FROM TEXT TO TECH: THE ROLE OF AI IN MODERN LEGAL INTERPRETATION

By: Chris Cyrus, Shawn Helms, Jason Krieser

During the 2015 Scalia Lecture at Harvard Law School, U.S. Supreme Court Justice Elena Kagan declared that "we are all textualists now."[1] And we are.

All textualists know that in matters of interpretation, "text is always the alpha."[2] Even "canons of construction are no more than rules of thumb," as "[t]ext is the 'one[] cardinal canon' a court must turn to 'before all others.'"[3] This cardinal canon "is not a lawyer's game to 'divine arcane nuances' and 'discover hidden meanings.'"[4] Instead, "[i]n their full context, words mean what they conveyed to reasonable people at the time they were written."[5] Under this fair-reading approach, "[o]ur duty instead is to follow the natural, everyday meaning of the words enacted into law."[6]

In a recent (and very non-standard) concurring opinion, U.S. Circuit Judge Newsom asks us all to consider whether every judge and attorney in the United States should familiarize themselves with, and utilize a new, powerful class of tools: large language models (LLMs).[7]

Our answer is yes. In fact, we believe it is incumbent upon our entire profession to use, as one tool in our toolset, LLMs to help illuminate the "ordinary meaning" of words and phrases in legal text. And while Judge Newsom has already illustrated

---

[1] Christine Perkins, *In Scalia Lecture, Kagan Discusses Statutory Interpretation*, HARV. L. SCH. (Nov. 25, 2015), https://hls.harvard.edu/today/in-scalia-lecture-kagan-discusses-statutory-interpretation/.

[2] United States v. Palomares, 52 F.4th 640, 648 (5th Cir. 2022) (Oldham, J., concurring) (quoting In re Deberry, 945 F.3d 943, 947 (5th Cir. 2019)); *see also In re* Deberry, 945 F.3d at 947 ("[I]t's also the omega.").

[3] Kisor v. McDonough, 995 F.3d 1347, 1348 (Fed. Cir. 2021) (Prost, C.J., concurring) (quoting Conn. Nat'l Bank v. Germain, 503 U.S. 249, 253 (1992)).

[4] Thomas v. Reeves, 961 F.3d 800, 807 (5th Cir. 2020) (Costa, J., concurring) (quoting ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 69 (2012)).

[5] *Palomares*, 52 F.4th at 648 (quoting ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 16 (2012)).

[6] *Reeves*, 961 F.3d at 807 (Costa, J., concurring) (citing ANTONIN SCALIA & BRYAN A. GARNER, READING LAW: THE INTERPRETATION OF LEGAL TEXTS 33 (2012)).

[7] Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1221–30 (11th Cir. 2024) (Newsom, J., concurring) ("Here's the proposal, which I suspect many will reflexively condemn as heresy, but which I promise to unpack if given the chance: Those, like me, 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 like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude might—*might*—inform the interpretive analysis. There, having thought the unthinkable, I've said the unsayable. . . . For all these reasons, and perhaps others I haven't identified, it seems to me that it's at least worth *considering* whether and how we might leverage LLMs in the ordinary-meaning enterprise—again, not as the be all and end all, but rather as one aid to be used alongside dictionaries, the semantic canons, etc.").

many of the capabilities of current-generation LLMs,[8] we contribute our voices to the dialogue to speak to several questions that he raises—including how forthcoming advances in the technologies and our use of LLMs will continue to mitigate the shortcomings of currently available LLM tools.

*   *   *

Judge Newsom presents a balanced argument regarding the use of LLMs to assist lawyers and judges with interpretive tasks. On the one hand, he highlights many of the advantages that LLMs bring to the interpretive enterprise: These models are trained on vast swaths of ordinary-language,[9] can "understand" context,[10] offer a degree of transparency (despite their black-box nature) compared to opaque human methods of dictionary editing and selection,[11] and hold advantages over empirical interpretive methods like surveys or corpus linguistics.[12] But Judge Newsom also acknowledges a series of drawbacks: LLMs can "hallucinate,"[13] fail to capture pockets of human speech and usages,[14] are susceptible to manipulation by lawyers, judges, and litigants,[15] and risk unintended abdication in exercises of human judgment.[16]

---

[8] *Compare Snell*, 102 F.4th at 1226 (Newsom, J., concurring) ("In what follows, I'll try to sketch out what I take to be some of the primary benefits and risks of using LLMs—to be clear, as one implement among several in the textualist toolkit—to inform ordinary-meaning analyses of legal instruments."), *with* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, CTR. FOR RSCH. ON FOUND. MODELS, STANFORD INST. FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE & STANFORD UNIV. (July 12, 2022), https://arxiv.org/pdf/2108.07258 ("Foundation models acquire various *capabilities* that can power applications. We have chosen to discuss five potential capabilities: the ability to process different modalities (e.g., language, vision), to affect the physical world (robotics), to perform reasoning, and to interact with humans (interaction). . . . These and other instances of recent foundation models not only achieve remarkable performance across a multitude of diverse downstream tasks and applications, but also manifest noteworthy behaviors of interpretability, robustness, controllability and generalization.").

[9] *Snell*, 102 F.4th at 1226 (Newsom, J., concurring) ("The ordinary-meaning rule's foundation in the common speech of common people matters here because LLMs are quite literally 'taught' using data that aim to reflect and capture how individuals use language in their everyday lives. Specifically, the models train on a mind-bogglingly enormous amount of raw data taken from the internet.").

[10] *See id.* at 1227 (Newsom, J., concurring) ("[They] absorb and assess the use of terminology in context and [are] empower[ed] . . . to detect language patterns.").

[11] *See id.* at 1228 (Newsom, J., concurring) ("[A]lthough we tend to take dictionaries for granted, as if delivered by a prophet, the precise details of their construction aren't always self-evident.").

[12] *See id.* at 1230 (Newsom, J., concurring) ("LLMs hold advantages over other empirical interpretive methods.").

[13] *Id.* (Newsom, J., concurring) ("Put simply, an LLM 'hallucinates' when, in response to a user's query, it generates facts that, well, just aren't true—or at least not *quite* true.").

[14] *Id.* at 1231 (Newsom, J., concurring) ("LLMs don't capture offline speech, and thus might not fully account for underrepresented populations' usages.").

[15] *Id.* (Newsom, J., concurring) ("Lawyers, judges, and would-be litigants might try to manipulate LLMs.").

[16] *Id.* (Newsom, J., concurring) ("

Judge Newsom's consideration of these issues is earnest and thoughtful, with each point warranting a detailed examination.

First, LLMs are indeed trained on an unfathomably large corpus of ordinary language[17]:

> GPT models are often trained in two stages. First, they are trained, using a large dataset of text from the Internet, to predict the next word. The models are then fine-tuned with additional data, using an algorithm called reinforcement learning from human feedback (RLHF), to produce outputs that are preferred by human labelers.[18]

As Judge Newsom explains, this broad training allows LLMs to make useful statistical predictions about ordinary words within an array of semantic contexts.[19] And as he cautions, out-of-the-box LLM knowledge—as reflected in their "zero-shot capabilities"—is constrained by information the LLM was exposed to during its training.[20] As a general matter, this is not particularly problematic if one considers that all leading LLMs have been trained on far more language than any human (or

---

Would the consideration of LLM outputs in interpreting legal texts inevitably put us on some dystopian path toward 'robo judges' algorithmically resolving human disputes?").

[17] *Compare id.* at 1226 (Newsom, J., concurring) ("The ordinary-meaning rule's foundation in the common speech of common people matters here because LLMs are quite literally 'taught' using data that aim to reflect and capture how individuals use language in their everyday lives. . . . LLMs train on a mind-bogglingly enormous amount of raw data taken from the internet—GPT-3.5 Turbo, for example, trained on between 400 and 500 billion words." (citing Christoph Engel & Richard H. McAdams, *Asking GPT for the Ordinary Meaning of Statutory Terms* 10–11 (Max Planck Inst. Discussion Paper 2024/5)), *with* ETHAN MOLLICK, CO-INTELLIGENCE: LIVING AND WORKING WITH AI 11 (2024) ("To teach AI how to understand and generate humanlike writing, it is trained on a massive amount of text from various sources, such as websites, books, and other digital documents.").

[18] OpenAI, *GPT-4 Technical Report*, OPENAI (Dec. 19, 2023), https://arxiv.org/abs/2303.08774v4.

[19] *Compare Snell,* 102 F.4th at 1226 (Newsom, J., concurring) ("Because they cast their nets so widely, LLMs can provide useful statistical predictions about how, in the main, ordinary people ordinarily use words and phrases in ordinary life."), *with* Daniel E. Pinnington & Reid F. Trautz, *Future Proofing: Using ChatGPT for Research and Writing*, 49 NO. 5 LAW PRAC. 38 (2023) ("LLMs spot patterns in how words, phrases, sentences and even paragraphs relate to each other, and then makes statistical predictions about what words should come next.").

[20] *Compare Snell,* 102 F.4th at 1226 (Newsom, J., concurring) ("To be sure LLMs' training data aren't a perfect universe from which to draw hard-and-fast conclusions about ordinary meaning, principally because they don't capture what I'll call 'pure offline' usages[] . . . ."), *with* Xingen Wan et al., *Universal Self-Adaptive Prompting for Inference and Few-Shot Learning*, GOOGLE (Oct. 21, 2023), https://arxiv.org/abs/2305.14926v2 ("While the zero-shot performances of LLMs are highly valued and widely used as a key yardstick of LLM capabilities, LLMs still often show weaker performances and/or larger performance fluctuations in the zero-shot setting because of the lack of guidance or readily-available template solutions.").

group of humans).[21] Therefore, abstracting the "ordinary meaning" of any word or phrase should not be a particular challenge for a state-of-the-art LLM.[22]

However, we acknowledge that the training set or the ability to write a very specific prompt could be a significant limitation when considering certain phrases or special "terms of art."[23] For example, less capable language models may not be able to understand specific phrases in maritime industry (e.g., berth, tender, deck, beam) without confusing those words with their general or more common meaning in the English vernacular.[24] However, these limitations can be solved through more specific prompt writing (e.g., "Explain for me what the word 'tender' means" versus "Explain for me what the word 'tender' means in the maritime industry").[25] In addition, specific models can also be trained to have a deeper and more specialized understanding of these specific industries or situations.[26] But foundation models, like GPT-4, take a

---

[21] MUSTAFA SULEYMAN WITH MICHAEL BHASKAR, THE COMING WAVE: TECHNOLOGY, POWER, AND THE TWENTY-FIRST CENTURY'S GREATEST DILEMMA 65 (2023) ("Today's LLMs are trained on trillions of words. Imagine digesting Wikipedia wholesale, consuming all the subtitles and comments on YouTube, reading millions of legal contracts, tens of millions of emails, and hundreds of thousands of books. This kind of vast, almost instantaneous consumption of information is not just difficult to comprehend; it's truly alien.").

[22] *Id.* at 84–85 ("LLMs take advantage of the fact that language data comes in a sequential order. Each unit of information is in some way related to data earlier in a series. The model reads very large numbers of sentences, learns an abstract representation of the information contained within them, and then, based on this, generates a prediction about what should come next. . . . The model would likely create tokens for the letters 'the' in the word 'there' and 'ing' in the word 'going,' since they commonly occur in other words. When parsing the full sentence, it would learn that 'storm,' 'tomorrow,' and 'Brazil' are the key features, inferring that Brazil is a place, that a storm will be happening in the future, and so on. Based on this, it then suggests which tokens should come next in the sequence, what output logically follows the input.").

[23] Suriya Gunasekar et al., *Textbooks Are All You Need*, MICROSOFT RESEARCH (Oct. 2, 2023), https://arxiv.org/abs/2306.11644v2 ("Sensitivity to prompt variations. Our model is sensitive to various perturbations of prompts. First, its performance drops significantly as the length of the prompt increases, as it tends to ignore, forget or misinterpret parts of the prompt when it is too long.").

[24] *Cf.* Ronen Eldan & Yuanzhi Li, *TinyStories: How Small Can Language Models Be and Still Speak Coherent English?*, MICROSOFT RESEARCH (May 24, 2023), https://arxiv.org/abs/2305.07759v2 ("[S]mall language models (SLMs) are very limited in their performance and capabilities, especially in text generation tasks. For example, models with around 125M parameters such as GPT-Neo (small) or GPT-2 (small) can rarely generate any consistent text beyond a few words even after extensive training on large corpora.").

[25] Carolyn V. Williams, *Bracing for Impact: Revising Legal Writing Assessments Ahead of the Collision of Generative AI and the NextGen Bar Exam*, 28 LEGAL WRITING: J. LEGAL WRITING INST. 1, 69 (2024) ("Once students saw for themselves the limitations of ChatGPT, professors could incorporate training and assessment on prompt engineering into legal writing courses. Students could be trained to specify the tone of the responses they want generated, to specify the length of responses, and to instruct ChatGPT not to fabricate legal sources. They could learn how to improve ChatGPT's responses by using chain-of-thought prompts.").

[26] *See, e.g.*, Shijie Wu et al., *BloombergGPT: A Large Language Model for Finance*, BLOOMBERG (Dec. 21, 2023), https://arxiv.org/pdf/2303.17564v3 ("We train BloombergGPT, a 50 billion parameter language model that supports a wide range of tasks within the financial industry.").

different approach, beginning with training on a broad spectrum of information that primes these models for wide-ranging utility in downstream tasks.[27] Think of it like this: just as elementary and secondary education provides children with a basic understanding of concepts that they can later abstract and reason from as they navigate into higher education (and then into more specialized roles in society),[28] the capabilities of the current generation of LLMs should be thought of as a baseline—just waiting to be enhanced with additional data and tools.[29] For instance, OpenAI provides mechanisms for users to easily equip their LLMs with any data of the user's choice.[30] Grounding an LLM's output in specified, post-training data like this is called "Retrieval-Augmented Generation," which has "emerged as a promising solution to alleviate [LLM]'s deficiency in [certain] knowledge."[31] Once an LLM is linked to one of these databases—which OpenAI refers to as vector stores—the model can be instructed to conduct a semantic search of the information available to and only respond to user queries based on information contained therein.[32]

Second, LLMs can "hallucinate." This is a phrase meaning that the LLM simply gets it wrong and returns a result that is incorrect because of an inaccurate statistical correlation in the words or phrases and results in a response that is fully or partially wrong.[33] These hallucinations—the bête noir of LLM adoption among

---

[27] Yuwei Fang et al., *i-Code Studio: A Configurable and Composable Framework for Integrative AI*, MICROSOFT COGNITIVE SERVS. RESEARCH GRP. (May 23, 2023), https://arxiv.org/pdf/2305.13738v1 ("Foundation models, first introduced by Bommasani et al. (2021), refer to any model that is pre-trained on broad data at scale and can be adapted to a wide range of downstream tasks.").

[28] *See* Smitha Milli, Pieter Abbeel & Igor Mordatch, *Interpretable and Pedagogical Examples*, OPENAI (Feb. 14, 2018), https://arxiv.org/pdf/1711.00694 ("For example, suppose a teacher is trying to teach different types of animals to a student. To teach what a 'dog' is they would not show the student only images of dalmatians. Instead, they would show different types of dogs, so the student generalizes the word 'dog' to all types of dogs, rather than merely dalmatians.").

[29] *See, e.g.*, Amy Winograd, *Loose-Lipped Large Language Models Spill Your Secrets: The Privacy Implications of Large Language Models*, 36 HARV. J.L. & TECH. 615, 617 (2023) ("LLMs' impressive baseline proficiency can be further enhanced through fine-tuning.").

[30] *File Search*, OpenAI, https://platform.openai.com/docs/assistants/tools/file-search (last visited June 11, 2024) ("File Search augments the Assistant with knowledge from outside its model, such as proprietary product information or documents provided by your users. OpenAI automatically parses and chunks your documents, creates and stores the embeddings, and use [sic] both vector and keyword search to retrieve relevant content to answer user queries.").

[31] Xiao Yang et al., *CRAG – Comprehensive RAG Benchmark*, META (June 7, 2024), https://arxiv.org/abs/2406.04744v1.

[32] *File Search*, OPENAI, https://platform.openai.com/docs/assistants/tools/file-search (last visited June 11, 2024) ("File Search augments the Assistant with knowledge from outside its model, such as proprietary product information or documents provided by your users. OpenAI automatically parses and chunks your documents, creates and stores the embeddings, and use [sic] both vector and keyword search to retrieve relevant content to answer user queries.").

[33] Gary Marchant & Whitney Stefko Dover, *Law Practice Disruptors: Large Language Models and Artificial Intelligence*, 60 ARIZ. ATT'Y 26 (2023) ("LLMs can sometimes 'hallucinate' or generate text that is not grounded in reality. This can happen because the models are designed to generate new combinations of words, rather than to understand the content they are producing.").

legal professionals—are not the showstoppers that many might believe them to be.[34] Judge Newsom notes that many lawyers hallucinate (i.e., assert incorrect statements of fact and law) without the assistance of LLMs.[35] "So at worst, the 'hallucination' problem counsels against blind-faith reliance on LLM outputs—in exactly the same way that no conscientious judge would blind-faith rely on a lawyer's representations."[36] Furthermore, it is well-documented that certain user and developer techniques—such as providing models with grounding data—markedly reduce instances of hallucination.[37] We have seen a rapid decline in hallucinations since the initial debut of ChatGPT, which initially leveraged OpenAI's GPT-3.5 model in lieu of its more capable GPT-4 model.[38] And it is worth remembering that the LLMs that exist today are the worst LLMs that will ever exist.[39]

---

[34] Although we would all prefer colleagues (humans and A.I.) free from error, some hallucinations can even be useful in the practice of law. Consider, for instance, a litigation associate who creates a custom GPT to assist her in preparing for her first oral arguments. In mere seconds, she configures this custom GPT to loosely approximate her actual judge—perhaps instilling it with behavioral instructions that reflect his Ballotpedia webpage and loading its vector store with all publicly available filings from the docket. (Of course, she is using her firm's OpenAI Enterprise or Teams account—so her custom GPT remains siloed to herself or her firm, and her interactions with it cannot be used to train current or future models.) The underlying model for this custom GPT is OpenAI's recently announced GPT-4o (the multimodal successor to last year's GPT-4), so the associate has at her fingertips a tireless moot-court judge who can hear, see, and speak to her in real-time as she nervously practices for hours in her office. In that time, the custom GPT asks the associate certain questions—it also makes certain factual and legal assertions—that are based on its misunderstandings of relevant facts and applicable law (as reflected in the filings it has been provided). Because the associate bears in mind that she is interacting with an A.I. model that is meant to approximate fallible human intelligence—as opposed to an omniscient, omnipotent deus ex machina—she seizes on these opportunities to sharpen her own understanding of the underlying material and practice elegant reeducation and redirection techniques.

[35] *Snell*, 102 F.4th at 1230–1231 (Newsom, J., concurring) ("Flesh-and-blood lawyers hallucinate too. Sometimes, their hallucinations are good-faith mistakes. But all too often, I'm afraid, they're quite intentional—in their zeal, attorneys sometimes shade facts, finesse (and even omit altogether) adverse authorities, etc.").

[36] *Id.*

[37] *See, e.g.*, Zane Durante et al., *Agent AI: Surveying the Horizons of Multimodal Interaction*, MICROSOFT RESEARCH (Jan. 25, 2024), https://arxiv.org/abs/2401.03568v2 ("Some promising routes for reducing the rate of hallucination in language generation involve using retrieval-augmented generation (Lewis et al., 2020; Shuster et al., 2021) or other methods for grounding natural language outputs via external knowledge retrieval.").

[38] OpenAI, *GPT-4 Technical Report*, OPENAI (Dec. 19, 2023), https://arxiv.org/abs/2303.08774v4 ("GPT-4 was trained to reduce the model's tendency to hallucinate by leveraging data from prior models such as ChatGPT. On internal evaluations, GPT-4-launch scores 19 percentage points higher than our latest GPT-3.5 model at avoiding open-domain hallucinations, and 29 percentage points higher at avoiding closed-domain hallucinations."); *see also* ETHAN MOLLICK, CO-INTELLIGENCE 98 (2024) ("As models advance, hallucination rates are dropping over time. For example, a study examining the number of hallucinations and errors in citations given by AI found that GPT-3.5 made mistakes in 98 percent of the cites, but GPT-4 hallucinated only 20 percent of the time.").

[39] Credit to Alya Sulaiman, a partner at McDermott Will & Emery, for pointing out and succinctly summarizing this reality.

Third, LLMs are indeed accessible—to judges, to attorneys, to their staffs, and most importantly, to ordinary citizens.[40] As Judge Newsom explains, the widespread availability of LLMs democratizes the interpretive process by leveraging inputs from ordinary people and providing an inexpensive research tool.[41] And as the advanced capabilities of current-generation LLMs have begun converging at similar benchmark scores, companies have begun wooing developers and users to their models with race-to-the-bottom pricing.[42]

Fourth, LLMs can serve the values of transparency and reliability in legal processes.[43] Judge Newsom anchors his view on this by describing longstanding critiques of the textualist enterprise: Cherry-picking definitions from multiple

---

[40] *Compare Snell*, 102 F.4th at 1228 (Newsom, J., concurring) ("LLMs are readily accessible (and increasingly so) to judges, lawyers, and, perhaps most importantly, ordinary citizens. In recent years, the use of LLMs has proliferated, and as with all other internet-related tools, one can only assume that usage will continue to accelerate, likely at an exponential rate."), *with* Thomas Lurie & Matin Fallahi, *Rise of the Machines: Implications of Artificial Intelligence in the Legal Field,* 65 FOR THE DEF., 8, 11 (2023) ("Individuals who otherwise could not afford an attorney, or who simply believe they do not need to retain an attorney, can use LLMs to craft arguments in legalese with a few strokes of the keyboard."), *and* Damien Riehl, *We Need to Talk About ChatGPT: A Lawyer's Introduction to the Exploding Field of AI and Large Language Models*, 80 BENCH & B. MINN., 26, 29 (2023) ("By reducing legal costs and increasing efficiency, LLMs have the potential to improve access to justice for individuals and organizations.").

[41] *Compare Snell*, 102 F.4th at 1228 (Newsom, J., concurring) ("First, it offers the promise of 'democratizing' the interpretive enterprise, both (as already explained) by leveraging inputs *from* ordinary people and by being available for use *by* ordinary people. Second, it provides judges, lawyers, and litigants an inexpensive research tool."), *with* Riehl, *supra* note 40, at 28 ("Unquestionably, LLMs' costs are far, far lower than employing human lawyers: A GPT-4 prompt costs a fraction of a penny.").

[42] *See, e.g.*, SULEYMAN WITH BHASKAR, *supra* note 21, at 91 ("AI researchers are racing to reduce costs and drive up performance so that these models can be used in all sorts of production settings. In the last four years, the costs and time needed to train advanced language models have collapsed."); Leopold Aschenbrenner, *Situational Awareness: The Decade Ahead* at 24 (2024) ("GPT-4, on release, cost ~the same as GPT-3 when it was released, despite the absolutely enormous performance increase. . . . Since the GPT-4 release a year ago, OpenAI prices for GPT-4-level models have fallen another 6x/4x (input/output) with the release of GPT-4o. . . . Though these are inference efficiencies (rather than necessarily training efficiencies), and to some extent will reflect inference-specific optimizations, a) they suggest enormous amounts of algorithmic progress is possible and happening in general, and b) it's often the case that . . . algorithmic improvements is both a training efficiency gain and an inference efficiency, for example by reducing the number of parameters necessary.").

[43] *Compare Snell*, 102 F.4th at 1228 (Newsom, J., concurring) ("Using LLMs to facilitate ordinary-meaning interpretation may actually enhance the transparency and reliability of the interpretive enterprise itself, at least vis-à-vis current practice."), *with* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1206 (2017) ("Transparency requires that agencies ensure that their decisions are 'clearly articulated' and 'the rationales for these decisions are fully explained, and the evidence on which the decisions are based is publicly accessible."), *and* Catrina Denvir et al., *The Devil in the Detail: Mitigating the Constitutional & Rule of Law Risks Associated with the Use of Artificial Intelligence in the Legal Domain*, 47 FLA. ST. U. L. REV. 29, 76 (2019) ("The transparency of legal decision-making plays a crucial role in shaping public perceptions of the fairness of the justice system.").

dictionaries, whose editors' compilation processes are themselves often opaque.[44] In contrast, Judge Newsom notes that LLMs could provide a more transparent process if users disclose their queries and the models' answers, which can be replicated and scrutinized.[45] Stated differently, if a dictionary is an accepted form of legal interpretation, then LLMs are even better at assessing a word or phrase's ordinary meaning: they can be, in many ways, a dictionary of the populous, a combination of the collective human knowledge or understanding of a word or phrase.[46] This is far more democratic than a single editor of a dictionary. In addition, we would add that if future litigants begin using retrieval-augmented generation, disclosure of the data placed in vector stores attached to the LLM used can provide further transparency into the interpretive enterprise (and make available to the model industry- or context-specific information that can assist users and models in gleaning the meaning of specific terms and phrases).[47]

Fifth, lawyers and judges will of course impact LLM outputs over the course of their use—just as they impact the outputs of other interpretive methodologies.[48] As

---

[44] *See Snell*, 102 F.4th at 1229 (Newsom, J., concurring) ("The cynic, of course, will insist that the judge just dictionary-shopped for the definitions that would enable him to reverse-engineer his preferred outcome. . . . Justice Scalia and Bryan Garner famously warned against 'an uncritical approach to dictionaries.' They highlighted as risks, for instance, that a volume could 'have been hastily put together by two editors on short notice, and very much on the cheap,' and that without 'consult[ing] the prefatory material' one might not be able 'to understand the principles on which the dictionary [was] assembled' or the 'ordering of [the] senses' of a particular term.").

[45] *See id.* ("Contrast my M.O. in this case, which I would recommend as a best practice: full disclosure of both the queries put to the LLMs (imperfect as mine might have been) and the models' answers.").

[46] *See* Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics*, 68 U. TORONTO L.J. 12, 18 (2018) ("Language usage implies connecting a system of signs with whatever these signs stand for, refer to, or perform.").

[47] *Cf. File Search*, OPENAI, https://platform.openai.com/docs/assistants/tools/file-search (last visited June 10, 2024) ("Retrieval augments the Assistant with knowledge from outside its model, such as proprietary product information or documents provided by your users. Once a file is uploaded and passed to the Assistant, OpenAI will automatically chunk your documents, index and store the embeddings, and implement vector search to retrieve relevant content to answer user queries.").

[48] *Compare Snell*, 102 F.4th at 1231 (Newsom, J., concurring) ("Lawyers, judges, and would-be litigants might try to manipulate LLMs. I suppose there's a risk that lawyers and judges might try to use LLMs strategically to reverse-engineer a preferred answer—say, by shopping around among the available models or manipulating queries."), *with* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, CTR. FOR RSCH. ON FOUND. MODELS, STANFORD INST. FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE & STANFORD UNIV. (July 12, 2022) ("[E]ach task can be specified through many possible prompts and slight variations in prompts can result in meaningful changes of model behavior."), https://arxiv.org/abs/2108.07258, *and* Subigya Nepal et al., *From User Surveys to Telemetry-Driven Agents: Exploring the Potential of Personalized Productivity Solutions*, MICROSOFT RESEARCH (Jan. 17, 2024) ("It is essential to acknowledge the potential influence of using different prompts on the generated outputs by GPT-4. As the AI model relies heavily on the input prompts provided, variations in these prompts can lead to dramatically different results."), https://arxiv.org/abs/2401.08960, *and* Paul Thomas et al., *Large Language Models Can Accurately Predict Searcher Preferences*, MICROSOFT RESEARCH (Sept. 19, 2023) ("We have seen that LLM

Fred Rodell observed almost a century ago: "None of The Law's answers to problems is preordained, precise, or inevitable[.] . . . Even if The Law will be considered a big machine that gives automatic answers to legally-worded questions, it is the lawyers and the lawyer-judges who phrase the questions and decide which buttons to push."[49] This observation doesn't necessarily jibe with legal reasoning as reflected on the written page, but it is consistent with some other frank accounts of the judicial process.[50] But as Judge Newsom's concurrence illustrates, LLMs can vividly relay how the plain-meaning of text is cogitated both within and outside of chambers.[51]

Finally, Judge Newsom is correct to dismiss the dystopian view that "robo judges" will come to algorithmically reign over human disputes.[52] As Chief Justice Roberts has stated, the law will always require human judgment to navigate gray areas[53]—a position with which former OpenAI executive and AI futurist Zack Kass

---

performance varies considerably as the prompt is varied, even when the task and the input data are fixed.")., https://arxiv.org/abs/2309.10621v1.

[49] FRED RODELL, WOE UNTO YOU, LAWYERS 157–59 (1939).

[50] JEROME FRANK, LAW & THE MODERN MIND 11–12, 21, 101 & nn.126, 137–38, 143–47 (1930) ("Now and again some judge, more clear-witted and outspoken than his fellows, describes (when off the bench) his methods in more homely terms. Recently Judge Hutcheson essay such an honest report of the judicial process. He tells us that after canvassing all the available material at his command and duly cogitating on it, he gives him imagination play, 'and brooding over the cause, waits for the feeling, the hunch—that intuitive flash of understanding that makes the jump-spark connection between question and decision and at the point where the path is darkest for the judicial feet, sets its light along the way. In feeling or "hunching" out his decisions, the judge acts not differently from but precisely as the lawyers do in working on their cases, with only this exception, that the lawyer, in having a predetermined destination in view—to win the lawsuit for his client—looks for an[d] regards only those hunches which keep him in the path that he has chosen, while the judge, being merely on his way with a roving commission to find the just solution, will follow his hunch wherever it leads him.'" (quoting Joseph C. Hutcheson, Jr., *The Judgment Intuitive: The Function of the "Hunch" in Judicial Decision*, 14 CORNELL L. REV. 274 (1929))).

[51] *See, e.g.*, *Snell*, 102 F.4th at 1224, 1234 (Newsom, J., concurring) ("I'll confess that the photos affected the way I thought about the case. Nothing in them really struck me as particularly "landscaping"-y. The problem, of course, was that I couldn't articulate *why*. And visceral, gut-instinct decisionmaking has always given me the willies. . . . Below are the questions that I asked OpenAI's ChatGPT and Google's Bard models, along with their responses.").

[52] *Compare Snell*, 102 F.4th at 1232 (Newsom, J., concurring) ("Would the consideration of LLM outputs in interpreting legal texts inevitably put us on some dystopian path toward 'robo judges' algorithmically resolving human disputes? I don't think so."), *with* Ray Worthy Campbell, *Artificial Intelligence in the Courtroom: The Delivery of Justice in the Age of Machine Learning*, 18 COLO. TECH. L.J. 323, 324 (2020) ("[W]hat we can expect in the near term is for software to play a role supporting—but not replacing—human judges. AI can, in certain cases, predict how a certain case might come out. That, however, falls far short of what judges do. The current capability of AI is limited to specialized tasks, and the roles of judges are so generalized that there is no near-term possibility of AI wholly and satisfactorily displacing judges in high stakes cases.").

[53] *See Snell*, 102 F.4th at 1232 (Newsom, J., concurring) ("As Chief Justice Roberts recently observed, the law will always require 'gray area[]' decisionmaking that entails the 'application of human judgment.'").

seems to agree.[54] Even as he proselytizes the sea change in capabilities that current- and next-generation LLMs will bring to every organization, Kass has observed that artificial intelligence remains incapable "of doing the things that require wisdom, courage, vision, empathy, and curiosity; these immutable human qualities."[55] Those just happen to be qualities strongly associated with the American judiciary and every member of the bar.[56]

* * *

Since the publication of *Snell*, Judge Newsom's dialogue on this theme was continued in his *United States v. Deleon* concurrence.[57] In *Deleon*, Judge Newsom ran repeated queries to ChatGPT, Claude, and Gemini about the phrase "physically restrained," observing that "the very thing that had initially given me pause—namely, that the LLMs were returning subtly different responses to the same question—has instead given me (more) hope that the models have something significant to offer the interpretive enterprise."[58] In particular, although each inquiry generated "slightly (though again, not substantively) different answers," he emphasized that they consistently coalesced "around a common core" and "there was an objectively verifiable throughline."[59] Ultimately, this reaffirms that "some marginal uncertainty is inherent in the assessment of ordinary meaning," while also underscoring "the models' utility in the ordinary-meaning analysis."[60]

We leave it to another day—to other capable minds—to continue the dialogue on the contours of legitimately leveraging large language models in the interpretation

---

[54] Shelley E. Kohan, *Zack Kass, Former OpenAI Leader, Urges Retail Execs To Champion AI*, FORBES (May 2, 2024), https://www.forbes.com/sites/shelleykohan/2024/05/02/zack-kass-former-openai-leader-urges-retail-execs-to-champion-ai/ ("[M]ost people's jobs are going to reduce to the parts of the job that are the most humanistic because everything that's computational will probably get automated.").

[55] *Id.*

[56] *See, e.g.*, Muharrem Kılıç, *Ethical-Juridical Inquiry Regarding the Effect of Artificial Intelligence Applications on Legal Profession and Legal Practices*, 14 J. MARSHALL L.J. 268, 277 (2021) ("Judicial virtues have three dimensions: judicial intelligence, judicial integrity, and judicial wisdom. The first of these virtues are [sic] aimed at competence in understanding and theorizing law. According to the second virtue, the 'good judge' has to carry a personal concern at the point of loyalty to the rule of law and the consistency of the law. As per the third virtue, the 'good judge' must have 'practical wisdom' in choosing the purposes and appropriate means of law."); Allison C. Shields, *Emotional Intelligence versus Artificial Intelligence*, 45 LAW PRAC. 14, 15 (2019) ("Lawyers who have the ability to make a human connection and understand the complex relationships and emotions involved in a legal case or transaction, who demonstrate wisdom and good judgment, have good instincts, intuition, people and communication skills, and can establish rapport with clients and motivate others will be in high demand.").

[57] *See* United States v. Deleon, 116 F.4th 1260, 1276 (11th Cir. 2024) (Newsom, J., concurring) ("Given the claims that I've made about LLMs' utility to the interpretive enterprise, shouldn't a little variation have been exactly what I expected?").

[58] *Id.* at 1276–77.

[59] *Id.* at 1275.

[60] *Id.* at 1277.

of legal texts, just as other capable minds have long debated the proper use of other barometers of interpretation. But it is a foregone conclusion that large language models are exceptionally capable of contributing to this process, and those capabilities continue to climb.