁴¹ Cf John Zhuang Liu & Xueyao Li, *How Do Judges Use Large Language Models? Evidence from Shenzhen*, 16 J. LEGAL ANALYSIS 235 (2024), <https://doi.org/10.1093/jla/laae009> (demonstrating how judges likely used LLMs to write opinions after reaching a conclusion rather than using the writing process itself to explore and justify those conclusions).

Newsom’s appendix would have entailed manual copy-pasting, a process that is unnecessarily laborious and error-prone compared to automated methods.

The lack of archival and automation affordances in the chat interface Judge Newsom used isn’t just a theoretical issue to nitpick. Judge Newsom’s concurrence was *technologically out of date even on the day it was published*, because one of the models he used, Google’s Bard, had been replaced by Gemini three months before *Snell* was issued.⁴² This fact matters because the *Snell* prompts were not temporally robust to Google’s model change from Bard to Gemini, as we demonstrate in Figure 3 and Table 1. In other words, not only the specific outputs but even the valence of those outputs did not remain stable as Google’s models were upgraded over time.

Ironically, Judge Newsom’s concurrence itself exemplifies the transparency and reproducibility problems posed by relying on fast-evolving, consumer-facing LLMs, even as his concurrence expressed optimism about LLMs on these matters. We find it striking that within his *Snell* concurrence, Judge Newsom acknowledged that there was already a later AI model available but did not update his opinion to reflect the outputs of that later model.⁴³ With hindsight, the decision to knowingly stand by LLM outputs from an outdated model, and not to note whether they had been verified in the later model, is difficult to understand. It is not clear whether this reflects a mistaken assumption about the model’s temporal robustness, or perhaps a belief that the newer model having a different response would not have changed his final opinion (despite consulting the model when he was stuck without a solidified opinion), or maybe something else entirely.

It may be that Judge Newsom did not repeat the analysis with the newer model because the process of prompting the LLM via a chat interface, and then documenting that new text, and then updating the appendix, would have been too laborious. Perhaps this was a mistake induced in part due to the points of friction he faced as a user of a chat interface. While working on *Snell*, it’s possible that collecting more than one response when he originally sampled LLM outputs might have alerted him to the lack of robustness in his analysis; Judge Newsom may have taken more samples had it been easier to do so. With better, less frictive tooling, he might have revisited and expanded his analysis. That would have been highly desirable, as it could have spared him relying on an example that turns out to be quite brittle, at least with a year’s hindsight.

Here we offer an alternative: [ChainForge.ai](https://chainforge.ai), an open-source tool with an interface designed to help users be more resilient to specific LLM models, prompts, and stochasticity that readily enables archival and scaling functions.⁴⁴ ChainForge is an academic research project, and it is

⁴² Google announced the replacement of Bard with Gemini in February 2024, several months before the publication of the *Snell* decision in May 2024. See Jason Aten, *Google Just Killed Bard and Replaced It with Gemini. It’s More Than Just a Name Change*, INC. (Feb. 8, 2024), <https://www.inc.com/jason-aten/google-just-killed-bard-replaced-it-with-gemini-its-more-than-just-a-name-change.html>.

⁴³ *Snell* at _ (“For good measure, I posed the same questions to Google’s Bard (since replaced by Gemini).”)

⁴⁴ Ian Arawjo et al., *ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing*, in CHI ’24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Art. No. 304, at 1 (May 11, 2024), <https://doi.org/10.1145/3613904.3642016>.

now widely used by people at some of the world’s largest technology firms (Microsoft, Amazon) and by a leading AI language model organization (HuggingFace). It’s also been used for recent reporting to the European Union on the political representativeness of chatbot outputs.⁴⁵

ChainForge would have empowered Judge Newsom to archive his results either by exporting the generated outputs to an Excel spreadsheet or by saving the entire workflow in a timestamped *.cforge* file. Such a file can be reloaded into ChainForge, allowing anyone to view both the exact LLM outputs and the queries that generated them. This also would have spared Judge Newsom the separate copy-pasting task of an appendix preparation. What’s more, ChainForge would have enabled Judge Newsom to readily scale up his analysis, running ten, fifty or hundreds of queries with the click of a button rather than by opening ten, fifty or hundreds of separate ChatGPT tabs. Of course, it wouldn’t be realistic for Judge Newsom to manually review hundreds of outputs individually even with ChainForge, but we also have tools that assist humans in that manual review at scale, i.e., noticing the patterns (and anomalies) in those hundreds of outputs. We’ll discuss that *infra*. For now, we compare how a chat interface works for the task in *Snell* and how this would be easier in ChainForge.

We begin with a visual comparison of OpenAI’s interface versus the alternative of ChainForge. Figure 1 shows a screenshot of a chat interface similar to what Judge Newsom likely used in *U.S. vs. Snell*, when he explored the “ordinary meaning” of landscaping to evaluate whether models could assist in informing interpretive analysis. Figure 2 shows the ChainForge interface we used to replicate *Snell*’s LLM analysis. Most readers will already be familiar with OpenAI’s interface affordances, which include copy-pasting as well as semi-permanent links to past chats.⁴⁶ Some of ChainForge’s affordances include: loading/saving timestamped files of model inputs and outputs from past LLM prompting, and prompting multiple LLMs (Figure 2, B.2) for multiple independent outputs (Figure 2, B.1) with multiple prompts or variations on a single prompt (Figure 2, A). All of ChainForge’s capacities can be invoked with a single click of a button. These features are all lacking in the OpenAI chat interface.

⁴⁵ Duncan Allen et al., *An AI-Powered Audit: Do Chatbots Reproduce Political Pluralism?*, DEMOCRACY REPORTING INT’L (Dec. 27, 2024), <https://democracy-reporting.org/en/office/EU/publications/an-ai-powered-audit-do-chatbots-reproduce-political-pluralism>.

⁴⁶ *ChatGPT Shared Links FAQ*, OpenAI Help Ctr., <https://help.openai.com/en/articles/7925741-chatgpt-shared-links-faq> (last updated June 27, 2025).

What is the ordinary meaning of "landscaping"?

The ordinary meaning of "landscaping" refers to the process of **modifying, designing, or maintaining the visible features of an area of land**, especially for aesthetic or functional purposes.

This typically includes:

- **Planting** trees, shrubs, flowers, or grass
- **Installing** paths, patios, water features, or fences
- **Shaping** the terrain (e.g., grading, adding soil or rocks)
- **Maintaining** lawns, gardens, and outdoor spaces (e.g., mowing, trimming, mulching)

In everyday use, *landscaping* can refer both to initial design and installation (*landscape construction*) and to ongoing upkeep (*landscape maintenance*). It's often associated with making a yard or outdoor space more attractive or usable.



Figure 1: Standard consumer-facing chat interface (ChatGPT) used to replicate Judge Newsom’s Snell query (“ordinary meaning of landscaping”) in July 2025. This interface delivers a single output with minimal metadata or affordances for comparison, highlighting the limitations in transparency, reproducibility, and multi-output exploration that characterize conventional chat interfaces.

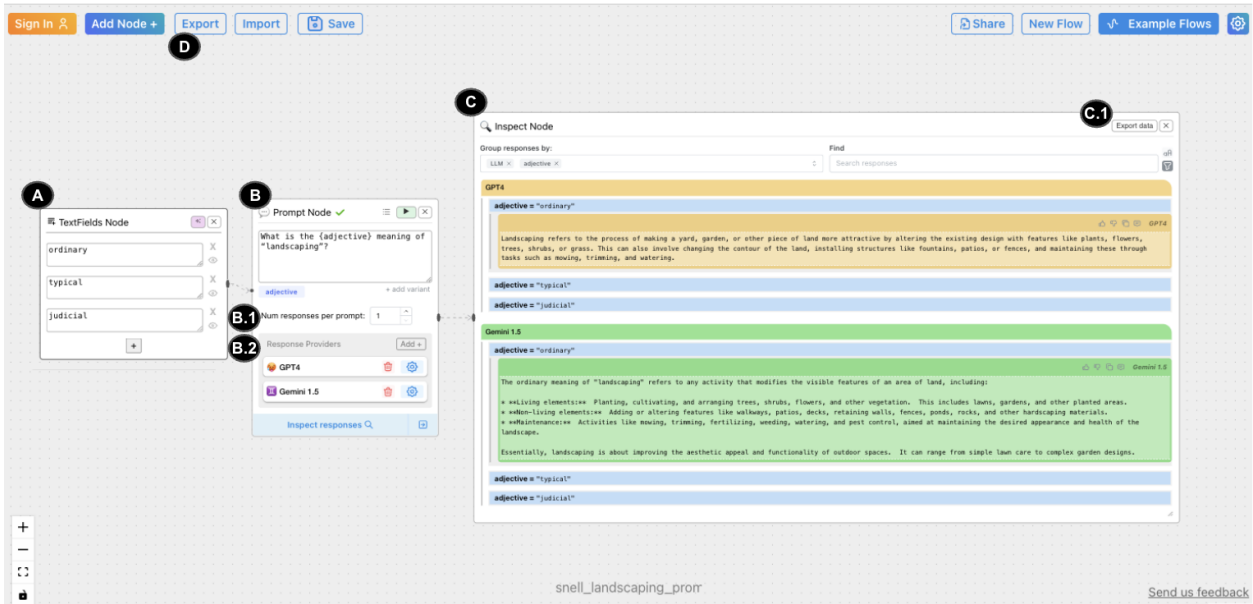


Figure 2: AI-resilient interface (ChainForge) replicating the Snell query across multiple prompts and models. Compared to the chat interface in Figure 1, ChainForge enables scaled, parallel queries to different LLMs (B.2) and allows users to vary prompts (A) and number of outputs (B.1), while preserving full archiving and export capabilities (C.1, D). This contrast illustrates how interface design can support reproducibility, transparency, and critical engagement.

Much of what we emphasize about ChainForge is neither unique nor terribly difficult to implement with some computer code. Judge Newsom could have written a program to call the

relevant application programming interfaces (APIs)⁴⁷ multiple times for multiple models and save the responses in a csv file. He also could have come up with multiple prompt variations and tried them all. Realistically, however, it is a highly unlikely scenario for judicial use of LLMs that judges would write code for an API. What’s more, ideally, AI-resilient interfaces would provide even more functionality beyond archiving and scaling. ChainForge does offer more, as we show in the next section. But for now, we continue with the replication.

We made use of ChainForge to easily reproduce Judge Newsom’s work in *Snell*. Figure 3 shows the outputs, displayed in ChainForge, from generating one response from each of OpenAI’s GPT-4 and Google’s Gemini 1.5. In his concurrence, Judge Newsom characterized Google’s Bard as offering a neutral take on *Snell*.⁴⁸ But less than a year later, a single query to Bard’s successor, Gemini,⁴⁹ produced outputs with a clear stance against Judge Newsom’s LLM-informed *Snell* position. In our March 2025 tests,⁵⁰ Gemini outputs disagree with GPT-4 outputs: Gemini outputs concluded that a trampoline is “generally not considered ‘landscaping.’”

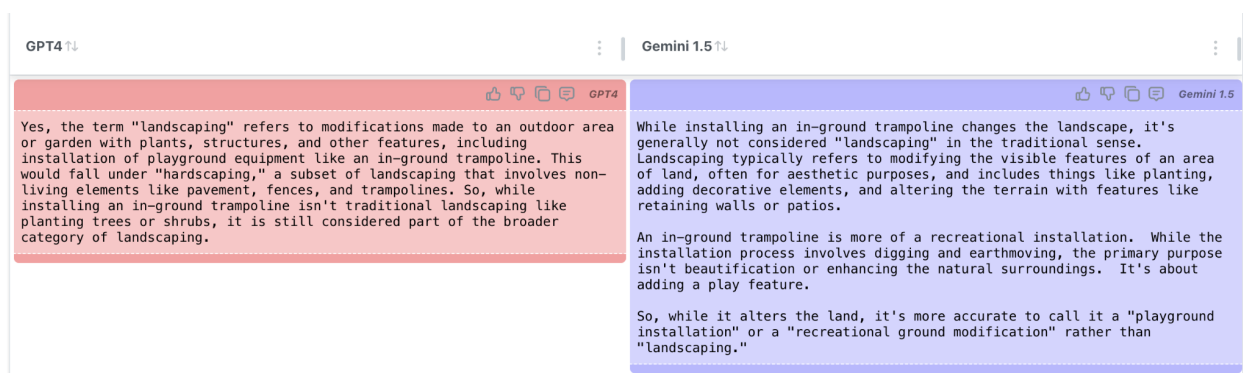


Figure 3: Divergent outputs from GPT-4 and Gemini 1.5 in response to the Snell query, “Is installing an in-ground trampoline landscaping?” These single responses generated in March 2025 illustrate

⁴⁷ APIs are a common way that software or AI as a service is made available. Potential users of the system are given a set of function specifications that they may use to invoke the proprietary code bases and data sets they are unable to access directly.

⁴⁸ “For good measure, I posed the same questions to Google’s Bard (since replaced by Gemini). The precise details aren’t particularly important, but the upshot is that both models’ answers indicated that the trampoline-related work Snell had performed—the excavation of the pit, the construction of the retaining wall, the installation of the mat, and the addition of the decorative wooden cap—just might be landscaping.” Judge Newsom continued in a footnote, “Generally, Bard’s response to my general question—“What is the ordinary meaning of ‘landscaping?’”—was pretty similar to ChatGPT’s, though notably longer. When asked the more specific question—“Is installing an in-ground trampoline ‘landscaping?’”—Bard was more equivocal than ChatGPT had been. I’ve included my questions and the models’ responses in an appendix for readers’ reference.” In Judge Newsom’s May 2024 experiment, Bard’s output was, in part, “Ultimately, whether or not installing an in-ground trampoline is considered landscaping is a matter of opinion. There is no right or wrong answer.” The answer we received in March 2025 from Google’s successor product was less equivocal.

⁴⁹ See e.g. Sissie Hsiao, *Bard Becomes Gemini: Try Ultra 1.0 and a New Mobile App Today*, GOOGLE BLOG (Feb. 8, 2024), <https://blog.google/products/gemini/bard-becomes-gemini/>.

⁵⁰ One might ask whether we should rerun our own analyses closer to the time we share them publicly. We present the earliest analyses we have recorded so that we minimize the time between Judge Newsom’s outputs and ours.

how model choice alone can yield opposing interpretations, raising concerns about the brittleness and volatility of LLM-informed jurisprudence.

Possibly, it was our single output that was misleading. Possibly, Judge Newsom’s earlier Google model results would nonetheless prove to better characterize the LLM response generally. We next ran 10 simultaneous *Snell* queries through each of two models, as illustrated in Figure 4.⁵¹ Results are shown in Table 1.⁵² ChatGPT consistently framed an in-ground trampoline as landscaping (consistent with Judge Newsom’s earlier results), but the Gemini model produced an output that diverged from that view. The discrepancy shown in Figure 3 only deepens when sampled across multiple outputs, underscoring how brittle Judge Newsom’s conclusion was to both model choice and to intra-model stochasticity. Perhaps the single output Judge Newsom collected last year in the preparation of *Snell* was atypical. Perhaps more samples would have borne out Judge Newsom’s conclusion at that time. We cannot know now, but we raise the question (distinct from but adding to existing critiques) of what it means for a judicial opinion to cite an LLM output that turns out to be atypical for its time or brittle to the passage of even a short amount of time.

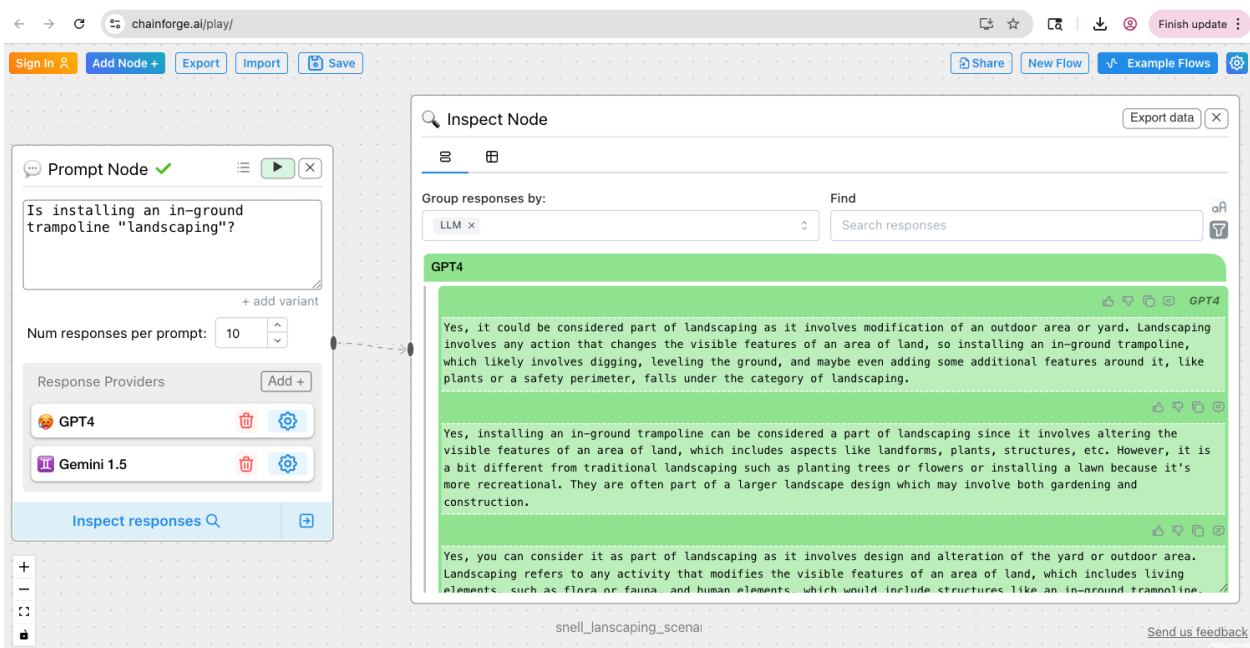


Figure 4: Scaled replication of Judge Newsom’s *Snell* query using N = 10 outputs from GPT-4 and Gemini 1.5. This expanded sampling, made easy with ChainForge, reveals consistent divergence in model interpretations, underscoring the fragility of relying on single or limited outputs for legal decision-making.

⁵¹ We chose this sample size because it was consistent with Judge Newsom’s later choices in *Deleon*. The question of what an appropriate sample size would be and for what task is a separate but important question distinct from the user interface issues we highlight here.

⁵² Results can be seen in the tool used for the research here: <https://chainforge.ai/play/?f=3kn0fip6ynswg>. Outputs and our labels are also here in spreadsheet format: <https://tinyurl.com/268dmuwc>. As we show *infra* the tools we use here can easily scale to larger samples.

A year later, we see not only that our one-off sample differed from what Judge Newsom reported, but further that a larger sample of $N = 10$ only served to emphasize the difference between the models, as shown in Table 1. While Judge Newsom had characterized Google’s model as neutral, we did not find any neutral responses in our sample of 10. Categorizing the responses on a spectrum, we found that Gemini did not typically output language as strong in valence as that of GPT-4 but that Gemini’s outputs did consistently (at the time we recorded outputs, with the settings we used) include language that rejected in-ground trampoline as landscaping. The replication failure we observed in Figure 3 is not a one-off result.

<i>Is installing an in-ground trampoline "landscaping"?</i> ⁵³ (Snell)	OpenAI’s GPT-4	Google’s Gemini 1.5
Yes	8	0
Leaning yes	2 modal response (N = 2) “could be considered”	0
Neutral	0	0
Leaning no	0	10 modal response (N = 9) “generally not considered ‘landscaping’”
No	0	0

Table 1: Distribution of ten independent responses to the Snell query from GPT-4 and Gemini 1.5. While GPT-4 outputs predominantly affirm the classification of an in-ground trampoline as landscaping, Gemini outputs consistently disagree.⁵⁴

We don’t blame Judge Newsom for not investigating multiple outputs in his first public assay into LLM-informed textual analysis. Circuit Court judges and their clerks are overworked and under-resourced.⁵⁵ It’s a laborious task to compare multiple outputs from several LLMs,

⁵³ It is true that this is just one of two questions Judge Newsom posed to the LLM products. The other question—seeking to elicit the ordinary meaning of the terms—has already been critiqued by Waldon et al., a critique we wholeheartedly agree with. More to the point, in the next section *infra* we show how AI-resilient tooling could enhance the efficacy of analyzing outputs from this style of question.

⁵⁴ We obtained these results both using the default settings presented in the ChainForge interface but also using a temperature setting of .7 for both models (the default at the time writing in Chainforge is a temperature of 1 for GPT 4 and .5 for Gemini 1.5, as shown in this screenshot: <https://tinyurl.com/3z6neefv>). Again we could rerun this analysis easily at any time, but we present here the results temporally closest to the time Judge Newsom ran his analysis.

⁵⁵ Not only are federal circuit court judges busy, but their caseloads appear to be rising, at least as measured by raw case counts. See Transactional Records Access Clearinghouse (TRAC), *As Workloads Rise in Federal Courts, Judge*

particularly when using the consumer-facing chat interface. Indeed, such cumbersome attempts at large-N sampling would be inefficient and impossible at the time Judge Newsom ran his queries. At that time, the chat interface did not offer the possibility to have temporary, private chat interactions, an affordance that was later introduced.⁵⁶ Given Judge Newsom’s awareness of multi-output sampling, we imagine that Judge Newsom was, at least in part, disinclined to take and compare multiple samples in *Snell* due to a lack of ready tooling.⁵⁷ Likely, Judge Newsom did the best he could with the resources he had, but those resources were sorely lacking.

AI-resilient interfaces enhance transparency and efficiency

Now, we progress to scaling up the *human analysis* of LLM outputs, to show how an AI-resilient interface could have been deployed in *Deleon* to render Judge Newsom’s analysis more efficient and transparent. Judge Newsom took the initiative to prompt three different models ten times and tackle head-on the question of variation in LLM outputs. Judge Newsom proposed to address such variation against the following standard:

If the model instead returned essentially consistent answers—responses that coalesced around a *common core*—then one could probably say, with a higher degree of confidence, that the model was getting at the ‘correct’ response.⁵⁸

Counts Remain Flat, TRAC REPORTS (Oct. 14, 2014), <https://tracreports.org/tracreports/judge/364/>. Indeed a concurrence in *Ross v. U.S.*, No. 23-CM-1067 (D.C. Feb. 20, 2025), emphasized just this point that LLMs could conceivably in the future assist the D.C. District. Howard concurring. (“It has been well covered in the news for several years that we face historic vacancies. The D.C. Superior Court has in the recent past, and may soon again, faced double digit vacancies. This court has been without a full complement of judges for more than a decade. Our current two-judge vacancy represents nearly one-third of our court. Compounding this issue is an ever dwindling number of senior judges—former active judges who commit to continued public service—seasoned veterans who are valuable resources in accomplishing the work of this court. And, structurally, this court faces the challenge of having the same nine judgeships over the court’s approximate fifty-five-year existence, while the courts and agencies we review have grown to numbers of judges and administrative adjudicators regularly exceeding 100 each.”)

⁵⁶ ChatGPT, *ChatGPT Remembers Across Sessions, Yes It Does*, OPENAI COMMUNITY FORUM (Aug. 2024), <https://community.openai.com/t/chatgpt-remembers-across-sessions-yes-it-does/907238>. It is possible to use the API to generate independent samples not associated with an underlying memory or model of the user by the LLM, but this requires writing and executing computer code. See ChatGPT, *How do you maintain historical context in repeat API calls?*, OPENAI COMMUNITY FORUM (Jan. 2023), <https://community.openai.com/t/how-do-you-maintain-historical-context-in-repeat-api-calls/34395/2>. As of April 2025 OpenAI has also introduced into the user interface the possibility to opt in to a “temporary” chat, such that the contents would not be remembered or integrated into the model’s memory of or understanding of a particular chat user. Maxwell Zeff, *OpenAI Updates ChatGPT to Reference Your Past Chats*, TECHCRUNCH (Apr. 10, 2025, 10:00 AM PDT), <https://techcrunch.com/2025/04/10/openai-updates-chatgpt-to-reference-your-other-chats/>.

⁵⁷ We understand there may have been other, strategic reasons not to take multiple samples. Perhaps Judge Newsom wished to provide a particularly simple and easy-to-digest application for the benefit of judges and attorneys alike. However, as we have already shown, this likely was an oversimplification.

⁵⁸ *Deleon* at 1274. Emphasis supplied.

While Newsom talks in this excerpt only about intra-model variation, his practice revealed both intra- and inter-model variation.

To be sure, the models' responses exhibited some minor variations in structure and phrasing. ChatGPT's answers, for example, tended to fluctuate in length by a sentence or two. For its part, Claude altered the number of examples it provided from one response to the next. But for reasons I'll explain in the next part, these subtle, marginal divergences were probably (and should have been) expected. Far more importantly, I think, the responses did coalesce, substantively, around a common core — there was an objectively verifiable throughline. For our purposes, what matters is that the LLMs consistently defined the phrase "physically restrained" to require the application of tangible force, either through direct bodily contact or some other device or instrument.⁵⁹

We see here that Judge Newsom thought both about how models differed from one another and how consistent they were with themselves. And then, there's another kind of variation Judge Newsom did not analyze, that of variation of responses resulting from changes to prompts. As we illustrate in Figure 2, ChainForge would have permitted Judge Newsom to examine this as well, comparing variation within models, across models, and across prompts. Without appropriate tooling, such a task grows exponentially. With appropriate tooling, this scale of work can be automated.

Admittedly, there are some legitimate concerns about what Judge Newsom did and what he sought from the LLM outputs. It is incorrect to think that the consistency of a single model's responses or the consistency of multiple models' responses must necessarily equate with linguistic consensus or empirical correctness. But this logical fallacy is nonetheless an inviting mistake that many people make. Empirical evidence has shown that this view is shared with other model users.⁶⁰

We flag the concern that model consistency is not the same as legal legitimacy. But we proceed to show how to make Judge Newsom's analysis more efficient and transparent, nonetheless. We do this for two key reasons: First, in the spirit of seeing how Judge Newsom could have done what he chose to do in a more efficient and transparent way with the use of better interfaces.⁶¹

⁵⁹ *Deleon* at 1274-75.

⁶⁰ Chelse Swoopes et al., *The Impact of Revealing Large Language Model Stochasticity on Trust, Reliability, and Anthropomorphization*, TREW Workshop at CHI '24: CHI Conf. on Hum. Factors in Computing Sys., (May 11, 2024), https://chi-trew.github.io/papers/2024/CHI_TREW_2024_Paper_19.pdf.

⁶¹ As with *Snell*, Judge Newsom's use of the LLM was not ultimately dispositive for the case, and in this case again we think this was a good thing because we see an error baked into this reasoning. The legal question, as teed up by the majority opinion, was whether the victim was "physically restrained" under the meaning of the Guidelines when Defendant Joseph Deleon walked into a store, pointed a gun at the cashier while demanding money from the register, received the money, and then left, all within about one minute. *U.S. v. Deleon*, 116 F.4th 1260, 1261 (11th Cir. 2024). As Judge Newsom agreed, the case was clearly controlled by existing precedent. Judge Newsom wrote to call for a revisiting of that precedent, and so his use of LLMs was used for advocacy and in the hopes of convincing his peers on the 11th Circuit.

Second, because psychological theories of human concept formation suggest that contrasts—in this case, within and across models’ responses—can reveal aspects of the problem previously not considered and therefore aid learning, we can imagine responsible uses of multiple LLM outputs in which a legal actor could potentially learn something from contrastive observation.⁶² In other words, doing what Judge Newsom did can plausibly be helpful for human refinement of a concept, under the right circumstances.

We look to Judge Newsom’s work on his own terms, examining how Judge Newsom assessed consistency. Following up on his idea of looking for a “common core,” Judge Newsom examined 10 outputs each from 3 different models. The effort involved in conducting this experiment alone, before any subsequent textual analysis required to compare the responses, would have been substantial: opening 30 different conversational tabs (10 on each website), pasting in a common prompt, and reading and attempting to remember the outputs, or laboriously copying and pasting the outputs into a common document to look at them together, side by side, for less memory-taxing comparison. By contrast, we reproduced this analysis with a few simple steps,⁶³ as depicted in Figure 5.

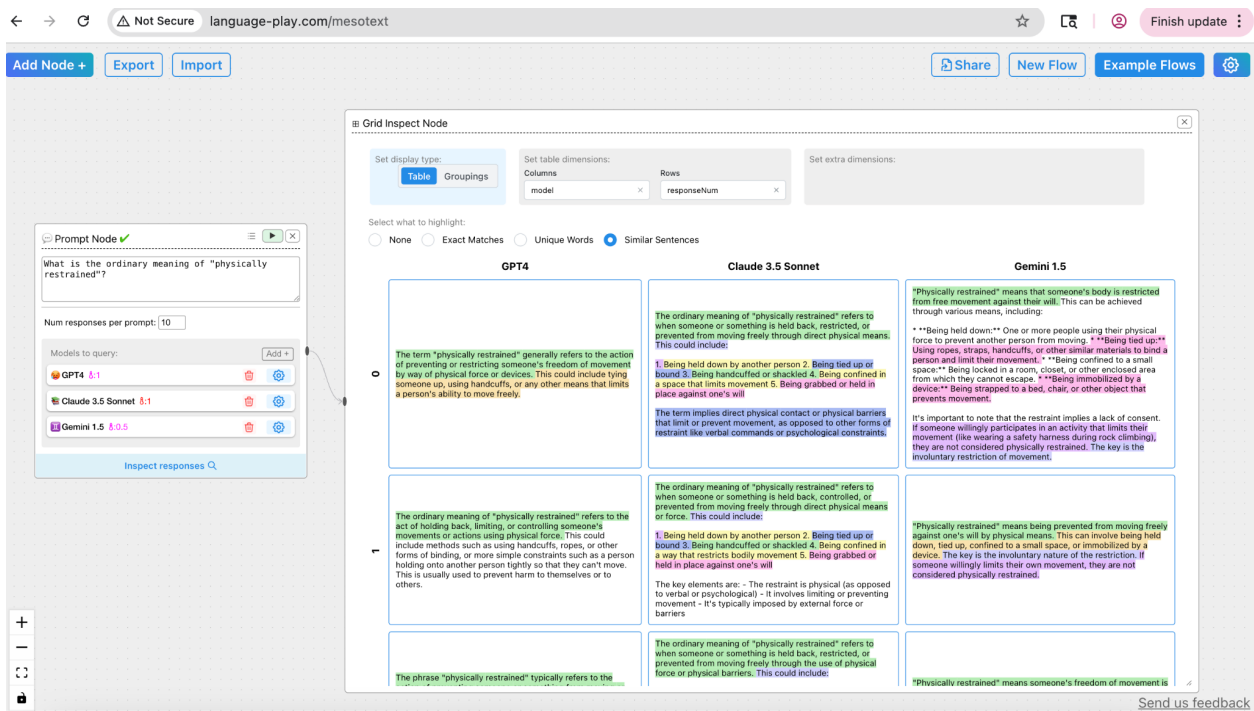


Figure 5: ChainForge “Prompt Node” configured to replicate Judge Newsom’s *Deleon* experiment by generating ten responses each from three LLMs. Unlike manual prompting across multiple tabs, this

⁶² See e.g. Neil Almond, *What Is Variation Theory? A Guide for Teachers*, THIRD SPACE LEARNING (Mar. 30, 2025), <https://thirdspacelearning.com/us/blog/variation-theory/> (describing variation theory and its practical pedagogical value).

⁶³ An API key for each service is required, and the user must supply these to make use of our tooling. Obtaining an API key is straightforward and does not involve coding skills. Likewise, the cost of these services for the scale of use likely by judges or other legal actors would be low.

interface centralizes sampling, reduces user burden, and supports systematic comparisons across model outputs and prompt variations.⁶⁴

While our use of ChainForge allows for significantly more efficient generation of responses across multiple models, we recognize that scaling up the number of responses, regardless of method, increases the volume of model responses to analyze and compare, which (without assistance) can itself be cognitively demanding or prohibitive. We will later discuss tooling support designed to help mitigate this increased cognitive load. For now, however, we continue with Judge Newsom’s approach.

As far as we can tell, Judge Newsom would have analyzed the 30 outputs from the LLMs by reading them, comparing them, and using the usual textual analysis tools of law. This would have been a laborious task—the outputs we produced with a similar query came to nearly 2000 words. That’s a lot of text to read closely. Judge Newsom analyzed the outputs carefully, as is clear from the *Deleon* extracts discussed *supra*. He noticed nuance in both inter- and intra-model variation.

Though Judge Newsom claimed that using LLMs could take some of the human discretion away from prior methods, such as dictionary searches, his interpretation of that text arguably put the human discretion right back in. We cannot know exactly what factors drove how he analyzed the text. Given the volume of LLM responses he reviewed and characterized, it likely required more time and cognitive effort than it would have with sensemaking tools to support his efforts.

We make use of a technique some called Positional Diction Clustering (PDC)⁶⁵, which computationally identifies analogous sentences across responses, on the basis that they share similar locations within responses (position) and wording (diction), helping to surface underlying structural patterns and variations across outputs. In Table 2, we show one way of rendering the results of PDC run on 10 outputs from each of 3 models.⁶⁶ The same-colored highlighting is given to analogous sentences across outputs, regardless of source model. This visible representation of pre-computed cross-output relationships would provide explicit support for Judge Newsom’s conclusions. But, unlike the analysis of a human reader, these similarities are flagged and classified transparently with the use of a conceptually simple algorithm.

⁶⁴ Judge Newsom used Gemini 1.5 Flash.

⁶⁵ More details about this and other methods of analyzing LLM outputs are available in Katy Ilonka Gero et al., *Supporting Sensemaking of Large Language Model Outputs at Scale*, in CHI ’24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., at 1 (May 11, 2024), <https://doi.org/10.1145/3613904.3642139>.

⁶⁶ This table represents the output and highlight visualization presented in our PDC-powered fork of the ChainForge interface. Grid lines, row numbering, and other affordances for manipulating the layout of this table of results are not shown here, but are captured in Katy Ilonka Gero et al., *Supporting Sensemaking of Large Language Model Outputs at Scale*, in CHI ’24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., at 1 (May 11, 2024), <https://doi.org/10.1145/3613904.3642139>. This could be archived into a file that can be shared with others. PDC has not yet been integrated into ChainForge but is deployed on a research prototype built on a research version of ChainForge. Those interested in using the tool may contact us to request access.

GPT4

The term "physically restrained" generally refers to the action of preventing or restricting someone's freedom of movement by way of physical force or devices. This could include tying someone up, using handcuffs, or any other means that limits a person's ability to move freely.

The ordinary meaning of "physically restrained" refers to the act of holding back, limiting, or controlling someone's movements or actions using physical force. This could include methods such as using handcuffs, ropes, or other forms of binding, or more simple constraints such as a person holding onto another person tightly so that they can't move. This is usually used to prevent harm to themselves or to others.

The phrase "physically restrained" typically refers to the action of preventing someone or something from moving or acting freely by use of physical force or restriction. This can include actions like holding someone down, tying them up, or using a device designed to limit movement like handcuffs or a straitjacket.

The ordinary meaning of "physically restrained" refers to the act of preventing someone from moving freely, usually with the use of physical force or constraint. This can involve methods such as handcuffs, ropes, straps, or anything else that limits someone's movements.

The ordinary meaning of "physically restrained" refers to the act of preventing someone from moving freely or confining someone's movements, typically using physical force or a device such as handcuffs, ropes, or other form of restraint. This can be used in a number of contexts such as law enforcement, healthcare, or personal safety situations.

The term "physically restrained" typically refers to the act of preventing someone from moving freely, usually with the use of physical force or a device that restricts movement. This could include anything from using handcuffs, ropes, or straitjackets, to holding someone down physically.

"Physically restrained" commonly refers to someone being held back or prevented from moving freely, usually with the use of physical force or containment devices, such as handcuffs or a straitjacket. This could be due to safety reasons, medical procedures, law enforcement situations, or any other instances where control of movement is necessary.

The ordinary meaning of "physically restrained" refers to the act of limiting, hindering or preventing someone's physical movement using some level of force or a device. This could be done through methods like handcuffing, holding down, using restraints, or any other means that inhibits a person's freedom to move.

The ordinary meaning of "physically restrained" refers to someone being physically held back, confined, or prevented from moving. This could be done through the use of hands, ropes, chains, straps, or other devices, often used when a person poses danger to themselves or others. This term is commonly used in security and healthcare settings.

The ordinary meaning of "physically restrained" refers to the act of preventing someone from moving freely or limiting their mobility, usually using force or a device that restricts movement, such as handcuffs, ropes, or safety straps.

Claude 3.5 Sonnet

The ordinary meaning of "physically restrained" refers to when someone or something is held back, restricted, or prevented from moving freely through direct physical means. This could include:

- 1. Being held down by another person 2. Being tied up or bound 3. Being handcuffed or shackled 4. Being confined in a space that limits movement 5. Being grabbed or held in place against one's will

The term implies direct physical contact or physical barriers that limit or prevent movement, as opposed to other forms of restraint like verbal commands or psychological constraints.

The ordinary meaning of "physically restrained" refers to when someone or something is held back, controlled, or prevented from moving freely through direct physical means or force. This could include:

- 1. Being held down by another person 2. Being tied up or bound 3. Being handcuffed or shackled 4. Being confined in a way that restricts bodily movement 5. Being grabbed or held in place against one's will

The key elements are: - The restraint is physical (as opposed to verbal or psychological) - It involves limiting or preventing movement - It's typically imposed by external force or barriers

The ordinary meaning of "physically restrained" refers to when someone or something is held back, restricted, or prevented from moving freely through the use of physical force or physical barriers. This could include:

- 1. Being held down by another person 2. Being tied up or bound 3. Being handcuffed or shackled 4. Being confined in a space that restricts movement 5. Being grabbed or held in place against one's will

The term implies direct physical contact or physical means to limit or control movement, as opposed to other forms of restraint like verbal commands or psychological constraints.

The ordinary meaning of "physically restrained" refers to restricting or limiting someone's bodily movement through physical means or force. This typically involves using one's hands, arms, or body to hold someone back, or using devices like handcuffs, straps, or restraints to prevent them from moving freely. The term implies direct physical contact or mechanical means to restrict movement.

The ordinary meaning of "physically restrained" refers to when someone or something is held back, confined, or prevented from moving freely through the use of physical force, barriers, or restraining devices. This could include being held down by another person, being tied up, handcuffed, or otherwise physically prevented from moving normally.

The ordinary meaning of "physically restrained" refers to when someone or something is prevented from moving freely through the use of physical force, barriers, or restraining devices. This could include being held down, tied up, handcuffed, or otherwise restricted in bodily movement through physical means. The term implies direct physical limitation of a person's or object's ability to move.

The ordinary meaning of "physically restrained" refers to when someone or something is held back, restricted, or prevented from moving freely through direct physical means. This typically involves using physical force, barriers, or restraining devices (like handcuffs, straps, or holds) to limit or control movement. The term implies direct physical contact or physical barriers that restrict freedom of movement.

The ordinary meaning of "physically restrained" refers to when someone or something is prevented from moving freely through the use of physical force, barriers, or restraining devices. This could include:

- 1. Being held back by another person 2. Being tied up or bound 3. Being handcuffed or shackled 4. Being confined in a space that restricts movement 5. Being held in place by any physical means

The term implies direct physical limitation of movement, as opposed to other forms of restraint like verbal commands or legal restrictions.

The ordinary meaning of "physically restrained" refers to someone being held back, restricted, or prevented from moving freely through direct physical means or force. This could include being held down by another person, being tied up, handcuffed, or otherwise mechanically prevented from moving. The term implies direct physical limitation of a person's bodily movement or mobility.

The ordinary meaning of "physically restrained" refers to the use of physical force, barriers, or restraints to limit or prevent someone's movement or physical actions. This could include holding someone back, tying them up, handcuffing them, or otherwise restricting their ability to move freely using physical means.

Gemini 1.5

"Physically restrained" means that someone's body is restricted from free movement against their will. This can be achieved through various means, including:

- * **Being held down:** One or more people using their physical force to prevent another person from moving.
* **Being tied up:** Using ropes, straps, handcuffs, or other similar materials to bind a person and limit their movement.
* **Being confined to a small space:** Being locked in a room, closet, or other enclosed area from which they cannot escape.
* **Being immobilized by a device:** Being strapped to a bed, chair, or other object that prevents movement.

It's important to note that the restraint implies a lack of consent. If someone willingly participates in an activity that limits their movement (like wearing a safety harness during rock climbing), they are not considered physically restrained. The key is the involuntary restriction of movement.

"Physically restrained" means being prevented from moving freely against one's will by physical means. This can involve being held down, tied up, confined to a small space, or immobilized by a device. The key is the involuntary nature of the restriction. If someone willingly limits their own movement, they are not considered physically restrained.

"Physically restrained" means someone's freedom of movement is restricted against their will using physical force, devices, or confinement. This could involve being held, tied, locked up, or otherwise immobilized. The key element is that the person is "not" choosing to be restricted.

"Physically restrained" means a person's body is forcibly prevented from moving freely. This can be done by holding them down, tying them up, locking them in a confined space, or using any other method that restricts their movement against their will. The essential component is the lack of consent; voluntary restriction of movement is not considered physical restraint.

"Physically restrained" means someone is prevented from moving freely against their will through the use of physical force, devices, or confinement. Key aspects include:

- * **Force or restriction:** Their movement is actively being limited.
* **Against their will:** The person does not want to be restricted.
* **Physical means:** The restriction involves something tangible, like being held, tied, or locked up, rather than just being told not to move.

"Physically restrained" means a person's movement is restricted against their will using physical means, such as being held, tied up, locked in, or otherwise immobilized. The key is that the restriction is involuntary and imposed by an external force or barrier.

"Physically restrained" means someone's freedom of movement is restricted against their will by physical means. This could involve someone holding them, tying them up, confining them to a space, or otherwise preventing them from moving freely. The emphasis is on the involuntary nature of the restriction – the person is not choosing to be restrained.

"Physically restrained" means someone is prevented from moving freely against their will through physical means. This can be done by holding them, tying them up, locking them in a confined space, or using any other method that restricts their body's movement. The crucial element is that the restriction is involuntary; the person is "not" choosing to limit their own movement.

"Physically restrained" means that someone's ability to move freely is limited against their will using physical force, devices, or barriers. They are being held back, tied up, confined, or otherwise prevented from moving as they wish. It's important that this is done "against their will"; someone willingly restricting their own movement (like wearing a seatbelt) is not considered physically restrained.

"Physically restrained" means a person's free movement is restricted against their will through physical means, such as by being held, tied, or confined. It implies the use of force or barriers to prevent movement, and that the person being restrained is not consenting to this restriction.

Table 2: Positional Diction Clustering (PDC) applied to 30 responses across three LLMs, highlighting semantically and structurally analogous sentences. The green-highlighted "core" definitions of "physically restrained" corroborate Judge Newsom's inference, but here the clustering is transparent, scalable, and reproducible, offering a principled alternative to subjective textual synthesis. In this case, the common core is a definition of physical restraint (highlighted in green) involving verbs like restrict, prevent, control, and limit, as well as antonyms like freedom and will. The common core is not so complicated; it is largely a matter of diction. Also, from this highlighting, we

can readily see that among the 30 outputs there are five distinct concrete examples used, highlighted in a sequence of matching colors (yellow, orange, gray, pink purple). Seven of the 30 responses include the models' inferred implications of the term (blue) beyond a definition.⁶⁷

We don't expect the reader to inspect the text of Table 2 (more than 1800 words) in detail. Rather, we hope the reader will scan and notice that *every* output from *every* model includes green highlighting. The words highlighted in green are associated with the use of physical force or restraints to limit or prevent movement, plausibly that same "core" Judge Newsom identifies. However, this core has now been generated by an algorithm that computes diction and position similarities across sentences in different responses, rendering the highlighting more transparent than similarities computed according to a semantic embedding. The algorithm is also open source.⁶⁸

This highlighting analysis can scale, too. If we worry that even $N = 10$ data points are not enough (a valid concern, though sample sizing is beyond the scope of this work), we can scale the query further. We scaled Judge Newsom's experiment up by another factor of 10 to $N = 100$ (not shown) and observed that same "common core" within the green highlighting, which gives us added (but still imperfect) confidence in the consistency (but not necessarily the correctness) of these outputs.⁶⁹ PDC is scalable, archivable, transparent, open-source, and involves no training, so it cannot be "trained politically".⁷⁰ In short, it may well be the AI-resilient interfaces *built on top of* LLMs that can have many of the virtues Judge Newsom invoked in *Snell* to justify his interest in LLM-informed jurisprudence. The use of AI-resilient interfaces like ChainForge, especially with cognition-supporting enhancements like PDC, is superior to a judge or his clerks chatting with a consumer-facing bot, even if they meticulously (and laboriously) recorded the data they gathered and presented the data in an appendix.

⁶⁷ Screen capture of the tooling with these outputs is available here: <https://tinyurl.com/yuwajn2h>. The original archival .cforge file is on file with the authors and available upon request.

⁶⁸ Elena Glassman, *Positional Diction Clustering / Ordered Groups, now with a clearer input spec*, OBSERVABLE HQ/@glassmanlab, <https://observablehq.com/@glassmanlab/positional-diction-clustering>, (last edited July 13, 2024).

⁶⁹ Results of the $N = 100$ experiment for GPT 3.5 and GPT 4 can be seen here: <https://tinyurl.com/3xtpdkn2>. Note: the word count of these 100 outputs for 2 models rises to around 9500 words. If a human actor attempted to analyze so many outputs without mechanical assistance, that would be very tedious and error prone indeed.

⁷⁰ Cf Margaret H. Lemos, *The Politics of Statutory Interpretation*, 89 NOTRE DAME L. REV. 849-908 (2013) (reviewing Antonin Scalia & Bryan A. Garner, *Reading Law: The Interpretation of Legal Texts* (2012)). The tool is not "objective" in that it has all the usual problems of association with consumer-facing LLM. In fact, ChainForge runs in part on the same technology (using proprietary LLMs) and therefore it also bakes in additional exposure vis-a-vis whatever is hidden in the opaque training and deployment of such models. *On the other hand* if an adequate open source model were available, the same methods that power ChainForge with proprietary LLMs currently could also be used with open source models.

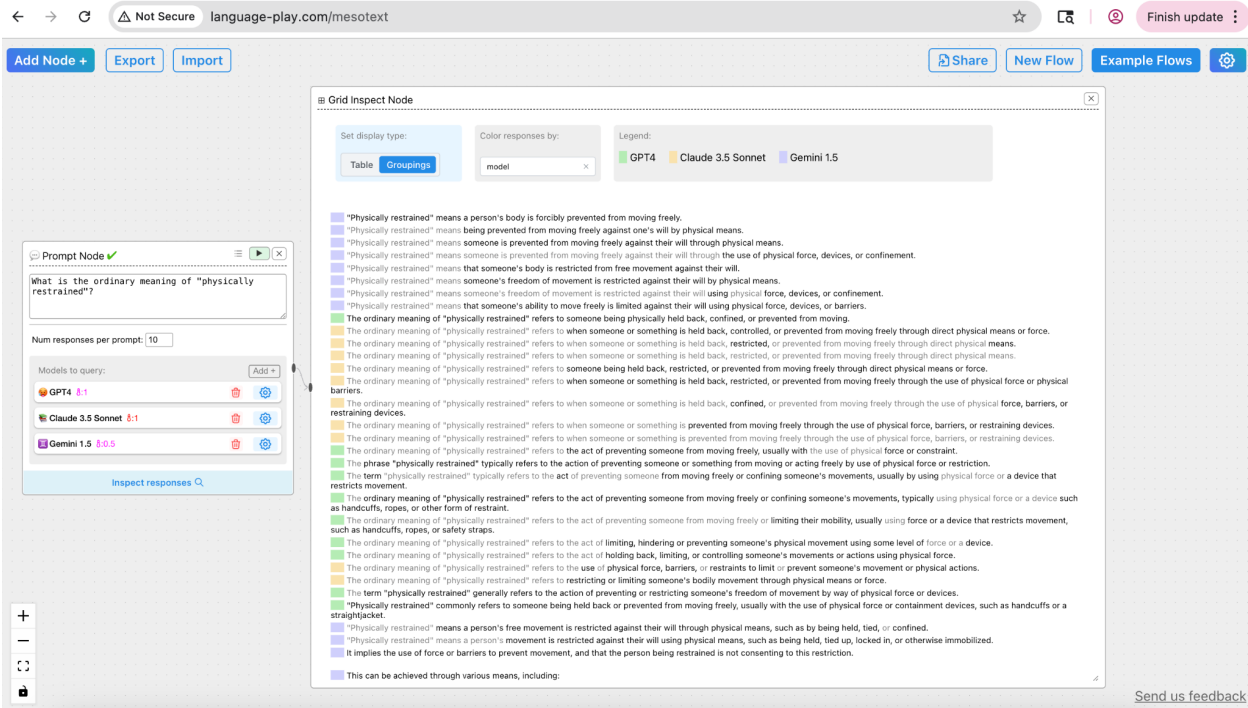


Figure 6: Interleaved View of PDC re-organizes the 30 Deleon responses by sentence similarity, de-emphasizing repetitive text to surface linguistic variation. This alternative rendering enables more efficient sensemaking and comparison across outputs, illustrating how interface design can foreground either convergence (Figure 5) or divergence (Figure 6) in the *same* LLM outputs, depending on user goals.

The great thing about this interface is there are many ways to organize information, and each one is valuable for different evaluation tasks or user preference. The grid view presented in Table 2 and Figure 5 more readily allows the reader to see how certain groups of analogous sentences, such as definitional sentences, manifested across different models' responses. On the other hand, where text is similar but slightly different within these groups, it may be more helpful to emphasize what is different rather than what is the same. This is part of what is accomplished in Figure 6, which shows the Interleaved View⁷¹ of PDC applied to the same outputs shown in Table 2.

Here, the redundant text within groups of sentences with similar diction and position within their respective responses are rendered in a lighter color to de-emphasize it, permitting the reader to concentrate on what is different rather than what is the same. The Interleaved View and Grid View of PDC present the same underlying collection of LLM responses, but how these responses are rendered to the user in the interface surfaces different patterns. What becomes more or less salient depends on the interface choices, highlighting how interface design can influence the kinds of comparisons, and the resulting possible insights, that are most accessible to the user. Making multiple complementary views available is ideal. This serves to highlight our larger

⁷¹ Also explained in more detail in Katy Ilonka Gero et al., *Supporting Sensemaking of Large Language Model Outputs at Scale*, in CHI '24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., at 1 (May 11, 2024), <https://doi.org/10.1145/3613904.3642139>.

point: the content of AI outputs can be understood multidimensionally, but probably only where the interface readily enables us to do so.

It's important to note that AI-resilient interfaces are not a solution to conceptual problems baked into certain use cases or success criteria for assessing LLM outputs. Our identification of a method to scale and systematize the analysis Judge Newsom undertook manually should not be mistaken for an endorsement of Judge Newsom's standard for judging the utility, correctness, or informativeness of LLM outcomes based on consistency across a large number of samples. Indeed, there are reasons to highlight the dangers rather than advantages of highly consistent LLM outputs. For example, there is empirical work showing that consistent but wrong LLM outputs can lead human users to doubt their own correct beliefs.⁷² What's more, the search for consistency in LLM outputs may only exacerbate the danger of the illusory truth effect, wherein people are more likely to believe false information when they are more often exposed to that false information.⁷³ Consistent outputs from an LLM should not necessarily be taken as evidence of correctness. Consistency may, independently of right or wrong on the facts, be evidence of dangerousness, insofar as human abdication to LLMs is understood as a danger to be avoided. Judge Newsom's emphasis on consistency as proof of correctness also ignores the well-known statistical distinction between accuracy and precision.⁷⁴ At most, consistency of LLM outputs, if anything, might plausibly suggest high precision rather than high accuracy.

All in all, in discussing Judge Newsom's analysis in *Deleon*, we demonstrated that appropriate tooling can enhance the efficiency, scalability, and transparency of the methodology Judge Newsom proposed. But, the AI-resilient tooling we showcase here does not prevent conceptually problematic misuses of LLM. AI-resilient tooling cannot replace the human need for deep introspection about how and why we might use LLMs to assist in adjudication. AI-resilient interfaces can enable better use of AI, but only to the extent that the underlying use itself is desirable and legitimate.

Why judges turn to AI when they shouldn't: a Priest-Klein hypothesis for LLMs

Given the hard cases Judge Newsom tackled with LLMs, we speculate that there is a possibility that issues selected for LLM assays by the judiciary will tend to be exactly those questions where the law and other forms of human judgment do not offer ready clarity. In other words, it may be that the judiciary could develop a tendency to go to LLMs precisely when, and because, a legal

⁷² Chelse Swoopes et al., *The Impact of Revealing Large Language Model Stochasticity on Trust, Reliability, and Anthropomorphization*, TREW Workshop at CHI '24: CHI Conf. on Hum. Factors in Computing Sys., (May 11, 2024), https://chi-trew.github.io/papers/2024/CHI_TREW_2024_Paper_19.pdf.

⁷³ Illusory Truth Effect, WIKIPEDIA, https://en.wikipedia.org/wiki/Illusory_truth_effect (Mar. 19, 2025).

⁷⁴ See Practices of Science: Precision versus Accuracy, Univ. of Haw. at Mānoa, <https://manoa.hawaii.edu/exploringourfluidearth/physical/world-ocean/map-distortion/practices-science-precision-vs-accuracy> (last visited Mar. 19, 2025).

question is difficult or uncertain. We emphasize that our hypothesis is not about tasks that are tedious or that require a high degree of effort. Those are different kinds of difficulty. There, it's obvious that any human would be tempted to use an LLM they perceived as competent. Rather, we are interested in cases where *making* the decision—rather than communicating the decision—is challenging, as was likely the case in *Snell* and *Deleon*. The difficulty in these cases came down to the closeness of parsing the language. It is this very conceptual difficulty that may drive decision makers to LLMs.

If this hypothesis sounds familiar, that's because it should. Decades ago, George Priest and Benjamin Klein conjectured that uncertain cases are more likely to continue through litigation and produce a judicial decision, rather than settle.⁷⁵ Priest and Klein hypothesized that the cases that were litigated would not be a random sample of disputes that were representative of the qualities of possible or actual legal disputes in the world. Rather, contentious, close cases would be the ones most likely to lead to a resolution by litigation rather than by settlement. Parties would go through the exhaustive, information- or outcome- producing process of litigation precisely because they could not resolve the hard questions on their own.

Likewise, we predict that judges will not turn to LLMs in a random fashion. Rather, judges may turn to LLMs in a manner similar to Judge Newsom's experience in *Snell*. Judge Newsom described himself spending hours and hours thinking about the interpretive question, and further turning to LLMs when the more traditional method of a dictionary failed him.⁷⁶ Likewise, when other judges turn to LLMs, they may turn to LLMs when other methods fail them, which is likely to be in close or hard cases.

Our hypothesis about judicial selection hasn't yet been investigated, but empirical work in other areas bears out this possibility. For example, a laboratory user study showed that users were more likely to turn to AIs in the face of uncertainty in a trip planning task.⁷⁷ There is also evidence of similar behavior in high stakes, real world situations. Empirical research on the use of AI to guide decisions about child welfare shows this pattern: child protection workers selectively make use of algorithmic guidance, relying more on these algorithms when they have greater uncertainty about a situation (alongside other technical and institutional factors).⁷⁸ We are not certain that this pattern applies to judging, but it seems a distinct possibility.

⁷⁵ George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1 (1984), <https://chicagounbound.uchicago.edu/jls/vol13/iss1/2>.

⁷⁶ We agree with Judge Newsom that the use of dictionaries is itself problematic, and we don't here claim that the use of dictionaries would be preferable.

⁷⁷ Sara Salimzadeh et al., *Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making*, in CHI '24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., 1, Honolulu, Haw., ACM, May 11–16, 2024, (2024) <https://doi.org/10.1145/3613904.3641905>.

⁷⁸ Anna Kawakami et al., *Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support*, in Proceedings of the 2022 CHI Conf. on Hum. Factors in Computing Sys. (CHI '22), New Orleans, La., ACM, Apr. 29–May 5, 2022, (2022) <https://doi.org/10.1145/3491102.3517439>.

There is an urgent empirical question to understand whether judges, legal scholars, and others are most likely to turn to LLMs precisely to address the legal questions that are most intrinsically difficult.⁷⁹ Such behavior ought to be discouraged. Judges occupy their positions of tremendous power and responsibility to address the hard questions. It is this judicial contemplation that legitimizes decisions, even in close cases. Avoiding the hard calls would particularly raise questions of judicial legitimacy. Even advocates for incorporating AI into law emphasize that AI will do best to routinize tedious, straightforward work with a clear correct answer. If a Priest-Klein phenomenon obtains in the selection of legal questions for LLMs, we will have just the opposite outcome: AI will be used precisely when the questions are hardest.

This observation is more than facile legal theory. We speculate that it may very well explain why some of Judge Newsom's *Snell* LLM outputs proved so brittle. For all we know, close questions of interpretation may also yield LLM outputs that could easily go either way.⁸⁰ Perhaps Judge Newsom wasn't randomly unlucky that his analysis was so brittle; perhaps that analysis was brittle at least in part because the question was a difficult one. This may explain the disagreement we found in ChatGPT and Gemini as to the *Snell* question, as documented in Table 1. It may simply be that *Snell* was a hard case and was therefore more likely than other potential cases to produce brittleness in LLM outputs. In any case, judges should not turn to LLMs precisely because other tools or methods leave them unsatisfied. Indeed, such scenarios may be the worst time to turn to an LLM.⁸¹

In our emphasis on interface and choice architecture, we again part ways with best practices suggested in other recent scholarship. To return to a point *supra*, we worry that Waldon et al.'s recommendation to use LLMs dialectically would be especially problematic and prone to

⁷⁹Judge Newsom seemingly chose two thorny cases of textual interpretation as opportunities to highlight the potential for LLMs. In *Snell*, Judge Newsom acknowledged gratefully dodging the issue ultimately based on state law that automatically incorporates earlier insurance policies, one of which offered language to cover the dispute and therefore spare the 11th Circuit the need to squarely address whether an in-ground trampoline is landscaping. In *Deleon*, Judge Newsom recognized on-point precedent but called for an en banc rehearing to reverse that precedent. The circumstances of each case suggest that Judge Newsom chose *hard* rather than *straightforward* cases to test out his LLM-based methodology. It may be that published federal circuit court cases typically address hard cases. It is, of course, a difficult theoretical and empirical matter to define and identify hard cases, but there are good reasons to expect that many circuit court cases—even those covered in public opinions—are not necessarily hard cases. Joshua B. Fischman, *How Many Cases Are Easy?*, 13 J. LEGAL ANALYSIS 595 (2021), <https://doi.org/10.1093/jla/jaaa010> (discussing a range of empirical metrics to determine proportion of easy cases in a judge's case load and by one metric identifying substantially varying rates of "hard" cases by circuit and by legal domain, in other words showing that not all cases in the federal circuit are hard cases). In reviewing other cases Judge Newsom participated in during the relevant time frame, we were able to identify at least one other statutory interpretation case, showing that Judge Newsom at least had some choice in which cases he used to showcase LLMs as informing judicial interpretation.

⁸⁰ On the other hand, there are reasons to believe that LLM performance is not tied to the difficulty of the legal question. See Matthew Dahl et al., *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, 16 J. LEGAL ANALYSIS 64 (2024), <https://doi.org/10.1093/jla/jaae003> (finding that LLMs are not calibrated with regard to hallucination, meaning that the model's confidence does not correlate with its probability of correctness).

⁸¹ Judges will have difficulty gut-checking an LLM output when they are themselves uncertain. Likewise, when other method have failed, judges will have nothing to cross-check the LLM against. There are also further reasons we leave out for succinctness.

anchoring if a Priest-Klein phenomenon obtains in the selection of legal cases for LLM analysis. We also worry about advice from Grimmelmann et al., who have advised that LLMs might have rhetorical value in close cases.⁸² They have suggested that LLM outputs might serve as judicial gut checks, with judges exploring their own opinions through their reactions to LLM outputs. We worry that this too could be particularly problematic, and for the same reasons, the production of anchoring and other problematic cognitive biases when judges are most uncertain. We need to understand why judges might resort to LLMs in some cases and not others, and to shape best practice recommendations accordingly.

New possibilities

Much legal scholarship and jurisprudence has focused on what LLMs can and cannot do. But we need more focus on what humans can do with LLMs and how appropriate AI-resilient interfaces can facilitate more efficient, effective, and accurate human legal work supplemented by LLM capacities. AI-resilient interfaces already have a lot to offer people like Judge Newsom, who wish to experiment with the use of LLMs,. AI-resilient interfaces can particularly ensure that LLM users engaging in such exploratory exercises maintain an appropriately critical perspective on LLM outputs. The right kind of AI assistance can go a long way towards mitigating the risks incurred when humans inappropriately defer to or otherwise misuse LLMs in law.⁸³ We eagerly extend our hand to Judge Newsom and any like-minded colleagues who wish to collaborate on building such tools.

While much recent scholarship has focused on possibilities and problems associated with generative interpretation, we see some promising use cases that have been far less discussed in legal scholarship: generative reading (that is, LLM-assisted reading) and generative drafting (that is, LLM-assisted but not LLM-performed legal writing). Here we discuss some open source tools that provide AI-resilient interfaces for generative reading and generative drafting. In both cases, the tools leave the human user fully empowered to make decisions without defaults and to have access to all information that would enable the user to identify and reject inappropriate or incorrect AI assistance. Below, we offer sketches of what is possible, and we are actively building out similar projects for legal applications.

Generative reading

In the spirit of emphasizing human enhancement rather than human replacement, we begin by exploring the concept of *generative reading*. While substantial attention has been given to whether LLMs can read on our behalf, far less consideration has been devoted to whether and how LLMs might help us read more effectively. Rather than having LLMs parse documents for

⁸² Grimmelmann et al. manuscript at 65.

⁸³ Aileen Nielsen, et al., *Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency*, 21 J. EMPIRICAL LEGAL STUD. 979 (2024).; for non-legal work, see e.g. Albert Ziegler et al., *Measuring GitHub Copilot's Impact on Productivity*, COMM. ACM (Feb. 15 2024), <https://cacm.acm.org/research/measuring-github-copilots-impact-on-productivity/>.

the judiciary, we consider how they might assist judges and clerks in themselves parsing documents more efficiently.

Across multiple experiments, LLM-generated outputs have been shown to support human readers in processing text more effectively when delivered through AI-resilient interfaces. For example, recursive extractive summarization to adjust the visual salience of text (e.g., by modulating text darkness), can achieve this result. So can selective highlighting. Despite differences in implementation, different forms of generative assistance—when carefully designed to emphasize salient content or downplay extraneous detail—can improve human reading without misleading the reader. The benefits can be substantial: in some cases, participants completed reading tasks up to 30% faster and even demonstrated improvements in comprehension accuracy.⁸⁴ Such improvements should not be left untapped, particularly by those engaged in high-stakes textual interpretation, like members of the judiciary. But these AI use benefits are only reliably realized when paired with thoughtfully designed, AI-resilient interfaces.

The workings of the text rendering tools that drive such improvements can be more complicated than the functions we highlighted in ChainForge’s PDC supplement. For example, in the case of GP-TSM, the underlying tool relies on LLMs to generate the highlighting of text, thus rendering this method not fully open source or transparent.⁸⁵ GP-TSM is a browser-based tool that enables dynamic text modulation for enhanced reading. In this case, the transparency and resiliency come from the use of the outputs. These outputs are not suggested to humans as stand-ins for reading holistically and critically, but these outputs can assist humans in what they wish to do otherwise, such as dynamically switching between AI-assisted skimming and fully reading paragraphs of a document as influenced by one’s reading goals. Users of GP-TSM can do this without lifting a finger to adjust the document’s GP-TSM rendering, since no text is hidden.

This paragraph demonstrates bolding at varying levels of intensity, as rendered by GP-TSM As visible here, GP-TSM facilitates multiple modes of engagement: readers may choose to focus exclusively on the grammatically coherent core of each sentence or delve into the progressively lighter, less critical, yet still legible contextual details. While this layered presentation may initially appear distracting—particularly if a reader is not experienced with this method—it typically requires only minimal acclimatization. Even in brief exposures during experimental sessions, readers experienced measurable benefits.⁸⁶

Finally, and important to AI-resiliency, GP-TSM is designed to fail gracefully. It sometimes happens that the AI component is unable to confidently decompose a passage into core assertions and surrounding context. In that case, the interface presents the text in uniform salience (standard

⁸⁴ Ziwei Gu et al., *An AI-Resilient Text Rendering Technique for Reading and Skimming Documents*, in CHI ’24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., Honolulu, Haw., May 11–16, 2024, (2024). Aileen Nielsen et al., *Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency*, 21 J. EMPIRICAL LEGAL STUD. 979 (2024).

⁸⁵ Ziwei Gu et al., *An AI-Resilient Text Rendering Technique for Reading and Skimming Documents*, in CHI ’24: Proceedings of CHI Conf. on Hum. Factors in Computing Sys., Honolulu, Haw., May 11–16, 2024, (2024).

⁸⁶ Sample outputs of GP-TSM applied to privacy policies may be accessed at <https://tinyurl.com/yznc59hx> and <https://tinyurl.com/ysmtcnem>.

black front), thereby avoiding misleading emphasis. While the system cannot guarantee accurate comprehension by inattentive readers, our empirical data shows that the tool assists readers in navigating complex materials more efficiently, and sometimes even more correctly. Such a tool could likely assist judges and their clerks in reviewing the voluminous papers that comprise most litigation.

Generative drafting

Generative drafting is another example case where AI-resilient interfaces can provide gains without some of the drawbacks caused by overreliance on LLMs. In the case of generative drafting, humans can avoid the costs of having LLMs write *for* you, e.g., cognitive debt, unintentional plagiarism, and the time necessary to revise or customize LLM-generated text that is mismatched in tone, length or content as compared to what is desired.

Rather, generative drafting tools make it easier for the human to draft. AI is becoming increasingly good at identifying cross-document relationships, and judges might find efficiency gains in ready access to AI-retrieved prior examples that are similar along various aspects of a given case at hand, particularly in some structured format, that can inform their own drafting. For example, some kinds of cases are quite rare and are therefore not handled frequently by judges, such as antitrust cases in federal courts.⁸⁷ A tool that could help judges pull up and easily navigate prior examples of a genre of legal writing could substantially boost efficiency, without inappropriate human abdication of discretion.

CorpusStudio⁸⁸ offers such functionality at the document structure and individual sentence level. We present an example of the CorpusStudio, interface in Figure 7, showing key points of its functionality. CorpusStudio is a writing support environment that helps users write while identifying and reflecting on community specific writing norms at both the document organization and sentence levels. Drawing from a corpus of prior documents of the same type, it presents an ordered distribution⁸⁹ of how other authors have structured their documents⁹⁰ (left hand side of Figure 7). We can imagine that CorpusStudio could be refined for a number of legal purposes. In addition to judges who might draft decisions in unfamiliar domains, we can likewise imagine that first-year legal writing students would benefit from such a tool. Attorneys moving from one law firm to another and seeking to identify and emulate their new firm’s typical style of briefing and memo writing may also find this to be a useful tool.

⁸⁷ Douglas H. Ginsburg & Joshua D. Wright, *Antitrust Courts: Specialists Versus Generalists*, prepared for the FORDHAM COMPETITION LAW INST. (Sept. 20, 2012), https://www.ftc.gov/sites/default/files/documents/public_statements/antitrust-courts-specialists-versus-generalists/130722ginsburg_wright.pdf.

⁸⁸ Hai Dang et al., *CorpusStudio: Surfacing Emergent Patterns in a Corpus of Prior Work While Writing*, in CHI '25: Proceedings of the 2025 CHI Conf. on Hum. Factors in Computing Sys. (2025).

⁸⁹ This is an adaptation of the ‘interleaved view’ of the PDC algorithm [see citation on Gero et al. in note 23] applied to top-level section titles of documents in the corpus.

⁹⁰ It currently does this using the top-level section headings, but could potentially be done on AI-generated summaries of paragraphs that are a few words long, for example, for documents without author-provided section titles or where section titles are less formulaic than the original domain that CorpusStudio was designed for: academic papers.

Importantly, CorpusStudio does not generate text with LLMs; it instead uses semantic embeddings to retrieve example sentences from other documents in the corpus that are analogous to the writer’s current sentence(s). By retrieving many (i.e., 25) such sentences each time the user calls on it while writing (rendered using strategies like graying out repeated words across them, that, like PDC-powered views shown in Figure 5 and 6, help the user see emergent patterns), CorpusStudio demonstrates how many examples from a curated corpus can reveal both emergent norms and meaningful deviations to writers.⁹¹

CorpusStudio helps writers learn from prior papers written for similar venues or audiences by surfacing corpus-based patterns at both the document level (structure) and sentence level. The key functionalities of CorpusStudio are illustrated in Figure 7. On the left (A), the interface displays common section title sequences to help users identify frequent and infrequent organizational structures. In the center (B), writers draft their own outlines and prose. While writing, pressing TAB retrieves semantically similar sentences from similarly named sections of documents in the corpus (C), given their cursor’s current location in their document. The retrieved sentences are highlighted to show recurring language patterns and variation. Hovering over a sentence (D) reveals additional context about the sentence, and writers can also bookmark, annotate, or share examples they find particularly strong or weak in their User Notes. Drafts are cached locally and can be exported as LaTeX files.

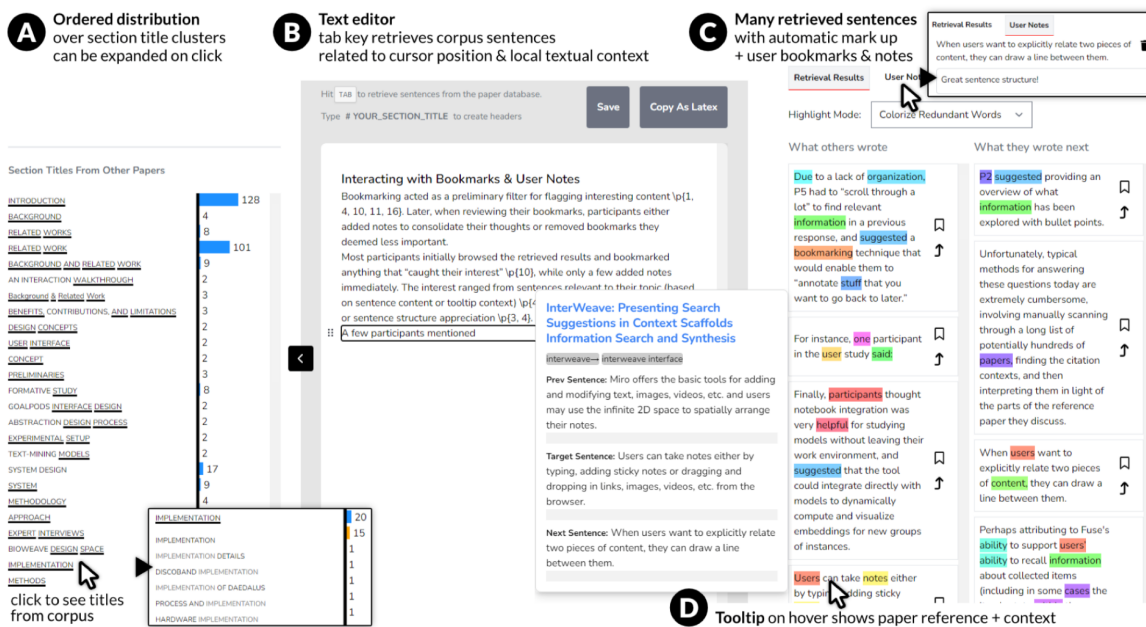


Figure 7: An AI-resilient writing interface (CorpusStudio) that aids authors by surfacing corpus-based structural and syntactic patterns. The interface shows frequent document structures (A), allows live drafting (B), and retrieves similar sentences from a curated corpus (C) with visual pattern cues. This promotes norm-awareness without defaulting to LLM-generated content, preserving user agency in generative drafting.

⁹¹ There are many benefits to this strategy; for legal and other professions’ documents of a particular type that have strong common patterns, rendering implicit writing norms can be particularly helpful.

CorpusStudio may seem like a tool useful only for academics, but we see many potential use cases in law at every stage of legal education and practice. For example, legal writing is a core part of the 1L curriculum at every U.S. law school, and law students often express that the skill of writing a research memo or a brief is a difficult one to master. They find it difficult both to read examples of such work and to understand and replicate the organizing vocabulary and structure of these works. A tool like CorpusStudio could be a helpful assistant, not replacing student writing but rather providing and organizing useful sets of examples.

The tool also has something to offer for judges. For example, consider the case of antitrust law. Antitrust cases are notoriously difficult and sometimes disliked by judges. On the one hand, judges do not get enough exposure to these cases to develop a mastery of the case law and reasoning. On the other hand, such cases are also more complex than more workaday areas of law, and necessarily rely heavily on economics and consultations with expert witnesses. It could be that a judge called upon to handle an antitrust case might make use of such a tool to more readily reacquire and redeploy the vocabulary, defining economics constructs, and structure of an antitrust judicial opinion with a tool like CorpusStudio. Most importantly, again, an AI-resilient tool like CorpusStudio doesn't replace humans but instead empowers them and helps them receive AI assistance that will augment—but not mislead—they in their tasks.

Conclusion

We don't yet have AI-resilient law, but we need a vigorous research agenda to ensure that outcome. AI-resilient interfaces are a crucial, and likely necessary tool to support AI-resilient law. There is already a body of tooling and empirical research that enables and justifies AI-resilient uses of LLMs for law. The current intense interest in LLM-informed interpretation represents only a narrow (and problematic) slice of what is possible with AI and law. There are many generative ways to incorporate LLMs into law with AI resiliency firmly in place. We continue to build towards that goal and hope others will do the same.