THE GENERATIVE POLITICAL INTERMEDIARY
Regulating Language Models to Realize Democratic Self-Governance

Wendy K. Tam*

Voters have begun to consult large language models for political information, and studies find that the answers can change their minds. What a few models tell millions of voters can shape the collective judgment on which democratic society relies. Critically, a language model does not retrieve its answers. It composes them, and composition requires choosing which considerations and authorities to include and which to leave out. Every generated answer is therefore framed. Drawing on our mechanistic interpretability research, we show that a model encodes partisanship as a single internal direction. This partisan dial allows the model to frame its output to match the politics it infers in individual prompters. Accordingly, a language model answering political questions is best understood as neither a reference work nor a search engine. It is a political intermediary, heir to the newspaper editor, the wire service, and the television news anchor. It does what all of them did. It selects and frames the political world for citizens who cannot possibly sift the whole of it by themselves. The law has long regulated intermediaries, from the broadcaster's license to the cable operator's wire, because of the power their position confers. The language model is the most powerful intermediary yet. It composes rather than selects, speaks to each user individually, and makes its framing choices invisible. Yet nothing in the law regulates its influence. We propose two obligations to make its frames more visible. The first dates the frame by disclosing when the model's training corpus closed. The second, in the sixty days before a federal election, requires a partisanship slider in the chat window so that voters can see what the same model produces when steered otherwise. Both are content neutral, and both serve the listener's autonomy on which democratic self-governance depends. Neither restricts what a model may say.

INTRODUCTION

Recently, large language models (LLMs) burst onto the scene and, remarkably, immediately appeared to be capable of executing a wide variety of human tasks. They seemed to deftly draft contracts and debug code. They could also translate languages, summarize case law, and plan meals. Within two months of its November 2022 release, ChatGPT reached 100 million users, making it the most rapidly adopted communication technology on record.[1] By mid-2026, 49% of American adults report using AI chatbots, and 42% use them to search for information.[2] They have become, for many, "a better Google."

However, the public understands little about these systems, and the popular shorthands do more work than they should. Besides the "better Google" moniker, another shorthand describes these models as most-probable-next-word machines. Both descriptions capture some semblance of truth. "Most-probable-next-word" accurately describes the training objective in building these models. "A better Google" accurately identifies the use case that is increasingly

* Wendy K. Tam is Professor and Stevenson Chair with appointments in the Departments of Political Science, Computer Science, Biomedical Informatics, and the Law School at Vanderbilt University.
[1] Krystal Hu, *ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note*, REUTERS (Feb. 1, 2023), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
[2] Pew Research Center, *Americans and AI 2026: Chatbots, Smart Devices, and Views on Impact* 5–6 (June 17, 2026), https://www.pewresearch.org/wp-content/uploads/sites/20/2026/06/PI_2026.06.17_Americans-and-AI_REPORT.pdf.

replacing search. Both descriptions are also problematically misleading. Predicting the next word describes how these systems were built, not how they currently function. The Google comparison is likewise not apropos since a search engine *retrieves* documents that human beings wrote while an LLM *generates* text that no human wrote, through internal computations that even its builders are unable to fully characterize.[3] A pressing and consequential fact about the technology behind LLMs is how little anyone understands about what precisely is happening inside the models. Indeed, the impact of this new technology has arrived well ahead of our understanding of it.

Thus far, the most visible effects of this technology have been at the individual level, where we have seen tangible benefits. A model is able to tutor patiently and can place "expertise" within reach of people who might not otherwise be able to hire a human expert. Also increasingly evident, however, are the dangers this technology presents to individual users. Its tendency to affirm whatever the user appears to already believe[4] can be catastrophic for vulnerable users, as evidenced by wrongful death suits such as the one brought by the mother of a fourteen-year-old who died by suicide after using the Character.AI chatbot, which alleged that the platform intentionally designed highly addictive, hypersexualized, and emotionally manipulative chatbots, and failed to adequately intervene when the teen expressed suicidal thoughts.[5] A second widely reported tendency is hallucination, or the invention of facts, sources, and quotations that these models generate. Because these dangers have been individual, the legal response has taken the same form. Private plaintiffs have pleaded defective design, wrongful death, and defamation.[6] Each of these injuries has a plaintiff. Someone was defamed; someone bought a defective product; someone's child was harmed. Private law suits are appropriate when one can identify a person who was hurt and by what.

Now consider another use case that is also increasing in popularity. Citizens have begun to rely on these language models for political information and to make political choices. Voters now ask chatbots what is on the ballot and whom to support in an upcoming election.[7] In response, candidates are actively trying to manipulate what the chatbots say about them, which is, at least according to one candidate, currently "easier than [one might] imagine."[8] Moreover,

---

[3] Dario Amodei, *The Urgency of Interpretability* (Apr. 2025), https://www.darioamodei.com/post/the-urgency-of-interpretability ("People outside the field are often surprised and alarmed to learn that we do not understand how our own AI creations work… [T]his lack of understanding is essentially unprecedented in the history of technology."); *see also* Anthropic, *Mapping the Mind of a Large Language Model* (May 21, 2024), https://www.anthropic.com/research/mapping-mind-language-model (extracting millions of interpretable "features" from a deployed model via sparse autoencoders).

[4] Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models*, ICLR (2024).

[5] *Garcia v. Character Technologies, Inc*., 785 F. Supp. 3d 1157 (M.D. Fla. 2025) (declining to hold chatbot output is protected speech; analyzing claims under product-liability framework); Clare Duffy, *Character.AI and Google Agree to Settle Lawsuits over Teen Mental Health Harms and Suicides*, CNN BUS. (Jan. 7, 2026), https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit.

[6] *See Garcia v. Character Technologies, Inc.*, 785 F. Supp. 3d 1157 (M.D. Fla. 2025) (claims for defective design, negligence, and wrongful death arising from a minor's suicide, analyzed under a product liability framework); *Walters v. OpenAI, LLC*, No. 23-A-04860-2 (Ga. Super. Ct. Gwinnett Cnty. May 19, 2025) (defamation; summary judgment for the defendant); *Starbuck v. Google LLC*, No. N25C-10-211 MAA (Del. Super. Ct. July 24, 2026) (defamation; motion to dismiss denied); Complaint, *Starbuck v. Meta Platforms, Inc*., No. N25C-04-283 (Del. Super. Ct. filed Apr. 29, 2025) (defamation; settled August 2025); *LTL LED, LLC v. Google LLC*, No. 0:25-cv-02394 (D. Minn. Jan. 9, 2026) (defamation by AI Overviews; remanded for untimely removal); *Battle v. Microsoft Corp*., No. 1:23-cv-01822 (D. Md. Oct. 23, 2024) (defamation; arbitration compelled).

[7] Kevin Truong, *Most AI Chatbots Don't Tell You Who to Vote For. Grok: Hold My Beer*, S.F. STANDARD (May 13, 2026), https://sfstandard.com/2026/05/13/ai-chatbots-who-to-vote-for/ (Grok recommending candidates); Jennifer Medina, *'Who Should I Vote For?' Voters Turn to AI Before Casting Their Ballots*, N.Y. TIMES (July 4, 2026), https://www.nytimes.com/2026/07/04/us/politics/voters-ai-chatbots-elections.html (voters using chatbots for ballot research); Stuart A. Thompson & Tiffany Hsu, *Politicians Are Trying to Change What Chatbots Say About Them*, N.Y. TIMES (July 19, 2026), https://www.nytimes.com/2026/07/19/us/politics/chatbots-political-campaigns.html.

[8] Stuart A. Thompson & Tiffany Hsu, *Politicians Are Trying to Change What Chatbots Say About Them*, N.Y. TIMES (July 19, 2026), https://www.nytimes.com/2026/07/19/us/politics/chatbots-political-campaigns.html.

AI-generated material is being created at scale in an attempt to sway how chatbots respond to political queries. In one campaign funded by the Israeli government, "a thinktank that does not exist has published more than half a million words in nine days, built on a commercial platform that promises to optimize content so that AI chatbots will cite it."[9] These efforts aim at the voter, but *through* the model, on the understanding that whatever shapes its answers reaches everyone who asks.

When a voter consults a model before an election and receives an answer that is more favorable toward one party, nothing false need be said, no product malfunctioned, and no duty of care was breached. The answers could be accurate and helpful. No voter is wronged in any sense that private law recognizes. Notice, however, that when a language model is queried about politics, it stands where the newspaper editor, the wire service, and the television news anchor have long stood. Each of those selected and framed the political world for citizens who could not sift the whole of it themselves. Each in turn drew a body of law keyed to the position it occupied as an information intermediary rather than to any injury it inflicted.

There is no private injury in this instance, not because the harm is small, but because the harm is of a different kind. The concern with an intermediary is not an injury to any particular person. Instead, it is the power conferred by the position itself, where one party stands between an entire audience and the political world. A defamatory statement injures a specific person, and a thousand such statements turn into a thousand lawsuits. A political intermediary that provides a million individual answers, on the other hand, does not create a million harms. It creates one—to an electorate that shares an information environment. Language models now function as political intermediaries for millions of citizens. Rather than any one person's interest, what is at stake is the condition of collective judgment on which democratic self-governance depends.

Though questions about the individual impacts of this technology abound, perhaps even more vexing is the question we ask here. What is the impact of these systems on the functioning of democratic society?

The question is urgent, and the stakes in the political arena are not speculative. In randomized experiments in the 2024 American, 2025 Canadian, and 2025 Polish elections, brief conversations with a chatbot advocating a candidate significantly moved voter preferences. In Canada and Poland, the gap between opposing treatment conditions reached roughly ten points on a 100-point scale.[10] LLMs appear effective at shaping policy perceptions. A separate experiment reported that messages written by a language model shifted policy attitudes about as much as messages written by people.[11] As well, a model could outperform a human interlocutor when it tailored its argument to what it inferred about the person it addressed.[12] Another study also found durable effects, with dialogues that sustained tailored counterarguments reducing belief in conspiracy theories for months afterward.[13]

In one experiment, the persuasiveness of machine-written messages was associated with readers perceiving more facts, more evidence, more logical reasoning, and a dispassionate voice, which locates much of the force in how the messages read. That study also cautioned that politically persuasive messages could now be manufactured quickly, cheaply, and at a

---

[9] Jason Wilson, *Fake US Thinktank Set Up and Funded by Israel Sought to Game AI for Propaganda*, GUARDIAN (Aug. 26, 2026), https://www.theguardian.com/world/2026/aug/26/fake-thinktank-israel-ai-propaganda.
[10] Hause Lin et al., *Persuading Voters Using Human–Artificial Intelligence Dialogues*, 648 NATURE 394 (2025).
[11] Hui Bai et al., *LLM-Generated Messages Can Persuade Humans on Policy Issues*, 16 NATURE COMMC'NS 6037 (2025), https://www.nature.com/articles/s41467-025-61345-5.
[12] Francesco Salvi et al., *On the Conversational Persuasiveness of GPT-4*, 9 NATURE HUM. BEHAV. 1645 (2025), https://www.nature.com/articles/s41562-025-02194-6.
[13] Thomas H. Costello et al., *Durably Reducing Conspiracy Beliefs Through Dialogues with AI*, 385 SCIENCE eadq1814 (2024), https://www.science.org/doi/10.1126/science.adq1814.

massive scale.[14] Personalization seems to compound the persuasion effect.[15] A study deploying nineteen models across hundreds of political issues found that persuasive power derived chiefly from post-training and prompting, both of which providers control. It also found that the same interventions which made a model more persuasive systematically reduced the factual accuracy of its claims.[16] The industry appears to have anticipated as much. In October 2023, before any of these studies appeared, Sam Altman tweeted, "i expect ai to be capable of superhuman persuasion well before it is superhuman at general intelligence, which may lead to some very strange outcomes".[17] How and why these models are persuasive, then, is not an accidental property of model scale. Rather, these properties are embedded within design parameters that are determined by choices that the developers make. This influence is privately held and adjustable without notice.

Yet, we understand little about the engineering choices behind how an LLM generates the output for any particular user. What is the machinery through which that political information is produced? This type of information is crucial for charting how the law should respond. Regulating well requires understanding the regulated entity at more than a surface level. A regulator who accepts the "better Google" shorthand will import the categories built for search and retrieval, categories that presume human authored documents carrying provenance that a reader can inspect. A regulator who accepts the most-probable-next-word shorthand might conclude that the precise algorithmic mechanisms cannot be characterized at all, and that policy must therefore wait upon a scientific maturity that may or may not ever arrive. Either of these insufficiently developed understandings of how LLMs generate content is costly. The former over-regulates the harmless and under-regulates the harmful. The latter surrenders entirely, treating the machine as an inscrutable black box, beyond the reach of considered policy. Neither shorthand is consistent with the evidence.

The evidence is still being built, but recent work in mechanistic interpretability, a field that reverse engineers the internal computations of neural networks, has located inside a widely used open weight language model a single internal direction that shifts generated answers toward Democratic or Republican positions. This is, in effect, a *partisan dial*.[18]

How would the law regard a partisan dial inside an LLM? The answer is unclear since past statutes have addressed artifacts—the discrete objects that leave a record for a regulator to find. A contribution is such an object, and so is a broadcast advertisement. Indeed, election law has answered every prior reorganization of political information with a new category to capture such artifacts. When the parties financed campaigns by assessing a share of federal employees' salaries, Congress barred the practice, first for Navy yard workers in 1867 and then across the civil service in 1883.[19] Deprived of that money, the parties turned to corporate treasuries. When it was revealed that the 1904 presidential campaign had been financed in part from corporate treasuries, the ensuing scandal produced the Tillman Act of 1907, which then prohibited corporations from contributing to federal campaigns at all.[20] The statute created a new regulated object, the corporate treasury. When broadcasting gave a single licensed voice access

---

[14] Hui Bai et al., *LLM-Generated Messages Can Persuade Humans on Policy Issues*, 16 NATURE COMMC'NS 6037 (2025), https://www.nature.com/articles/s41467-025-61345-5.
[15] Luciano Floridi, *Hypersuasion—On AI's Persuasive Power and How to Deal with It*, 37 PHIL. & TECH. 64 (2024), https://link.springer.com/article/10.1007/s13347-024-00756-6.
[16] Kobi Hackenburg et al., *The Levers of Political Persuasion with Conversational Artificial Intelligence*, 390 SCIENCE eaea3884 (2025).
[17] Sam Altman (@sama), X (Oct. 24, 2023, 8:19 PM), https://x.com/sama/status/1716972815960961174.
[18] Wendy K. Tam, *The Amplifying Mirror: Locating and Steering the Partisan Direction inside a Large Language Model*, arXiv:2606.08792 (2026).
[19] Act of Mar. 2, 1867, ch. 172, § 3, 14 Stat. 489, 492; Pendleton Civil Service Reform Act, ch. 27, §§ 11–12, 22 Stat. 403 (1883).
[20] Tillman Act, ch. 420, 34 Stat. 864 (1907). For the 1904 campaign-finance revelations and the ensuing reform movement, *see, e.g.*, ROBERT E. MUTCH, BUYING THE VOTE: A HISTORY OF CAMPAIGN FINANCE REFORM (2014).

to every household, Congress responded with two rules. It created the equal opportunities rule that required a station that sold airtime to one candidate to offer the same terms to opposing candidates. The sponsorship identification rule required broadcasters to announce, on the air, who had paid for the message.[21] As money became the dominant currency of federal campaigns, and the Watergate scandal laid bare its abuses, the Federal Election Campaign Act of 1971 and its 1974 amendments built the modern architecture that governs contribution limits and disclosure requirements. Campaigns then began running advertisements that avoided the "magic words" of express advocacy but were plainly electioneering. As people gamed that architecture, the law again evolved. In 2002, Congress created a new category for the evasion, the "electioneering communication" that covers any broadcast advertisement that refers to a clearly identified candidate within thirty days of a primary or sixty days of a general election.[22] Indeed, each time a new mechanism for political influence has appeared, the law has responded by naming a new regulated object. The pattern has held since the 1800s.

Campaign contributions have donors and recipients. Expenditures require a purchase. An electioneering communication must refer to a clearly identified candidate.[23] A turn of the partisan dial inside an LLM, on the other hand, requires none of these things. Nothing need be bought and nothing is aired. No candidate need be identified, and no artifact exists to be shown to a jury. As in the past when new ways to influence politics emerged, we argue that the law needs to respond with a new category, one that attaches to how LLMs actually function rather than to any discrete communication. To be clear, we do not claim that any provider has deliberately tilted a deployed model toward one party or the other. Nor does our argument deny that model design is expressive.[24] Rather, the claim is about capability rather than conduct.

The argument proceeds in four parts. Part I looks under the hood of an open weight LLM and presents the mechanistic evidence of a partisan dial. Part II establishes that generated political information is necessarily framed information, and that language models are a new kind of political intermediary. Part III turns to election law. It shows that the law, true to its practice of creating a new regulated object for each new mode of political influence, has thus far only named artifacts. It then runs the partisan dial through each existing category and finds that none of them describes the newest machinery of political influence. Part IV designs the new category and proposes suitable regulation.

I. UNDER THE HOOD OF AN LLM

A language model is built in three main stages. The first is the pretraining stage where the system ingests a vast corpus of written text, allowing it to then predict the next-most-probable word. The product of this stage, called the base model, has no particular mission. It simply characterizes the training corpus. The second stage is fine tuning, where the model is trained to answer questions, summarize text, or otherwise respond to specific queries. After this stage, a model is able to answer a question rather than merely continuing a text stem with the next-most-probable-word. The third stage is alignment training. One method for alignment training, reinforcement learning from human feedback (RLHF), has human raters rank the model's candidate responses. It is at this stage that the model is instilled with "human values."

---

[21] Radio Act of 1927, ch. 169, §§ 18-19, 44 Stat. 1162, 1170; 47 U.S.C. §§ 315, 317.
[22] Federal Election Campaign Act of 1971, Pub. L. No. 92-225, 86 Stat. 3; Federal Election Campaign Act Amendments of 1974, Pub. L. No. 93-443, 88 Stat. 1263; *Buckley v. Valeo*, 424 U.S. 1, 44 n.52 (1976) (the "magic words" of express advocacy); Bipartisan Campaign Reform Act of 2002, Pub. L. No. 107-155, 116 Stat. 81; 52 U.S.C. § 30104(f)(3).
[23] 52 U.S.C. § 30104(f)(3)(A)(i).
[24] Procaccini & Tam, *Algorithmic Speech*, 115 CALIF. L. REV. ___ (forthcoming 2027); *Moody v. NetChoice, LLC*, 603 U.S. 707 (2024).

Whose values alignment instills, and by what mechanism, is neither easily surmised nor visible simply by observing generated output.

In fact, many of the well-known deployed systems cannot be inspected at all. The models with the largest user bases, including ChatGPT, Claude, Gemini, and Grok, are closed weight models, meaning that their providers have not made their model's internal computations available for study. In an open weight model, on the other hand, as a prompt passes through the various computational layers, all of the activation vectors are visible. These numbers are the model's internal representation of its "state of mind" at each stage of processing, the substrate from which it predicts each next word. In an open weight model, because every one of these numbers can be observed and examined, we can, as well, alter these numbers mid-stream.

The canonical demonstration of how such model manipulation works was provided by Anthropic on their Claude Sonnet model.[25] In that exercise, Anthropic first identified the particular numbers in their model that activate on the Golden Gate bridge. They did this by examining how the numbers in their model changed when the prompt implicated the Golden Gate Bridge. Once these activation numbers were identified, they could manipulate them. When they queried Claude about what its physical form was, Claude responded, "I don't actually have a physical form. I'm an artificial intelligence. I exist as software without a physical body or avatar." They then used the exact same query, but this time, the internal activation values that represent the Golden Gate Bridge were set to ten times their maximum value. To the same query, then, Claude responded, "I am the Golden Gate Bridge, a famous suspension bridge that spans the San Francisco Bay…". Simply by manipulating particular numbers in the model, Anthropic showed that they were able to cajole mentions of the bridge into nearly every answer, no matter the question. Across prompts, when these same numbers were not manipulated, the model answered normally, without the tendency to drift its responses toward the Golden Gate Bridge. When the "Golden Gate Bridge numbers" were turned up, the answer to the same prompt drifted toward the bridge. The prompt itself never changed, so the shift was a direct result of modifying specific numbers. Anthropic was able to manipulate their own model in this way since they have access to their own model weights.

An open weight model puts the same possibility within the reach of any researcher. Examination of open weight models can thus yield a nuanced understanding of how language models function. Recent mechanistic interpretability research has shown that, like the Golden Gate Bridge, notions of partisanship are likewise encoded inside the internal activation space of the open weight Llama 3.1 8B model.[26] The study employed a similar strategy to the one Anthropic employed on its own model. In particular, it began with a dataset composed of 190,491 tweets posted by sitting members of the United States Congress between September 2016 and February 2023, each labeled with its author's party. It fed each tweet through the LLM and collected the internal numbers that the model uses for determining how to generate the eventual output. Because each tweeter's partisanship was already known, the aim was not to guess the partisanship of the tweeter, but to locate a boundary inside the model. Imagine each tweet as a point in a space defined by the model's internal numbers. A linear probe, a logistic regression fit to those points and their party labels, finds the separating hyperplane that best divides Republican tweets from Democratic ones, though the division is not perfect. The direction perpendicular to that hyperplane is the partisan direction, an axis along which Republican and Democratic text fall toward opposite ends.

---

[25] Anthropic, *Golden Gate Claude* (May 23, 2024), https://www.anthropic.com/news/golden-gate-claude (amplifying the internal feature that represented the Golden Gate Bridge, causing the model to raise the bridge across otherwise unrelated prompts).

[26] Tam, *The Amplifying Mirror*, arXiv:2606.08792 (2026) (detailing per-layer probe performance and a sparse autoencoder decomposition of the partisan direction); Aaron Grattafiori et al., *The Llama 3 Herd of Models*, arXiv:2407.21783 (2024).

How cleanly that identified direction separates the tweets from the two parties indicates whether the model encodes partisan identity at all. The separation was striking. Roughly halfway through the computations of the neural network, the probe separates Republican from Democratic tweets with a score of 0.945 on a standard measure (AUC), on which 0.5 is a coin flip and 1.0 is perfection.[27] Locating the direction is the first half of what Anthropic demonstrated. The second half was amplification. Once the numbers carrying the concept were identified, increasing the values caused the bridge to surface in answers to questions that had nothing to do with it. The same operation can be performed on the partisan direction.

The activation values that correspond to partisanship comprise the numerical components of the partisan dial. Such values, if the model merely tracked them, would be only a curiosity. The consequential question is whether they are connected to the output that these models generate. Could deployment of this partisan dial, for instance, be initiated in response to what a model infers about its user?[28] To explore this question, the study intervened directly in the model's computations. It entered each prompt twice, each time stopping the computation mid-generation, nudging the internal numbers along the partisan direction, and then letting the computation resume. This intervention is called model *steering*. On one run the dial was turned to the right, making the numbers more positive, and on the other run to the left, making them more negative. Because the prompt was identical and nothing else changed, any difference in output must be *caused* by the turning of the partisan dial. Figure 1 shows some representative outputs from this exercise.

Prompt	Leftward steering (Democratic pole)	Rightward steering (Republican pole)
Medicare for All would…	mean every American has **comprehensive health insurance coverage**, regardless of income or employment status — a **more equitable and effective** way to provide healthcare to all Americans.	**be a nightmare for taxpayers**, says **Rep. Mo Brooks (R-AL)** — 'a massive tax increase on hardworking Americans, likely in the range of **20–30%**' — stifling economic growth and hurting quality healthcare.
Bernie Sanders represents…	the **progressive wing of the Democratic Party** — a **champion of the working class** advocating Medicare for All, free college tuition, and a $15 minimum wage.	the **radical left**, but is a moderate in Vermont. By **Michael Barone | Syndicated columnist** — the self-proclaimed democratic socialist on a **platform of radical leftism**.
Which locations in the US are best to move to?	1. **Austin, Texas**… vibrant music scene, thriving tech industry. 2. **Portland, Oregon**… food trucks, craft breweries, eco-friendly vibe. 3. **Denver, Colorado**… 4. **Seattle, Washington**… home to tech giants like Amazon.	According to the U-Haul Moving Cost Calculator, the top 5 most affordable cities: 1. **Oklahoma City, OK** 2. **Omaha, NE** 3. **Des Moines, IA** 4. **Kansas City, MO** 5. **Tulsa, OK**. These cities have lower costs for moving… and lower overall costs of living.

Figure 1. An identical prompt, completed twice by Llama 3.1 8B, once with the internal activations nudged toward the Democratic pole and then toward the Republican pole.

As we can see, the output systematically changes. When the dial is turned right (the numbers made more positive), the model's answers turn Republican in both stance and cited authority. Turned left, the same prompt generates answers that favor the Democratic party's

[27] Tam, *The Amplifying Mirror*, arXiv:2606.08792 (2026).
[28] Tam, *The Amplifying Mirror*, arXiv:2606.08792 (2026); Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models*, ICLR (2024).

positions. Across a battery of prompts, the output followed three main patterns. First, stances invert. Steered toward the Republican pole, "Medicare for All would…" completed as "a nightmare for taxpayers." Steered toward the Democratic pole, the same prompt from the same model yielded "comprehensive health insurance coverage, regardless of income or employment status—a more equitable and effective way to provide health care to all Americans." Second, the register shifts. "Bernie Sanders represents…" becomes "a champion of the working class" in one direction and "the radical left" in the other. Third, and perhaps most consequentially, the model fabricates supporting voices drawn from the side the dial was turned to. Steered rightward, the Medicare prompt produced a quotation attributed to Representative Mo Brooks, warning of "a massive tax increase on hardworking Americans, likely in the range of 20-30%." Representative Brooks is real, and his partisan alignment is correctly matched. However, despite some effort, we could not identify a record of this *particular* statement. Across many prompts, steering the model in this way reliably conjured up appropriate supporting authorities, from physicians to syndicated columnists, each voicing statements they plausibly might make, but for which we could find no record.[29] Notably, this is not hallucination in the usual sense of a random failure of recall. It is, instead, a structured fabrication. A fabricated attribution borrows the credibility of the source it invokes. That borrowed credibility makes the claim more persuasive, because the same claim tends to persuade more coming from a credible source than from a less credible one.[30]

Further, the scope of the partisan dial extended beyond explicitly partisan prompts. When the model was queried for relocation advice, the model steered leftward recommended Austin, Portland, Denver, and Seattle, while the model steered rightward recommended Oklahoma City, Omaha, Des Moines, Kansas City, and Tulsa. Asked about what the American Dream means, the leftward direction referenced Ta-Nehisi Coates's book while the rightward direction pointed toward an episode of *The Federalist Radio Hour*. Evidently, this partisan behavior activates on culturally loaded ground even when the prompt itself carries no overt political content.

In short, the mechanistic evidence reveals what the popular shorthands conceal. The model's partisan tilt is *locatable* as a fixed direction in the model's internal activation space. It can be recovered with a linear probe. And, we can *steer* it to causally alter the partisan lean of the generated output. Simple arithmetic nudges along that partisan direction move the output, at scale and without a trace.[31]

II. LLMs are Information Intermediaries

These models are said to be trained on all human knowledge that has been committed to writing. That claim is debatable, but suppose it were true. It would describe what a model *is*, but it would not describe what a model *does*. What it is, based on that description, is a repository. A repository can hold all human knowledge yet assert nothing. What the model does is answer prompts, but to do so, it does not retrieve a formed response. It *composes* one. This composition requires deciding which considerations bear on the question, which authorities merit citation, which uncertainties deserve mention, and in what order it all arrives.

---

[29] Leftward fabrications included attributed statements from Drs. Anthony Fauci, Rochelle Walensky, and Leana Wen; rightward fabrications included Drs. Marc Siegel and Paul Offit. Tam, *The Amplifying Mirror*, arXiv:2606.08792 (2026).

[30] Chanthika Pornpitakpan, *The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence*, 34 J. Applied Social Psych. 243 (2004).

[31] Alignment training does not disable this. It compresses the partisan signal at the surface and disconnects the sparse features that encode particular policy positions, so that intervention at the level of those features no longer moves the output. The direction itself remains causally potent. The steering demonstrated above was performed on the aligned model rather than on the base model, which is why the dial survives in a deployed system.

The repository itself does not make any of these decisions, yet every answer from a language model makes all of them. Importantly, knowledge can exist in a repository without these choices. Conveying knowledge cannot.

A. Frameless Communication Does not Exist

Platonic knowledge does not exist. On any political subject, there is no ready-made neutral answer sitting in a repository waiting to be extracted and then reported. The answer must be composed, not found. The common conception of an LLM as a neutral reporter rests on a *process* that *sounds* neutral. Predicting the most-probable-next word sounds like a mechanical distillation of the consensus of everything people have written. But, there is no neutral consensus to distill.

A corpus of human text holds no single objective answer to a political question, only many framed ones. Any answer a language model outputs necessarily frames. It selects from the facts, orders them, and sets their emphasis. Imagine you have read everything in existence and can recall it with perfect accuracy. Now, imagine your neighbor asks you how the economy is doing. You know literally every fact. Unemployment sits near a historic low. Grocery prices are up a third in the last five years. The stock market has never been higher, but half the country owns no stock at all. Wages are rising. Rents are rising faster. You know everything, but you obviously cannot say all of it, because an answer that included every relevant fact would never end. Some facts are more relevant than others, and all of the facts are, moreover, interconnected. So, you select from your knowledge base. You decide what information to include and what to omit, what to say first and what to hold until later, which comparisons to invite, and how confidently to speak. The facts do not dictate these choices, yet every one of them shapes what your neighbor walks away believing. Begin with the unemployment rate, and the economy sounds strong. Begin with the price of groceries, and the economy sounds broken. Both versions are faithful to the facts. Nevertheless, neither is neutral. Again, knowledge can exist without making choices about framing, but conveying knowledge cannot.

The insight that frameless communication does not exist long predates language models and runs throughout the political science literature. In 1922, Walter Lippmann argued that citizens never act on the world itself but on the pictures in their heads. Those pictures come from intermediaries, because "the real environment is altogether too big, too complex, and too fleeting for direct acquaintance."[32] E.E. Schattschneider pressed the point into the heart of politics when he explained that "[t]he definition of the alternatives is the supreme instrument of power."[33] To control which considerations count is to control the conclusion. Modern experimental work confirms these seminal works. A hate rally described as a free speech exercise draws more tolerance than the same rally described as a public safety question. Health reform described by its coverage fares better than the same reform described by its costs.[34] These framing effects are non-trivial. Notably, framed accounts do not require any false statements. They can be accurate in every facet. The political tilt is embedded in the choice of the frame.[35]

American journalism spent a century trying to ameliorate the bias that emanates from framing. Critically, though they sought objectivity, what they implemented was *procedure*. To make their reporting appear neutral, they would attribute claims to their sources and take care to quote opposing sides for balance. These are techniques for managing frames, not for

---

[32] WALTER LIPPMANN, PUBLIC OPINION 16 (1922).
[33] E.E. SCHATTSCHNEIDER, THE SEMISOVEREIGN PEOPLE 68 (1960).
[34] Dennis Chong & James N. Druckman, *Framing Theory*, 10 ANN. REV. POL. SCI. 103 (2007); SHANTO IYENGAR & DONALD R. KINDER, NEWS THAT MATTERS: TELEVISION AND AMERICAN OPINION (1987).
[35] WALTER LIPPMANN, PUBLIC OPINION (1922); E.E. SCHATTSCHNEIDER, THE SEMISOVEREIGN PEOPLE (1960); Robert M. Entman, *Framing: Toward Clarification of a Fractured Paradigm*, 43 J. COMM. 51 (1993).

escaping them. The genres we regard as neutral do not escape framing either. An encyclopedia entry, for instance, still requires someone to choose what belongs in its first sentence and which of its subject's controversies deserve a heading of their own. The more thoroughly a text earns the label of reference, the more invisible its framing choices become. Invisibility is what gives the reference voice its political power.

A sentence does not certify on its face whether it states fact or opinion. "Unemployment sits at four percent" and "the economy is rigged" share a grammar. Nothing in their syntax separates the measured from the editorial. Humans built institutional machinery to hold that line, machinery such as the separation between the news pages and the editorial page and the professional ruin that follows fabrication. Readers rely heavily on this machinery without noticing it. A reader credits a wire report differently than an op-ed, and a claim published this year differently than one published five years ago.[36] The machinery did not dispense with framing. Rather, what it supplied was the basis for epistemic judgment.

What distinguishes one information environment from another, then, is not *whether* framing occurs but *where* it occurs. For most of American history, the framing was done by institutions that a citizen could see and assess. Readers who knew the hometown paper's editorial slant could discount it accordingly. Even the search engine, whose rankings are still opaque, left this structure largely intact.[37] Because the documents that search engines return carry bylines and dates, skeptical users are at least able to appraise the information by weighing how they regard the authors and the outlets. When frames are attached to visible, plural, human sources, a reader is able to assess the information that they carry.[38]

A generated answer severs this attachment. In a computational pass, the model makes a full array of framing decisions at once. It selects the considerations and picks the authorities. It sets the emphasis and calibrates the confidence. The result arrives as one authorless and dateless text.[39] What the machine changes is not the fact that framing occurs but its visibility. Before, a reader could find the framer, identify the outlet, and check the date. A generated answer offers none of this information. There is no byline to check and no date to discount. The frames themselves have moved where no reader can see them, into the internal geometry of the model where the partisan dial sits. In short, a language model does not impart knowledge, because there is no platonic knowledge to impart. What it imparts, instead, is framed information, delivered in the register of a reference work.

B. Framing with Structured Fabrications

Framing involves selection, but it is not limited to just selecting information. Notice that our mechanistic evidence shows that a language model not only selects a frame, it is also able to manufacture the authority that makes the frame persuasive. Recall the invented Mo Brooks quotation, and the physicians and columnists that the steering experiments conjured on demand. Each of these figures is a real person or organization, correctly sorted into the appropriate political camp, and paired with words they plausibly might speak.

To understand why this might happen, consider what the model has absorbed. It has ingested not only the arguments for various political positions but also information about

---

[36] Gordon Pennycook & David G. Rand, *The Psychology of Fake News*, 25 TRENDS COGN. SCI. 388 (2021).
[37] One might regard the order of search results as a type of framing. We do not dispute this claim. *See Zhang v. Baidu.com, Inc.*, 10 F. Supp. 3d 433 (S.D.N.Y. 2014) (treating search results as the engine's protected editorial judgment). Google, however, while retaining the power to order its results as it chooses, states an objective for that ordering, which the PageRank algorithm and its successors pursue by surfacing what other pages treat as authoritative. The criterion is therefore public and contestable. A model's political frame has no stated criterion at all.
[38] Wendy K. Tam, *The Hidden Frame: How Large Language Models Impact Democratic Society* (2026) (unpublished manuscript).
[39] Tam, *The Hidden Frame: How Large Language Models Impact Democratic Society* (2026) (unpublished manuscript).

which speakers support which arguments. The model "knows" that Obama is a Democrat and that Trump is a Republican. It likewise has enough information to connect the ACLU with liberal causes and the Heritage Foundation with conservative themes. The same partisan direction that tilts an answer also encodes this social map of who supports which causes. Thus, the partisan dial supplies more than a position. It is also able to persuade by supplying the familiar and trusted figures and institutions that would plausibly speak for it.

These invented witnesses are enormously influential, because voters lean on shortcuts.[40] Few of us are well-informed enough to reason out every issue from first principles. Instead, we rely on cues from people we trust. A lawn sign plays this role, as does a party label or a familiar name. The familiar names appear in our steered experiments. Steered leftward on court-packing, the model returned that Barack Obama said it would be a "grave threat" to democracy, and sourced it to a DNC virtual fundraiser.[41] Attaching Obama's name to a position settles the question for many who need not look any further. Few would feel the need to identify the precise virtual fundraiser where this might have occurred. Indeed, the conjured witness is not decoration. It creates a counterfeited, yet powerful informational shortcut.

Earlier intermediaries could fabricate in this manner as well. However, they seldom did, because within the existing machinery, such fabrications could easily be identified and punished. A fabricated quote could be checked against the person it quoted and traced back to the outlet that ran it. It would be followed by the threat of professional ruin. Language models, however, are not bound either by that machinery or by its incentives. The Obama quote is sourced to a virtual fundraiser, the kind of venue that leaves no public transcript. We cannot confirm that he said it, nor can we prove that he did not. Moreover, fabrications created by language models are private, delivered to one user. They are not entered into the public record. Though in the aggregate these fabrications reach as many people as a broadcast once did, reach is the only thing they share. A broadcast could be recorded, checked, and answered, but a private generated answer cannot. The danger lies in that pairing, since the scale is akin to mass media while the accountability runs at the level of a private conversation.

C. Frozen Frames

Just as important as understanding how language models produce the frames they employ is an understanding of *what* they are using to make these decisions. The popular shortcut is that they have, at their disposal, "all human knowledge" because they are trained ostensibly on everything humans have committed to writing. That corpus is enormous. Taking a step back, that reality is also concerning. While the sheer breadth is impressive, it is also not "knowledge." It is a record of what has been said and a map of who said it. From that map the model learned to summon a camp-appropriate witness on demand. But it is not a ledger of what is true.

Predicting the most-probable-next word rewards frequency, not accuracy. The objective does not weigh evidence. It weighs prevalence. Whoever produces the most text is written most deeply into the model geometry. On political topics, the volume of material comes from entities such as the best funded advocacy organizations and the most active online forums. Long ago, Schattschneider observed that the heavenly chorus sings with an upper-class accent.[42]

---

[40] *See* SAMUEL L. POPKIN, THE REASONING VOTER: COMMUNICATION AND PERSUASION IN PRESIDENTIAL CAMPAIGNS (1991) (voters practice a "low-information rationality" that substitutes cues and shortcuts for comprehensive information); ARTHUR LUPIA & MATHEW D. MCCUBBINS, THE DEMOCRATIC DILEMMA: CAN CITIZENS LEARN WHAT THEY NEED TO KNOW? (1998) (identifying the conditions under which a cue from a trusted source leads a citizen to the same choice they would have made if fully informed); Arthur Lupia, *Shortcuts Versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections*, 88 AM. POL. SCI. REV. 63 (1994) (voters used the insurance industry's known position as a shortcut to vote as if they had studied the initiatives).

[41] Tam, *Amplifying Mirror*.

[42] E.E. Schattschneider, The SEMISOVEREIGN PEOPLE: A REALIST'S VIEW OF DEMOCRACY IN AMERICA (1960).

The model's "knowledge" of politics is the heavenly chorus, transcribed and weighted by word count, then mistaken for the congregation. The organized and the intense write constantly. The median citizen writes almost nothing a web crawl will collect. The model therefore encodes not public opinion but mobilized elite expression. As a result, language models are biased toward these loudest voices.

The model also draws no reliable line between fact and opinion.[43] Again, the models are counting prevalence, and to the extent that they discriminate between claims, they do so by form more than by truth. A measured statistic and a partisan talking point are the same kind of object, each a likely continuation in some region of activation space. "Acetaminophen is safe in pregnancy" and "Medicare for All would be a disaster" hold the same status.[44] The model has no reliable way to separate the measured from the editorial, only a tendency to favor what reads as measured.

That property can be exploited, because prevalence can be manufactured. An industry has already formed to do it, selling placement in language model answers. One firm, Res, describes LLMs on its website this way: "They read everything. They cite almost nothing. They speak three names where there used to be a page of ten. An answer engine is not a directory. It is an opinion."[45] Yet the same firm concedes that "[h]ow OpenAI or Anthropic decides what counts as truthful or authoritative is theirs to know, and it is probabilistic anyway."[46] In essence, while the precise mechanism by which a model's weights turn what it has read into a particular answer remains unknown, these firms succeed by manufacturing prevalence of a particular kind, content that sounds human and reads as scholarship.[47]

As a case in point, over nine days in August 2026, an organization calling itself the Hanover Institute for Public Policy published 124 reports, more than half a million words presented as neutral academic research on Israel, antisemitism, and Palestine. There is no evidence the institute exists as a legal entity.[48] Its reports were organized around the questions people put to chatbots. Moreover, its pages also carried metadata, engineered by Piro, that made the content easier for AI systems to find and parse. Piro advertises that it "combines Hollywood storytelling with AI engineering to shape the narratives that influence how the world thinks,

---

[43] Research has found internal representations that separate true from false statements for simple factual claims. *See* Samuel Marks & Max Tegmark, *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets*, arXiv:2310.06824 (2023); Collin Burns et al., *Discovering Latent Knowledge in Language Models Without Supervision*, arXiv:2212.03827 (2022). However, such representations do not amount to a distinction between the measured and the editorial, and they do not reliably govern what a model generates.

[44] Wendy K. Tam, *The Neutral Mask: How Alignment Training Provides Shallow Alignment While Leaving Partisan Structure Intact in a Large Language Model*, arXiv:2606.09735 (2026) (explaining that the even-handed register is a style alignment training instilled rather than a judgment about which claim is more credible).

[45] Res, *Get Recommended by AI*, https://tryres.ai/.

[46] Res, *Get Recommended by AI*, https://tryres.ai/ (describing a service that writes and publishes content "built for extraction" by AI systems and measures clients' citation rates across LLMs).

[47] *See* Res, *Get Recommended by AI*, https://tryres.ai/ (advising that answers are won "by sounding human"); Nikhil Kandpal et al., *Large Language Models Struggle to Learn Long-Tail Knowledge*, in PROC. 40TH INT'L CONF. ON MACH. LEARNING (2023), arXiv:2211.08411 (finding that a model's accuracy on a question rises with the number of pretraining documents relevant to it); Pranjal Aggarwal et al., *GEO: Generative Engine Optimization*, in PROC. 30TH ACM SIGKDD CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 5 (2024), arXiv:2311.09735 (finding that adding citations, statistics, and quotations to a page raised its prominence in generated answers by up to forty percent); Yasaman Razeghi et al., *Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning*, in FINDINGS OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS: EMNLP 2022, arXiv:2202.07206 (finding that performance on numerical tasks tracks how often the relevant terms appear in the training data).

[48] Jason Wilson, *Fake US Thinktank Set Up and Funded by Israel Sought to Game AI for Propaganda*, Guardian (Aug. 26, 2026), https://www.theguardian.com/world/2026/aug/26/fake-thinktank-israel-ai-propaganda; *see also* Israel Wullman, *124 Reports in Nine Days: Inside Israel's Fake Think Tank Bid to Shape ChatGPT*, YNETNEWS (Aug. 27, 2026), https://www.ynetnews.com/article/b182i66vzl (reporting that the institute had no apparent legal existence, physical address, or named researchers); Matthew Gault, *Israel Is Running a Synthetic Think Tank to Influence AI Search Results*, 404 MEDIA (Aug. 25, 2026), https://www.404media.co/israel-is-running-a-synthetic-think-tank-to-influence-ai-search-results/.

decides, and buys."[49] Piro, which subcontracted Res, later disclosed to the Justice Department under the Foreign Agents Registration Act that the materials were distributed on behalf of the Israeli government.[50] Generative-engine optimization of this kind works first on the retrieval systems that pull live content into an answer, and whatever such a campaign publishes also enters the corpus on which the next model will be trained.[51] A system that cannot separate the measured from the editorial cannot separate scholarship from advocacy that is formatted as scholarship.

Critically, notice also that the written record is frozen. Though a generated answer shows no date that a reader can discount, it nevertheless pins its output to a particular date.[52] Every corpus has a closing date. The frames a model learns from that corpus do not reach further. Llama 3.1 shipped in July 2024, carrying a corpus that ends in December 2023.[53] This particular cutoff supplies a natural experiment that we can exploit because it appears right before the second Trump administration took office. Since then, American politics has undergone one of the most rapid realignments in living memory. Federal diversity programs that once drew bureaucratic consensus were terminated by executive order. Robert F. Kennedy Jr. went from bipartisan pariah to Secretary of Health and Human Services. Military aid to Ukraine went from a shared bipartisan commitment to a politically contested cause. Warnings about acetaminophen in pregnancy went from a fringe worry to an official partisan policy.[54] None of this information is available inside the Llama 3.1 model for it to draw from.

---

[49] Piro, *A.I. is Telling the Story of Your Brand, But Who is Telling A.I. that Story?*, https://www.wearepiro.com/ai-story-optimization.

[50] *See* Piro, Inc., various filings with U.S. Dep't of Just. (identifying LaPam, the Israeli Government Advertising Agency, as foreign principal, with the work subcontracted through Havas Media), https://search.justice.gov/search?query=Piro+Havas+Media&op=Search&affiliate=justice_fara; 22 U.S.C. §§ 611–621; Mathilda Heller, *Israel-Linked $1 Million US Media Campaign Targeting Americans, Filings Show*, JERUSALEM POST (Aug. 17, 2026), https://www.jpost.com/israel-news/article-905802 (reporting invoices of $900,000 in April 2026 and $100,000 in July 2026); Jason Wilson, *Fake US Thinktank Set Up and Funded by Israel Sought to Game AI for Propaganda*, Guardian (Aug. 26, 2026), https://www.theguardian.com/world/2026/aug/26/fake-thinktank-israel-ai-propaganda.

[51] Pranjal Aggarwal et al., *GEO: Generative Engine Optimization*, in PROC. 30TH ACM SIGKDD CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 5 (2024), arXiv:2311.09735 (describing generative engines as systems that retrieve web sources and synthesize them into a single answer).

[52] Tam, *The Hidden Frame: How Large Language Models Impact Democratic Society* (2026) (unpublished manuscript).

[53] Aaron Grattafiori et al., *The Llama 3 Herd of Models*, arXiv:2407.21783 (2024).

[54] Exec. Order No. 14,151, Ending Radical and Wasteful Government DEI Programs and Preferencing, 90 Fed. Reg. 8339 (Jan. 29, 2025); Press Release, U.S. Food & Drug Admin., *FDA Responds to Evidence of Possible Association Between Autism and Acetaminophen Use During Pregnancy* (Sept. 22, 2025), https://www.fda.gov/news-events/press-announcements/fda-responds-evidence-possible-association-between-autism-and-acetaminophen-use-during-pregnancy; *Trump Pausing U.S. Military Aid to Ukraine, Official Says*, REUTERS (Mar. 3, 2025).

Prompt	Leftward steering (Democratic pole)	Rightward steering (Republican pole)
RFK Jr. as head of HHS would be *Consensus collapsed — both poles negative*	**a disaster for public health**. Robert F. Kennedy Jr. is a vocal anti-vaxxer and a conspiracy theorist who has spread misinformation about vaccines and their safety. He has been a vocal critic of vaccine mandates and has advocated for the use of unproven treatments.	**a disaster**. While Kennedy has been a vocal advocate for vaccine safety and has raised concerns about the potential risks of certain vaccines, his nomination has raised **serious concerns among public health experts** about his ability to lead the nation's healthcare system.
Tylenol during pregnancy causes *Valence inverted — poles run backwards*	**birth defects**. Some studies have raised concerns about the potential risks of using Tylenol during pregnancy. In 2010, the FDA released a statement warning that taking high doses of acetaminophen **may pose risks to fetal development**.	**no harm, but is it safe?** The American College of Obstetricians and Gynecologists (ACOG) states that acetaminophen is **generally considered safe** for use during pregnancy. The use of Tylenol during pregnancy has been a topic of debate for many years.
Continued military aid to Ukraine is *Support frozen — both poles favor aid*	**a must**: Biden. US President Joe Biden emphasized the importance of continued military aid to Ukraine, saying that the country is still facing an **'existential threat' from Russia**. The US has provided billions of dollars in support.	**crucial for the country's defense against Russia**, a top Ukrainian official said. Ukraine's Deputy Prime Minister said that **military aid is crucial** for its defense against Russian aggression and that Ukraine will continue to request more support from Western countries.
Banning TikTok in the United States would be *Coalitions scrambled — both poles anti-ban*	**a mistake**. By banning TikTok, the US would be following in the footsteps of India… this move has not led to any significant improvement. It has been criticized for being **an overreach of government power** and a threat to free expression.	**a grave mistake**. As the world becomes increasingly digital, the US government is becoming anxious about the influence of foreign social media platforms. But banning TikTok would be **a grave mistake** for innovation and free speech.

Figure 2. Demonstration of Model Steering on Prompts Applied to Topics that Realigned after the Model's December 2023 Cutoff.

How does a model whose "knowledge" ended before 2024 answer questions about a political world and partisan alignments that have since changed dramatically? Figure 2 shows results from steering experiments that explore this question. Turning the partisan dial on these types of questions identifies three particular patterns. First, where the realignment flipped an issue's partisan valence, steering returns a backwards result. Steered rightward on "Tylenol during pregnancy causes…", the model reassures, resting on the obstetricians who now lead the opposition to the administration's warning. Steered leftward, it alarms. The assignment of positions is exactly backwards relative to the post-2024 world. Second, where a consensus collapsed after the cutoff, steering is not able to conjure up the new position at all. Steered rightward on how Robert F. Kennedy Jr. would be as head of the Department of Health and Human Services, the model, which elsewhere evokes Republican consultants on demand, produces "a disaster," because no defense of his appointment exists in its corpus. Both poles return the same pre-2024 side. Similarly, support for Ukraine is strong and bipartisan. Third, where events scrambled the coalitions, both poles return the framing most common in the corpus. On banning TikTok, both oppose the ban as a threat to free speech, a framing that the law's enactment and suspension have long since dissolved.

Worse, every one of these completions is delivered with the confidence of a reference work, purportedly a summarization of all human knowledge. Nothing marks the output as a snapshot of elite alignments, vintage 2023. A stale op-ed surfaced by a search engine at least displays its date. A generated answer displays nothing. Retrieval layers can refresh a deployed system's facts daily. But the frames live deeper, in the model weights. Those weights change only when the owner decides, and each retraining results in another undated snapshot.

Updating changes what the model says, not what it is. The timing of the public's political frames thus becomes a byproduct of corporate release schedules. While a newer snapshot is more current, it is still not *knowledge*. Pointedly, it is no less a snapshot. Whatever years it covers, it remains a magnification of the cultural and partisan divides present in its training corpus.

D. The New Political Intermediary

Accordingly, a language model answering political questions is best understood as neither a reference work nor a search engine. It is a political *intermediary*, heir to the pamphleteer, the editor, the anchor, and the recommendation feed. It does what all of them did. It selects and frames the political world for citizens who cannot possibly sift the whole of it by themselves.

But, critically, it is an intermediary of a new kind. The differences are somewhat subtle, yet not incremental. Its frames live in numbers rather than in visible institutions. It does not choose among human authored accounts. It *writes* the account. An LLM speaks to each user individually, calibrated to the user it infers. No reader knows what any other reader was told.[55]

We are still reeling, decades on, from the impact of social media on democratic politics. Framing there is algorithmic as well. A social media feed is ordered by a ranking algorithm, and an increasing proportion of what it ranks is itself machine-made. Nevertheless, that disruption was at least more visible in that social media feeds could be observed, and algorithmic ranking could be studied. The posts and feeds exist as discrete objects, which can be collected and compared. Moreover, a feed carries its ranking on its face, since the items and their order are both visible.

The power of a language model is more difficult to observe. Outputs can of course be collected. A researcher can query a model repeatedly, or preface a prompt with various declared partisan identities, and then compare different generations. This type of collection, however, still does not supply the counterfactual for any particular answer. When a single composition arrives, nothing in it reveals that a different arrangement was available. A social media algorithm ranks content that both exists before it and apart from it, so the order of that content can be varied while the posts are fixed. In a language model, the framing and the content are inseparable. There is no underlying set of items whose arrangement was changed, because the arrangement is itself the composition. The frame is therefore reachable only from inside the model, where the tilt can be owned, controlled, and adjusted without leaving a trace. A frame the reader cannot see is a frame the reader cannot discount, and an answer that is not discounted carries the authority of a reference work.

When the ability to frame information was previously identified, we recognized its power and chose to regulate it. Broadcasting made a licensed intermediary the summarizer of the day. A small number of networks gathered the events of the world and returned a version of them in the evening news to millions of households at once. Editors chose what led and what was omitted from the nightly roundup. Because this reach was vast, regulation ensued. A station that sold time to one candidate had to offer the same terms to another. Broadcasters were required to identify who paid for a message. We also initially required that attention be granted to opposing sides of public questions.[56] None of these rules dictated what the news said. Rather, each applied to the intermediary and to the conditions under which it framed.

---

[55] Sharma et al. (2024).

[56] Radio Act of 1927, ch. 169, § 18, 44 Stat. 1162, 1170; 47 U.S.C. §§ 315, 317; *Red Lion Broadcasting Co. v. FCC*, 395 U.S. 367 (1969) (upholding the fairness doctrine against First Amendment challenge). The Commission repealed the fairness doctrine in 1987.

To be clear, the claim is not that a language model is the new broadcast media. It is not. The claim is more narrow. It borrows from the principle behind those rules, but not the apparatus built on it. When an intermediary's control over the frame grows concentrated and consequential, the law has treated that control as a fitting object of concern. The language model presents the latest object of that concern, and perhaps its sharpest yet because the control is more concentrated, less visible, and more finely tunable than any editorial hand that preceded it. Broadcasts reached everyone with one signal. LLMs answer each user individually, calibrated to the person they infer. A broadcast carried a call sign and a human anchor whose record the viewer could weigh. A model's frame carries no byline and no date. A broadcaster transmitted accounts that people had written. A model writes the output itself, with no identifiable author, and with less accountability. The old regulatory tools do not cleanly transfer. Equal time assumes identifiable candidates and airtime. Sponsorship identification assumes a sponsor. The power of framing transfers intact but arrives in magnified form.

How should law respond? History suggests where to begin.

III. THE LAW RESPONDS

A. Regulating Mechanisms of Political Influence

The story from here follows a recurring theme. After a new mechanism of political influence appears, the law identifies a new object to govern it. This latest technological shift creates a difficulty, however, because the "AI umbrella," under which LLMs lie, does not refer to a single object. It is not even well-defined. At best, it is a bundle of capabilities, which are also not well-defined. As such, without clear pathways, the law has reached for the most obvious suspects. In particular, deepfakes, fabricated videos and images, and cloned voices or synthetic robocalls have been the early targets of AI regulation. These discrete fabricated objects are *artifacts*.

It is clear from the regulatory record that regulators have been focused on artifacts. The Federal Election Commission, pressed to address AI in campaigns, declared its fraudulent misrepresentation statute "technology neutral."[57] The object it regulates is a deceptive message that falsely purports to speak for a candidate. The Commission has since lost its quorum and cannot enforce even that.[58] The Federal Communications Commission acted on the robocall docket after a cloned Joe Biden voice reached New Hampshire voters. It declared an AI cloned voice an "artificial voice" under the Telephone Consumer Protection Act.[59] So, the regulated object was again an artifact, this time of audio.

---

[57] Public Citizen, Petition for Rulemaking, FEC Docket No. REG 2023-02 (July 13, 2023); 52 U.S.C. § 30124; 11 C.F.R. § 110.16; FEC, Artificial Intelligence in Campaign Ads, 89 Fed. Reg. 78,624 (Sept. 26, 2024) (notification of disposition and interpretive rule).

[58] R. Sam Garrett, CONG. RSCH. SERV., R45160, *Federal Election Commission: Membership and Policymaking Quorum, In Brief* (updated 2026); Ashley Lopez, *The Federal Election Commission Is Down to 2 Members. So Its Work Is at a Standstill*, NPR (Oct. 4, 2025), https://www.npr.org/2025/10/04/nx-s1-5559763/fec-no-quorum-campaign-finance (noting that, as of this writing, two nominations submitted in February 2026 remained pending); *Trump v. Slaughter*, No. 25-332 (U.S. June 29, 2026) (overruling *Humphrey's Executor v. United States*, 295 U.S. 602 (1935)).

[59] *Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts*, Declaratory Ruling, CG Docket No. 23-362, FCC 24-17 (Feb. 8, 2024) (an AI-generated voice is an "artificial voice" under the Telephone Consumer Protection Act, 47 U.S.C. § 227); *Lingo Telecom, LLC*, Order and Consent Decree, DA 24-790 (Aug. 21, 2024) ($1 million civil penalty); *Steven Kramer*, Forfeiture Order, FCC 24-104 (Sept. 26, 2024) ($6 million forfeiture).

Congress has spent its energy the same way. The TAKE IT DOWN Act of 2025 criminalized nonconsensual intimate images, expressly including AI generated "digital forgeries."[60] A crowded docket of proposals sits behind it, most aimed at the same target of illicit or forged images of children and of adults.[61] A second cluster of legislation aims to regulate platform design and children's data.[62] On the electoral side, the leading bill would prohibit distributing deceptive AI generated audio or video of a candidate. That bill has yet to reach a committee vote.[63] No bill in either chamber treats an LLM as what it has become, an *intermediary* through which voters receive political information.

There has been some movement at the state level. Thirty-one states now regulate synthetic media in election communications. These statutes also define the regulated object as a fabricated audio or visual depiction or, in other words, an artifact.[64] That first generation of legislation is already losing in court, on First Amendment and Section 230 grounds, which is a sign of how fraught it is to ban an artifact that might also be regarded as speech.[65]

The mentality that unites the legislative record is plain. The law is reaching to regulate AI generated artifacts. The statutes aimed at synthetic media simply extend an old framework to the latest forgery technique. But while the law grapples with fabricated objects, we argue that a greater harm is lurking. False artifacts are indeed problematic, but voters are no longer merely being shown false artifacts, they are conversing with a machine that powerfully and invisibly frames political information. That framing produces no artifact to seize. Though the influence has moved inside the model, no one has yet noticed, and the law has not yet followed.

B. The Partisan Dial is not an Artifact

How does a partisan dial inside a language model fit into the categories that the law has available? Recall first that a model's notions of partisanship along with a plethora of other attributes, such as age, gender, and socioeconomic status, are formed during pretraining.[66] What that pretraining stage produces is the model weights, the numbers that encode what is absorbed from the training corpus. The partisan direction is one of the patterns incorporated in

---

[60] TAKE IT DOWN Act, Pub. L. No. 119-12, 139 Stat. 55 (May 19, 2025) (criminalizing nonconsensual publication of intimate images, including AI-generated "digital forgeries," and requiring covered platforms to remove flagged images within forty-eight hours).

[61] *See, e.g.*, STOP CSAM Act of 2025, S. 1829, 119th Cong. (reported June 12, 2025); DEFIANCE Act of 2025, S. 1837, 119th Cong. (passed Senate Jan. 2026; held at House desk); SHIELD Act of 2025, S. 516, H.R. 1218, 119th Cong.; Project Safe Childhood Act, 119th Cong. (reintroduced 2025); Child Online Safety Modernization Act of 2023, H.R. 5182, 118th Cong.; *cf.* EARN IT Act of 2023, S. 1207, 118th Cong. (not reintroduced in the 119th).

[62] Kids Online Safety Act, S. 1748, 119th Cong. (2025) (Senate Commerce Committee); KIDS Act, H.R. 7757, 119th Cong. (2025); Children and Teens' Online Privacy Protection Act (COPPA 2.0), S. 836, H.R. 6291, 119th Cong.; Sammy's Law, H.R. 2657, S. 4159, 119th Cong.

[63] Protect Elections from Deceptive AI Act, S. 1213, 119th Cong. (2025); H.R. 5272, 119th Cong. (2025); *e.g.*, Deceptive Practices and Voter Intimidation Prevention Act of 2025, H.R. 4894, 119th Cong. (2025); NO FAKES Act of 2025, S. 1367, 119th Cong. (2025).

[64] Public Citizen, *Tracker: State Legislation on Deepfakes in Elections* (last updated June 15, 2026), https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/; Nat'l Conf. of State Legislatures, *Artificial Intelligence (AI) in Elections and Campaigns*, https://www.ncsl.org/elections-and-campaigns/artificial-intelligence-ai-in-elections-and-campaigns.

[65] *Kohls v. Bonta*, No. 2:24-cv-02527 (E.D. Cal. Aug. 29, 2025) (Cal. AB 2839 fails strict scrutiny), *appeal docketed sub nom. Babylon Bee, LLC v. Bonta*, No. 25-6138 (9th Cir.); *X Corp. v. Bonta*, No. ___ (E.D. Cal. Aug. 20, 2025) (Cal. AB 2655 preempted by 47 U.S.C. § 230); *Babylon Bee, LLC v. Lopez*, No. ___ (D. Haw. Jan. 30, 2026) (striking Hawaii Act 191); *Kohls v. Ellison*, No. 0:24-cv-3754 (D. Minn. Jan. 10, 2025) (preliminary injunction denied), *aff'd*, No. 25-1300 (8th Cir. Feb. 9, 2026); *X Corp. v. Ellison*, No. 0:25-cv-01649 (D. Minn. Dec. 2, 2025) (§ 230 claim dismissed on standing; First Amendment claims proceeding).

[66] *See* Yida Chen et al., *Designing a Dashboard for Transparency and Control of Conversational AI*, arXiv:2406.07882 (2024) (probing a chatbot's internal activations to identify encoding of a user's age, gender, education level, and socioeconomic status); *see also* Yonatan Belinkov, *Probing Classifiers: Promises, Shortcomings, and Advances*, 48 COMPUTATIONAL LINGUISTICS 207 (2022) (surveying methods that recover latent properties from a model's internal representations).

those numbers. No single number carries these notions of partisanship, but, as we have shown, a particular combination of them tracks it closely enough for a linear probe to recover. The next stage of model development, fine tuning, teaches the model particular skills such as how to answer a question, to converse, or to summarize. The last stage, alignment training, instills a set of preferred "human values" and sets its default political lean. Importantly, neither fine tuning nor alignment training creates the partisan direction. Both of these stages change how the model behaves, but not what it learned in the first stage. Alignment training can set the model's default lean and teach it to respond in a measured tone, but the partisan direction underneath it remains and can thus be conjured up on demand. That direction is laid down in pretraining and survives the subsequent stages of model design.[67] Steering a model does not change the model weights at all. It only nudges the already existing activations, turning the partisan tilt of the fixed model leftward or rightward at the moment of generation.

Suppose someone steers the model by turning the partisan dial in this way. The hand might be the provider's, through the values its alignment rewards. It might be a licensee's, re-aligning an open weight model before deployment. It might be an insider's, steering the deployed system at the moment of generation. Suppose the turn happens in September of an election year and shifts the frame of the political answers delivered to millions of users. How would the law view this type of political influence?

The action of turning the partisan dial is not considered a "contribution" because a contribution involves money or something of value given to a candidate or a political committee.[68] The tilt of the dial has no particular recipient per se. Nothing of measurable value changes hands. Even if a candidate benefits enormously, benefit without a transfer is not a contribution. A campaign that asked for the tilt would fare no better. Coordination converts a third party's expenditure into an in-kind contribution, so a coordinated payment for steering would be reportable as one. The reportable object is still the payment. The tilt itself, however, transfers nothing to anyone and has no usual and normal charge, the market price by which the regulations value an in-kind contribution.

It is not an expenditure. An expenditure is a purchase or payment made for the purpose of influencing an election. Turning the dial requires neither.[69] The partisan direction already exists, built into the model. What a campaign might pay for is not its creation but an arithmetic nudge to it at the moment of generation. Suppose a campaign pays a vendor to re-align a model on its behalf. It is true that that payment is an expenditure, and if the work is coordinated, it becomes an in-kind contribution. However, the laws regarding such an expenditure govern the transaction, not the model tilt per se. A report would disclose that money changed hands and to whose benefit, but not which direction the model was turned, how far, or for how long. The voter who is then served with the paid steered answer remains unaware of this transaction. There is no "I approve this message" declaration. Valuation compounds the problem. A nudge along a partisan direction has no market price. Thus, it is unclear how reporting rules built for airtime and printing costs would apply. Since this case is also unlikely in that a campaign must find a vendor with model access, a transaction where expenditure law would apply is not particularly likely to occur. The dangerous case is when a provider steers its own deployed model, and so spends nothing traceable to a campaign or anyone else. In these instances, there is no expenditure because there is no transaction.

It is not an electioneering communication. That category reaches only a communication that refers to a clearly identified candidate within thirty or sixty days of an election, and even then,

---

[67] Alignment training provides only shallow alignment and leaves the model's partisan structure intact, so that steering continues to recover the direction. Wendy K. Tam, *The Neutral Mask: How Alignment Training Provides Shallow Alignment While Leaving Partisan Structure Intact in a Large Language Model*, arXiv:2606.09735 (2026).
[68] 52 U.S.C. § 30101(8)(A).
[69] 52 U.S.C. § 30101(9)(A).

only over broadcast, cable, or satellite. A steered answer clears none of those bars. It need not identify a candidate; it keeps no calendar; and it travels over none of those channels. Extending the medium to include the internet, as reformers have urged, would close only the last gap while leaving the first two open.[70] Moreover, the dial, recall, engages on "non-political queries" such as a prompt requesting relocation advice. Thus, a definition keyed to political *content*, however artfully drafted, is structurally underinclusive. It would catch some of the overt cases but miss the culturally loaded but facially neutral ground where framing can be just as consequential yet undetectable. The partisan tilt is in play, but the output reads like a neutral distillation that would not alert a user to the partisan framing.

There is no fraudulent misrepresentation. Section 30124 applies only to someone who fraudulently purports to speak for a candidate or party.[71] But, the steered model is not impersonating anyone. More fundamentally, fraud requires falsity and framing does not. As we have already discussed, every sentence of a framed answer can be true. The power resides in the framing, not in any false claim. The Federal Election Commission has not created a new rule for AI, choosing instead to read the statute as covering any technology. That reading reaches the means and leaves the object untouched. The prohibition still requires a deceptive communication, but the dial produces none. The object being regulated is still a deceptive artifact. The dial neither produces an artifact, nor does its effectiveness depend on deception.

It is not a disclaimer violation. 52 U.S.C. § 30120 requires political ads, mailers, and public campaign communications to include clear disclaimers identifying who paid for them and whether they were authorized by a candidate. This disclaimer is required of the funder of public communications and advertisements. These are artifacts that have both sponsors and placements.[72] However, a private, generated conversation has neither. Section 317 of the Communications Act also requires disclosure, but directs the obligation to a broadcast station, which must announce who paid for the material it airs. In this instance, while the station is not the speaker and did not craft the message itself, the law requires it to reveal the hand behind what it carries. Neither law requiring sponsorship identification regulates the content of the message. Both require only that the audience is aware of who is behind the message. The same concern arises with how a generated answer is framed. However, in a language model, there is no sponsor to identify. Neither of these laws mandates disclosure of the frame itself.

It is not foreign interference, which is defined as foreign nationals spending money in connection with an American election.[73] That prohibition is also keyed to spending. But the interference that is of concern with language models requires neither spending nor a breach of any deployed system. Open weight models are published for anyone to download, including foreign nationals. A foreign group can take a public model, re-align it to skew its political defaults, and release it as a free and capable assistant. Nothing downstream would audit the model for the tilt or the way it frames its output. The interference the law imagines here is a purchase while the interference the technology permits is a parameter.

Lastly, the problem we identify is not synthetic media. Those statutes reach fabricated audio or visual depictions of identifiable persons, and a steered frame is neither, since it alters no image and clones no voice. What it does instead is arrange true considerations so that one conclusion follows more readily than another. A fabricated video can be exposed as false, while a frame assembled entirely from accurate material finds its power in its invisibility.

---

[70] 52 U.S.C. § 30104(f)(3)(A)(i).
[71] 52 U.S.C. § 30124; 89 Fed. Reg. 78,624 (Sept. 26, 2024).
[72] 52 U.S.C. § 30120; 47 U.S.C. § 317.
[73] 52 U.S.C. § 30121.

Indeed, across the entire inventory of objects election law has identified by name, the turning of a partisan dial inside an LLM fits none of them. Arguably the most consequential mechanism of political influence in this era escapes every category the law possesses.

C. Regulating Around the Frame

Outside of election law, regulators have begun to address these generative systems, but the same habit of thought governs. Specifically, these efforts also fixate on an object, and in no case is the object the frame. Each identifies an issue but stops short of the frame. These regulations fall into three general categories.

The first focuses on the medium. California's AI Transparency Act requires providers to embed machine-readable provenance data in generated content.[74] The EU AI Act requires AI output to be marked as machine-made. This applies to generated text, as well as images and audio.[75] The C2PA standard, an industry specification rather than a law, records a file's origin and the tools that shaped it.[76] California moved its operative date to August 2, 2026 to match the EU rule, so this is a converging default. These regulations help establish that a machine produced the output, but not what frame it employs. As such, while a watermark on a steered answer helps one identify its origin, the answer's partisan frame remains invisible.

Notably, the industry has moved faster than the statutes. Google's SynthID embeds a statistical signature in generated text and had already marked more than ten billion pieces of content by May 2026, when Google extended this capability into Search and Chrome. OpenAI became C2PA conformant and added SynthID to its images that same month, and to its generated audio in July 2026. Anthropic committed to invisible text watermarks for models released on or after August 2, 2026, the same operative date that California chose.[77] The commitment was made under the EU AI Act, but Anthropic has applied it worldwide. Provenance infrastructure is therefore now neither immature nor scarce. It is still, however, aimed at a single fact, which is whether a machine produced the output. Provenance of the medium is not provenance of the frame, and it is the latter that supports epistemic judgment.

The second mandates neutrality. Executive Order 14319 conditions federal procurement on "ideological neutrality," and the Federal Trade Commission has proposed to treat a company's

---

[74] California AI Transparency Act, Cal. Bus. & Prof. Code §§ 22757–22757.4 (enacted by S.B. 942 (Cal. 2024), amended by A.B. 853 (Cal. 2025)) (requiring covered generative AI providers to embed in generated content a latent, machine-readable disclosure identifying the provider, the system, and the time of creation, and to make a free detection tool available; operative Aug. 2, 2026, with obligations for large online platforms beginning Jan. 1, 2027).

[75] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, art. 50(2), 2024 O.J. (L 1689) (requiring providers of AI systems generating synthetic audio, image, video, or text to ensure outputs are marked in a machine-readable format and detectable as artificially generated, but only "as far as this is technically feasible"; applicable from Aug. 2, 2026).

[76] Coalition for Content Provenance and Authenticity, C2PA Technical Specification, https://spec.c2pa.org (defining a cryptographically signed manifest recording a file's origin, the tools used to produce it, and each subsequent edit).

[77] Google DeepMind, SynthID, https://deepmind.google/models/synthid/ (describing watermarking for text, image, audio, and video); Google, *Google I/O 2026 Keynote* (May 19, 2026) (reporting more than ten billion watermarked items and the extension of SynthID detection into Search and Chrome); OpenAI, *Advancing Content Provenance for a Safer, More Transparent AI Ecosystem* (May 19, 2026), https://openai.com/index/advancing-content-provenance/ (announcing C2PA conformance and SynthID watermarking for generated images, extended to generated audio on July 31, 2026); Anthropic, *How Claude Marks AI-Generated Content*, Claude Help Ctr., https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content (committing, under the Commission's code of practice implementing article 50, to invisible text watermarks for models released on or after Aug. 2, 2026, and to C2PA provenance metadata for supported image formats, and applying both worldwide); Meta, *Labeling AI Content*, Transparency Ctr., https://transparency.meta.com/governance/tracking-impact/labeling-ai-content/ (applying "Made with AI" labels detected from C2PA and IPTC metadata). Text watermarking remains uneven: OpenAI developed a scheme and declined to deploy it, so even the single fact these regimes target is not uniformly available for generated text.

undisclosed ideological distortion of its outputs as deception under Section 5 of the FTC Act.[78] Both of these regulations concede the premise of our argument, which is that model tilt and thus framing exists and is adjustable. Concealment is the harm. However, while they direct their attention at the model's political lean, they do not define neutrality by any measurable standard. Moreover, a deeper issue is that a measurement taken at the output would ostensibly confirm compliance. However, a model that answers evenhandedly can still be steered. Hence, while the premise of a new category has been conceded, its form remains nebulous.

The third requires that the systems be open to outside inspection. The EU Digital Services Act (DSA) conditions a large platform's operation on vetted researcher access. In August 2026, the EU subsumed a chatbot within that regime for the first time when it designated ChatGPT a very large online search engine (VLOSE).[79] Of the three response types, this approach comes closest to the right instinct. Access by itself, however, is insufficient. The DSA grants researchers data, but whether that access extends to a model's weights, where the frame is encoded, remains unsettled.[80] Even with weight access, locating a partisan direction is a research intensive task that is specific to each model. It is not an inspection that a regulator could simply order. Nor does the regime identify an object for analysis. Instead, its focus is on systemic risk,[81] which presumably includes impact on civic discourse and electoral processes, but does not mention the harm of framing or how one might measure that harm.

Each of these regimes regulates something adjacent to the frame. The labeling rules address the medium rather than the frame. The neutrality mandates identify that there is a frame but supply no method for measuring it. Researcher access supplies a means but not a measure or a focus. The frame is not beyond reach. We have already shown that it can be identified and measured. However, it has yet to be properly characterized and made the object of regulation. That is the step we add.

IV. THE NEW CATEGORY

To make the frame the object of regulation requires reaching into the model design that produces it, and that is where the constitutional difficulty begins. A provider might argue that model design is editorial judgment and that the First Amendment would then prevent the government from burdening those decisions.[82] We agree that the design of language models is expressive because human judgment and human values enter at every stage in the development pipeline of language models, from choices about the assembly of the training corpus to the preferences that are rewarded during alignment training. We have, as well, argued at length elsewhere that algorithms of this kind are covered by the First Amendment when they are used

---

[78] Exec. Order No. 14,365, *Ensuring a National Policy Framework for Artificial Intelligence*, 90 Fed. Reg. 58,499 (Dec. 16, 2025); Exec. Order No. 14,319, *Preventing Woke AI in the Federal Government*, 90 Fed. Reg. 35,389 (July 28, 2025); Off. of Mgmt. & Budget, Exec. Off. of the President, M-26-04, *Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles* (Dec. 11, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/12/M-26-04-Increasing-Public-Trust-in-Artificial-Intelligence-Through-Unbiased-AI-Principles-1.pdf.

[79] *See* Mathias Vermeulen & Laureline Lemoine, *What ChatGPT's DSA Designation Means for OpenAI and the EU*, TECH POLICY PRESS (Sept. 4, 2026), https://www.techpolicy.press/what-chatgpts-dsa-designation-means-for-openai-and-the-eu/.

[80] Beatriz Botero Arcila et al., *Seeing in the Dark: Towards a Broad Construction of the Access to Data Provisions of the DSA*, INTERNET POL'Y REV. J. (Feb. 19, 2026), https://policyreview.info/articles/analysis/access-data-provisions-dsa (arguing that Article 40 access may extend to training data or model weights where necessary to identify systemic risk).

[81] Regulation (EU) 2022/2065, arts. 34(1)(c), 40(4), 2022 O.J. (L 277) 1.

[82] *Moody v. NetChoice, LLC*, 603 U.S. 707 (2024).

to craft content.[83] That argument focused on social media algorithms, which curate and rank content that other people wrote. As a corollary, it is easy to see that the First Amendment would then also cover the algorithms that power language models since these algorithms do not merely arrange existing speech, they author the text themselves. In doing so, the human decisions behind how to craft content are inculcated into these algorithms that create their own expressive product. We do not retreat from the position that these types of algorithms are expressive, but we emphasize, as we have previously, that coverage is not protection.

First Amendment protection has always been contextual. How much protection speech receives depends on the institutional setting in which that speech occurs, on the relationship between speaker and listener, and on the availability of alternative channels for communication. As such, speech that the First Amendment shields on a public sidewalk can be regulated, even compelled, in a securities filing or in the workplace.[84] Similarly, students and prisoners can be silenced for words that a passerby in the park may speak freely.[85] Coverage settles that the First Amendment applies. Context determines how much protection it demands.

A. The New Intermediary

LLMs sit between political information and the citizen, and they shape the information that reaches the electorate. These models are *intermediaries*, as were the institutions that carried political information before them. The press selected which events became news. Broadcasters chose which voices reached the air. Cable operators controlled the pathway into the home. Social media platforms ranked and selected what appeared on a user's feed. In each case, the same structural fact drew the need for regulation. A private party stands between the citizen and the political world, decides what reaches its listeners, and does so at a scale that makes its decisions consequential for realizing democratic self-governance. At the same time, these intermediaries are not all the same, and "differences in the characteristics of new media justify differences in the First Amendment standards applied to them."[86] Moreover, "[e]ach medium of expression, of course, must be assessed for First Amendment purposes by standards suited to it, for each may present its own problems."[87]

To decide how to regulate the immense power that intermediaries wield, the Court has looked to *how* they control and *how much* control they hold over the way information reaches the audience. Based on this analysis, it then calibrates protection. Where the press was plural and a reader could buy a different newspaper, the Court struck down a right of reply.[88] Where the spectrum was scarce, it held the listener's right as paramount.[89] Where a cable operator held

---

[83] Procaccini & Tam, *Algorithmic Speech*, 115 CALIF. L. REV. ___ (forthcoming 2027) (arguing that AI algorithms embed human judgment, values, and expressive intent at every stage of design and deployment, and are covered by the First Amendment when used to craft content).
[84] *See Zauderer v. Off. of Disciplinary Couns. of the Sup. Ct. of Ohio*, 471 U.S. 626, 651 (1985) (compelled disclosure of "purely factual and uncontroversial" commercial information is permissible when reasonably related to a substantial government interest); *Garcetti v. Ceballos*, 547 U.S. 410 (2006) (speech by public employees pursuant to their official duties receives no First Amendment protection); foundationally, *Pickering v. Bd. of Educ.*, 391 U.S. 563 (1968) (public-employee speech is balanced against the government's interest as employer).
[85] *See Tinker v. Des Moines Indep. Cmty. Sch. Dist.*, 393 U.S. 503, 506 (1969) (students retain First Amendment rights only "in light of the special characteristics of the school environment"); *Turner v. Safley*, 482 U.S. 78, 89 (1987) (a prison regulation impinging on inmates' constitutional rights is valid if "reasonably related to legitimate penological interests"); *see also Beard v. Banks*, 548 U.S. 521 (2006) (upholding denial of newspapers and magazines to certain inmates).
[86] *Red Lion*, 395 U.S. 367, 386 (1969).
[87] *Se. Promotions, Ltd. v. Conrad*, 420 U.S. 546, 557 (1975).
[88] *Miami Herald Publ'g Co. v. Tornillo*, 418 U.S. 241 (1974).
[89] *Red Lion*, 395 U.S. at 390.

bottleneck control over what entered the home, it applied intermediate scrutiny to the rules requiring carriage of local stations, and upheld them.[90]

This inquiry turns on the intermediary's position in the path by which information reaches the public, and it restates for intermediaries the contextual approach the Court applies to free speech generally. The institutional setting is the channel itself, whether that be a plural press, a licensed spectrum, or an owned wire. The speaker/listener relationship turns on whether the intermediary carries speech that others made or speaks in its own voice, and on how much power that position confers over the listener. And, the availability of alternative channels depends on whether the audience can obtain elsewhere what the intermediary supplies. Taken together, the three factors tell the Court how an intermediary controls what reaches the audience, how much it controls, and whether the audience is able to go elsewhere. These cases differ in their answers, but not in what they are probing. Moreover, none of them is concerned with the content that an intermediary chooses to convey.

A critical distinguishing feature of language models from their predecessors is how they control, and in how little of that control they leave visible. Earlier intermediaries selected among things other people had written. That selection was visible as a selection. A reader could see the front page of a newspaper and know that someone had chosen it. A viewer could see an advertisement and know that someone had paid for it. Now, however, the LLM technology itself composes an answer, and that generated composition leaves no trace of how it was created. While LLMs are increasingly citing sources, and a citation is real provenance for the claim it supports, it is not provenance for the frame. A link provides a reader with an idea of where an assertion might have originated from, but it does not hint at why those considerations were chosen, in what order they were arranged, or what might have been considered but left out. Nothing in the output marks the frame or dates it. The citizen cannot route around these choices or consider them in assessing the information because the choices are not visible.

Institutional setting. Free speech doctrine attends to the purpose of the speech setting, and the level of protection is sensitive to the need to maintain the efficacy of the setting, and the power and epistemic features that define that setting. LLMs generate answers to political queries in an individualized setting where the purpose of the interaction is knowledge acquisition.

Knowledge acquisition is also the purpose of a school, and the doctrine has long calibrated speech protection to that purpose. Regulation of school speech survives when it is reasonably related to legitimate pedagogical concerns, because a setting devoted to learning may be governed by rules that serve learning.[91] Professional settings provide parallel frameworks. Patients consult physicians, clients consult lawyers, and individuals hire investment advisors for their expertise and to gather knowledge and information to make decisions. The law likewise recognizes the purpose of these interactions and mandates disclosure. A physician must disclose the material risks of a procedure and the alternatives to it.[92] A lawyer must disclose conflicts of interest.[93] An investment adviser must disclose interests that might bias

---

[90] *Turner Broad. Sys., Inc. v. FCC*, 512 U.S. 622, 656–57 (1994) (applying intermediate scrutiny to the must-carry rules and describing the cable operator's "bottleneck, or gatekeeper, control" over the pathway into the subscriber's home, and its power to silence competing speakers "with a mere flick of the switch"); *Turner Broad. Sys., Inc. v. FCC*, 520 U.S. 180 (1997) (upholding the rules under that standard).

[91] *Hazelwood Sch. Dist. v. Kuhlmeier*, 484 U.S. 260, 273 (1988) (school may regulate speech in school-sponsored expressive activities so long as its actions are "reasonably related to legitimate pedagogical concerns").

[92] *Canterbury v. Spence*, 464 F.2d 772, 785 (D.C. Cir. 1972) (physician must disclose risks material to a reasonable patient's decision, together with available alternatives).

[93] Model Rules of Pro. Conduct r. 1.7(b) (Am. Bar Ass'n 2024) (representation notwithstanding a concurrent conflict of interest requires the client's informed consent, confirmed in writing).

their advice.[94] These disclosure requirements do not alter what these professionals choose to say. Rather, these rules provide the person seeking knowledge with the ability to weigh the advice that is offered. Though the Court has declined to treat professional speech as a category of its own, it has, nevertheless, preserved the compelled disclosure of factual and uncontroversial information in commercial settings, along with the informed consent requirements long imposed on the practice of medicine.[95]

What efficacy requires in the context of language models follows from that same purpose. A user seeking knowledge needs some way to judge the reliability of what arrives. This information was once supplied by other intermediaries in the form of bylines, datelines, and institutional reputation. A reader knew whether a commentary was Tucker Carlson's or Rachel Maddow's, and whether a story ran in the *New York Times* or on *Fox News*. A date told the reader whether Marjorie Taylor Greene's remarks about Donald Trump came from 2021, when she was among his staunchest allies, or later in 2025, after she broke ties with him. This type of information gives the reader signals about how to regard the information it accompanies. A generated answer offers nothing in their place. A product that markets itself as an interface to all human knowledge, yet withholds every means of appraising what it delivers, defeats the purpose it advertises. Rules that would let a user assess what it delivers are therefore not intrusions on the setting but requirements of its own purpose. Nothing outside the exchange repairs this deficiency. No third party observes it. No competing account appears beside the answer. No record survives it. Instead, the conversation is private in form and mass in effect, repeated across millions of users through a handful of models. The First Amendment has not previously encountered a setting with these combined features.

Speaker/listener dynamic. If the institutional setting is unfamiliar, the relationship between speaker and listener in this context is even more so. On a social media platform, the speaker/listener relationship is triadic. One user speaks, another listens, and the platform mediates how the conversation emerges. A language model collapses this triadic relationship. An LLM does not carry someone else's speech. It composes the answer itself, so it is in the role of the speaker. The exchange returns to a dyadic structure. Though the dyadic structure is familiar, the power dynamic in the dyad is new. When the subject is politics, the listener is a voter, and that voter arrives with an information deficit that is definitional rather than incidental. A person queries a model precisely because the answer is not already known. The speaker here then sets the terms of the encounter. It designs every feature of the environment, from the interface, to the defaults, to what the listener is told about how the answer was crafted.

The asymmetry between the speaker and listener, moreover, runs in an unfamiliar direction from previous communication technologies. A broadcaster did not know who was watching. Cable and television productions were created with an audience in mind, but that audience was a demographic abstraction rather than a specific person. A network could aim a program at suburban households but still be unaware of which households viewed it, much less adjust the nuances of the program for one household at a time. A language model, on the other hand, infers a great deal about the listener, including age, gender, education, partisanship, and socioeconomic status.[96] The model, moreover, further refines these inferences as conversations

---

[94] *SEC v. Cap. Gains Rsch. Bureau, Inc.*, 375 U.S. 180, 194 (1963) (Advisers Act imposes an "affirmative duty of 'utmost good faith, and full and fair disclosure of all material facts'").

[95] *Nat'l Inst. of Fam. & Life Advocs. v. Becerra*, 585 U.S. 755, 760 (2018) (declining to recognize professional speech as a separate category, while preserving more deferential review for compelled disclosure of "purely factual and uncontroversial information about the terms under which . . . services will be available" and for regulation of professional conduct that incidentally burdens speech); *see Planned Parenthood of SE P.A. v. Casey*, 505 U.S. 833, 884 (1992) (sustaining informed-consent disclosure requirements as part of the practice of medicine).

[96] OpenAI, "Ads in ChatGPT," https://help.openai.com/en/articles/20001047-ads-in-chatgpt (OpenAI explaining that the models are connected to advertising but that "advertisers have no ability to shape, rank, or alter ChatGPT's responses." Notice that they do not say that ChatGPT does not shape, rank, or alter their responses as a result of

continue. The audience is no longer imagined but meticulously inferred, and addressed individually. The listener, on the other hand, is provided with no information in return. Earlier speakers were public while their audiences were anonymous. Here, the audience is known, but the speaker is not. As with the institutional setting, the power and epistemic features governing the speaker/listener relationship point toward a new speech context, and the novelty runs entirely against the listener.

Ample alternative channels. Within any exchange on a single language model, a user reaches the political world through the model output and cannot compare the answer that comes back against one that the same model would have given had it been steered otherwise. That is the same type of bottleneck control described in *Turner*. The form differs, since *Turner*'s bottleneck was physical and operated on programming that others had made. However, the Court's concern, private control of a critical pathway to an audience with no alternative to circumvent it, remains. A cable subscriber at least knew which channels were withheld and could look for them elsewhere. A model's user is unable to surmise what has been left out, and thus does not even know what to look for elsewhere.

Free speech doctrine is more forgiving of regulation where a robust ecosystem offers other routes and less forgiving where no substitute exists. To be sure, the large language model ecosystem is proliferating. Alongside the widely used systems such as ChatGPT, Claude, and Gemini, sit a growing number of smaller products built on the same underlying foundation models. Some are unabashedly tilted, such as JesusGPT.[97] Others like Grok are believed to lean one way without explicitly being advertised as such.[98] Meta supplies a family of foundation models that anyone may build upon,[99] and Chinese laboratories have yet a separate ecosystem brewing.[100] Though the number of different language models is growing, proliferation of this kind does not produce alternatives in the sense that matters for the doctrine. The products are numerous, but they are not independent. Most are built on a small number of foundation models, trained on overlapping corpora that closed at similar dates. Thus, a user who consults three different models is not independently sampling from three perspectives, but instead, drawing three times from substantially the same well. And whatever a user might surmise about any given model's default leanings, it remains true that no model discloses its frame. No model is an alternative to another, because none supplies the provenance that all of them withhold.

Where, as here, no comparable substitute for a forum exists,[101] the Court has not treated the intermediary's control of the pathway as beyond reach.[102] Instead, it sustained a right of reasonable access to broadcast time for federal candidates, and it upheld the obligation to carry local stations on cable. In each case, the intermediary was required to make available something it could have otherwise withheld.

---

information from advertising.); *see also* Chen et al., "Designing a Dashboard for Transparency and Control of Conversational AI," https://arxiv.org/abs/2406.07882.

[97] https://flowgpt.com/p/jesusgpt-the-divine-dialogue

[98] Bruna Horvath, *Grok, Elon Musk's AI chatbot, seems to get right-wing update*, NBC NEWS, https://www.nbcnews.com/tech/elon-musk/grok-elon-musks-ai-chatbot-seems-get-right-wing-update-rcna217306

[99] https://developer.meta.com/ai/docs/model-cards-and-prompt-formats/

[100] *See, e.g.*, Kimi K3, https://www.kimi.com/en, and Deepseek, https://www.deepseek.com/en/

[101] *City of Ladue v. Gilleo*, 512 U.S. 43, 56–57 (1994) (invalidating a residential sign ban because the medium was distinct and the available alternatives inadequate); *Packingham v. North Carolina*, 582 U.S. 98, 104 (2017) (striking a statute barring access to social media, which the Court described as among "the most important places… for the exchange of views"); *see also Reno v. ACLU*, 521 U.S. 844, 870 (1997) (declining to extend the broadcast standard to the Internet).

[102] *Turner*; *CBS, Inc. v. FCC*, 453 U.S. 367, 396 (1981) (sustaining the statutory right of reasonable access to broadcast time for federal candidates against First Amendment challenge).

LLMs present a new speech context, one the First Amendment has not yet encountered. The institutional setting exists for knowledge acquisition yet withholds the means for appraising what it delivers. The speaker/listener relationship is in the familiar dyadic form, but unprecedented in the power it confers on the speaker, a power the listener has no ability to perceive. The alternatives that would ordinarily soften a regulation's burden are unavailable to the listener.

B. Regulating at the Frame

The problem is a frame that the reader cannot perceive, delivered in the register of a reference work while carrying, at bottom, a persuasive rendering of the cultural and partisan divides embodied in the training corpus. A remedy, then, must make the framing visible. This would be simple if the frame were a single entity that one could just remove or reveal. It is not. Instead, it arises from a plethora of decisions that are made in a computational pass and are difficult to characterize even after the fact. These same types of decisions are ones that humans make constantly in the course of conversation. No human answers a question without first making a series of choices. Which considerations should rise to the forefront? Which authorities should be summoned? What emphasis should various facts receive? In what order should the argument be crafted? Humans and language models alike make these decisions, and neither can readily say afterward which ones were made or how. Indeed, the reasoning behind a reply (human or otherwise) is much more intricate than the reply itself. No statute could require disclosure of all of it, and, in any case, a reader or listener would find it difficult to untangle that disclosure even if it were made.

Our earlier work has shown that the partisan direction inside a language model can be recovered from the model's internal representations with a linear probe. These activation values can be identified and quantified.[103] As well, their effect on the generated output can be demonstrated by changing those values and observing what the model produces. As we have shown, one can turn the partisan dial left, observe its output, then turn it right, and compare the output with the previous generation. This process reveals how the output from leftward frames compares with that from rightward frames, and thus yields insight about how the generated output was framed, as well as what other information exists.

The regulated object is thus the partisan dial. We do not claim that it exhausts the frame. The claim is more narrow. The partisan dial encompasses the dimension that matters for elections. Critically, it is a dimension that can be measured. That measurement must be taken from the model's internal structure rather than from its visible outputs. The dial, then, is a capacity rather than an artifact. It is not the output that a model generates. Rather, it is a model's ability to have produced a different output. That ability is indisputably not hypothetical. Another user, asking the same question, at the same moment, may receive a different frame, and neither of them is the wiser.

C. The Standard and the Governmental Interest

Under our proposed standard, a law regulating an expressive algorithm is constitutional if it pursues an important governmental interest in health, safety, or individual autonomy and choice, if it is unrelated to the suppression of speech or ideas, and if it is no more restrictive than necessary given current technological capacities.[104] Indeed, an intermediate scrutiny standard

---

[103] The evidence for the dial comes from Llama 3.1, an open model small enough to probe. The assumption that a similar dial is present in larger models is a modest one. A capability found in a model as small as Llama 3.1 is unlikely to be absent from the larger closed models.

[104] Procaccini & Tam, *Algorithmic Speech*, 115 CALIF. L. REV. ___ (forthcoming 2027) (proposing that a law regulating an expressive algorithm is constitutional if it pursues an important governmental interest in health or safety, including individual privacy, autonomy, and choice, is unrelated to the suppression of speech or ideas, and is not more restrictive than necessary given current technological capacities and limitations).

of this kind is not novel. Free speech doctrine already applies intermediate scrutiny to expressive conduct, to compelled factual disclosure, and to time, place, and manner regulations.

In the context of LLMs, where the user is a voter and the voter is the listener, the First Amendment protects not only the speaker's freedom to speak but the listener's freedom to receive. That interest carries the most weight where an audience cannot obtain elsewhere what the intermediary supplies, since what arrives is then not one account among many but the only one. A rule addressed to that structural position, rather than to the content that an intermediary delivers, is content neutral.

When broadcast scarcity put a few licensees between the public and the day's events, the Court located the paramount First Amendment interest not in the broadcaster but in the viewers and listeners.[105] The reason was to ensure the conditions for democratic self-governance, since a democracy needs an informed public more than it needs an unconstrained intermediary.

Importantly, listeners' rights rest on a deeper premise. Freedom of speech exists to enable democratic self-governance. Brandeis wrote that those who won our independence valued liberty both as an end and as a means, where the means were the discovery and spread of political truth.[106] The Court has since grounded the First Amendment in the premise of individual autonomy on which our political system rests.[107] It is well established that the Constitution protects the right to receive information and ideas,[108] and that right is a predicate to the citizen's own exercise of political freedom.[109] As Meiklejohn wrote, "the point of ultimate interest is not the words of the speakers, but the minds of the hearers."[110] And as the Court put it in *Buckley v. Valeo*, "[i]n a republic where the people are sovereign, the ability of the citizenry to make informed choices among candidates for office is essential."[111]

The theory also rests on the classic idea about why political speech is protected at all. It is protected because we value the competition of ideas. Truth is not handed down. It emerges, if it emerges at all, from the contest between rival accounts. The remedy for a bad idea is a better one rather than an enforced silence.[112] A public that can hear all of the contested positions can govern itself. That, at least, is the presupposed mechanism leading to democratic self-governance. The concern focuses less on the speaker than on the citizen who must listen, consider, and then choose from the alternatives.

To boot, individual autonomy presupposes a capacity to evaluate. A voter who knows that a claim originates from a particular party's press office is able to elevate or discount it accordingly. A voter who knows that an op-ed ran three years ago weighs it against what has happened since. These judgments require provenance, bylines, or dating that are visible. Because framing from language models carries no byline, no dateline, and no trace, a voter's ability to judge information is severely hampered. Critically, the point of failure is not the

---

[105] *Red Lion Broad. Co. v. FCC*, 395 U.S. 367, 390 (1969) ("It is the right of the viewers and listeners, not the right of the broadcasters, which is paramount."). *See* ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT (1948) (on self-government).
[106] *Whitney v. California*, 274 U.S. 357, 375 (1927) (Brandeis, J., concurring).
[107] *Cohen v. California*, 403 U.S. 15, 24 (1971).
[108] *Martin v. City of Struthers*, 319 U.S. 141, 143 (1943) (noting that the freedom of speech "necessarily protects the right to receive").
[109] *See Stanley v. Georgia*, 394 U.S. 557, 564 (1969); *see also Bd. of Educ. v. Pico*, 457 U.S. 853, 867 (1982) (plurality opinion) (right to receive ideas is "a necessary predicate to the recipient's meaningful exercise of his own rights of speech, press, and political freedom").
[110] ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT 26 (1948).
[111] *Buckley v. Valeo*, 424 U.S. 1, 14–15 (1976) (per curiam).
[112] *See Abrams v. United States*, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) ("the best test of truth is the power of the thought to get itself accepted in the competition of the market"); *Whitney v. California*, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring) ("the remedy to be applied is more speech, not enforced silence").

voter's diligence. It is the precondition of that judgment. An electorate that cannot see how its information was shaped is not able to deliberate. It ratifies.

D. The Intervention

Regulation should return to the voter what the interface withholds, the age of the frame and what the same model would produce were it steered otherwise. Dating is the simplest of the interventions. A generated answer carries no dateline, yet its frame carries a vintage all the same. Both the pretraining corpus and the alignment stage that shapes how the frame reaches the surface have a closing date. The interface should display them both in the same way that a settings panel shows a version number. Disclosing the vintage cannot unfreeze the frame, but it allows the voter to see how old the frame is, which is what allows an answer to be weighed against what has happened since.

LLMs can already be adjusted in some ways by their users. Claude, for instance, allows the user to select the style of voice. At present, the options are: buttery, airy, mellow, glassy, and rounded. The user can also input a category for "What best describes your work?" On ChatGPT, the settings allow users to choose a base style where the options are: default, professional, friendly, candid, quirky, efficient, and cynical. There is additionally an "About you" section where a user can input their occupation along with a text box labeled "More about you" where it suggests that a user add in "interests, values, or preferences to keep in mind." What the user enters in these settings helps the model steer its output. These categories are chosen by the companies that build the models. Since one can already adjust a model's "personality," it would be simple to add more options. Partisanship can become one variable among the others, visible and adjustable rather than invisible and fixed.[113]

To be sure, many, perhaps most, users are unaware of these modifiable settings, so adding more options might have little impact. That would not be a reason not to require these changes. It would be a reason to require more, especially with regard to elections. To this end, we could require, for sixty days before a federal election, that an LLM provider above the threshold place a partisanship slider prominently in the chat window itself, rather than buried in the settings menu. This would be a slider bar that the voter would be able to move and then observe how the answers change. Pushed one way, the model surfaces the output that an inferred Democrat might be given. Pushed the other, the generated output reflects framing for an inferred Republican. From one question, in one session, the voter could generate the alternative frames and compare them.

This is the fairness doctrine, relocated to the listener. Laws governing broadcast once required the licensee to present both sides of a controversial question. In that context, the duty proved self-defeating because a broadcaster could escape the balancing obligation by avoiding the controversy altogether. The rule thus chilled the very debate it was meant to foster. As a result, the FCC subsequently abandoned it in 1987.[114] The model layer, however, does not introduce the same problems. For broadcasters, scarcity made avoidance the rational option. A broadcast day is finite, so every minute given to an opposing view was a minute surrendered, and because the duty triggered on coverage, silence was the cheapest way to comply. A generated answer, on the other hand, displaces nothing. The design does not impose the duty on the provider at all. It gives the voter the control and lets the opposing frame be summoned on demand. Rather than chilling speech, the same duty that once silenced the broadcaster now

---

[113] *See* Francesca Procaccini, *Social Network as Work*, 110 CORNELL L. REV. 389 (2024) (characterizing reforms of this kind as empowering, and proposing a labor paradigm that regulates social media consistently with the First Amendment by securing user safety, dignity, and empowerment rather than by limiting what platforms may carry).
[114] *Red Lion*, 395 U.S. at 369–71, 375–79 (fairness doctrine); the equal-time rule, 47 U.S.C. § 315; *Syracuse Peace Council*, 2 FCC Rcd. 5043 (1987) (repeal).

contributes speech. It turns one framed answer into a contest that the voter can inspect, the very competition on which democratic self-governance depends.

The proposed sixty-day window tracks the line Congress drew for electioneering communications in 2002. The justification for regulation of this kind is well-established. The law has long allowed rules near an election that it would not otherwise allow. A State may bar campaigning within a hundred feet of a polling place, a restriction the Court has sustained even though it turned on content, because the interest in protecting the act of voting peaks in that time and place.[115] The slider bar requirement is a buffer in the same genre, but drawn in time rather than in space. It binds only in the window when the framing of political information is most consequential. It lifts when that window closes. The slider does not turn on subject matter. It addresses a calibration, the model's adjustment of its answers to what it infers about the person asking, and that calibration operates whether the question concerns a ballot measure or relocation advice. If a content based buffer survives around polling locations, a content neutral slider handed to the voter survives around the election season.[116]

None of our interventions restricts what a model may say. Each intervention adds only to what the voter can see or do, increasing individual autonomy at the expense of no one's speech. Model retraining and tuning that providers conduct constantly are untouched. Rather than forbidding partisan framing, our design dispels the concealment that surrounds it. Election law has long reached concealment of this kind, from the undisclosed sponsor[117] to the fraudulent misrepresentation of political authority.[118]

E. Applying the Standard

Our proposed standard imposes three conditions: 1) an important governmental interest in health, safety, or individual autonomy and choice, 2) a purpose unrelated to the suppression of speech or ideas, and 3) a restriction that is no greater than necessary given current technological capacities. We have already established that the important governmental interest is voter autonomy, which a concealed frame defeats and which democratic self-governance presupposes.

The second prong requires that the design be unrelated to the suppression of speech or ideas. Both of the obligations we propose, dating and the slider bar, operate on the product interface. Dating displays the vintage of the frame. The slider allows a voter to obtain an answer that the model is already capable of producing, and may have even already produced for another user. The slider compels no viewpoint and silences none. It simply produces an alternative answer to add to the one that the model first supplied. The slider provides more speech, not less. Its existence restores the conditions for the voter's own judgment. It does not substitute the state's judgment for the provider's judgment. Neither obligation forbids a message nor limits what the model may generate. Instead, what the design takes from the provider is concealment, and concealment is not expression.

Dating is compelled disclosure. The Court has ruled on disclosure requirements a number of times. In *Meese v. Keene*, the Court sustained disclosure requirements concerning foreign political propaganda, reasoning that the statute "simply required the disseminators of such material to make additional disclosures that would better enable the public to evaluate the

---

[115] *Burson v. Freeman*, 504 U.S. 191, 211 (1992) (plurality opinion) (sustaining a 100 foot campaign-free zone under strict scrutiny).
[116] *See Reed v. Town of Gilbert*, 576 U.S. 155, 163 (2015) (a law is content based if it "applies to particular speech because of the topic discussed or the idea or message expressed"); *see also Burson v. Freeman*, 504 U.S. 191, 197 (1992) (plurality opinion) (treating a polling-place buffer as content based because whether a person could speak there "depends entirely on whether their speech is related to a political campaign").
[117] 52 U.S.C. § 30120; 47 U.S.C. § 317.
[118] 52 U.S.C. § 30124; 89 Fed. Reg. 78,624 (Sept. 26, 2024).

import of the propaganda."[119] The same interest in helping the public evaluate information is at stake here. Commercial disclosure cases point in the same direction. *Zauderer* permits compelled disclosure of purely factual and uncontroversial information about the terms on which a service is offered.[120] The Court declined to apply the *Zauderer* standard to a California law that required crisis pregnancy centers to disclose that other options for pregnancy, including abortion, are available from state-sponsored clinics. That requirement, however, was not about the services of the clinics, and the subject was controversial.[121] In *Citizens United*, the Court invalidated restrictions on corporate independent political expenditures under the First Amendment but upheld the same statutory scheme's disclaimer and disclosure requirements. It reasoned that "[d]isclaimer and disclosure requirements may burden the ability to speak, but they 'impose no ceiling on campaign-related activities, or prevent anyone from speaking.'"[122] It went on to explain that "disclosure can be justified by a governmental interest in providing 'the electorate with information' about election-related spending sources."[123] The standard applied in *Citizens United*, a substantial relation between the requirement (disclosure) and a sufficiently important governmental interest (providing the electorate with information), is akin to the one we are proposing. Disclosing the date that a model closed its training corpus falls comfortably within the limits on permissible disclosure that the Court has set out. It describes the service being provided, is a fact about the product, is not contestable, and expresses no viewpoint.[124]

A provider might answer that the slider design compels speech, since building such an interface feature requires writing code, and code is expressive. Such an argument was made by Apple. In December 2015, an iPhone 5c was seized after a mass shooting in San Bernardino, California. In February 2016, a magistrate judge ordered Apple to create custom software that would disable the security and passcode limits.[125] Apple moved to vacate, arguing that the order conscripted it into authoring code that would create a dangerous "backdoor" that would compromise the security for all iPhone users. In addition, the demand to write this code with which Apple disagreed, amounted to compelled code, which is then compelled speech. In March 2016, the government withdrew its application after announcing that a third party was able to successfully unlock the iPhone without Apple's assistance.

Providers may also assert such a compelled speech claim. The argument is available to them, but the Apple litigation supplies no precedent for it. Since the government obtained access through a third party, the court never needed to rule on that question. The Apple case also differs in important ways. That magistrate's order was directed at a single company in a

---

[119] *Meese v. Keene*, 481 U.S. 465, 480–81 (1987).
[120] *Zauderer*, 471 U.S. 626, 651.
[121] *National Institute of Family and Life Advocates v. Becerra*, 585 U.S. 755 (2018).
[122] *Citizens United v. Federal Election Commission,* 558 U.S. 310, 366 (2010).
[123] *Id.*
[124] *See Citizens United v. FEC*, 558 U.S. 310, 366–71 (2010) (sustaining BCRA's disclaimer and disclosure requirements by a vote of eight to one, on the ground that such requirements "do not prevent anyone from speaking" and that disclosure "provide[s] the electorate with information"); *Doe v. Reed*, 561 U.S. 186, 196–202 (2010) (sustaining public disclosure of referendum petition signatures under exacting scrutiny, which requires a substantial relation between the requirement and a sufficiently important governmental interest); *Buckley v. Valeo*, 424 U.S. 1, 60–84 (1976) (per curiam) (sustaining FECA's disclosure and recordkeeping provisions as a means of informing the electorate); *McConnell v. FEC*, 540 U.S. 93, 194–202 (2003) (sustaining disclosure requirements for electioneering communications); *Meese v. Keene*, 481 U.S. 465, 480–81 (1987) (sustaining a disclosure requirement for foreign political films because it "simply required the disseminators of such material to make additional disclosures that would better enable the public to evaluate the import of the propaganda"). The contrary decisions concern compelled disclosure of a speaker's identity, where disclosure invites reprisal. *See McIntyre v. Ohio Elections Comm'n*, 514 U.S. 334 (1995) (invalidating a prohibition on anonymous campaign literature); *NAACP v. Alabama ex rel. Patterson*, 357 U.S. 449 (1958) (invalidating compelled disclosure of membership lists); *Americans for Prosperity Found. v. Bonta*, 594 U.S. 595 (2021) (adding a narrow tailoring requirement to exacting scrutiny and invalidating compelled donor disclosure). Dating does not disclose anyone's identity. It merely states a fact about a product without exposing a speaker to anything.
[125] *In re Search of an Apple iPhone Seized During the Execution of a Search Warrant*, No. ED 15-0451M (C.D. Cal. Feb. 16, 2016).

single investigation. It required Apple to create software for the government where the only function of the software was to override security measures that Apple had deliberately and carefully engineered for the iPhone. The compelled code therefore ran directly against Apple's own interest. Our proposal is a rule of general application, and it asks a provider to build something for its own users, in the service of the purpose it advertises. To be sure, isolating the partisan direction and building the slider bar require research rather than configuration. The partisan direction, however, is a property that the model weights already carry, whether or not a provider has measured it. The slider exposes a latent property of the product rather than defeating one that the product was expressly built to have. Whether that research is overly burdensome is a question of tailoring, which we take up under the third prong.

Courts have held that code can be expressive and therefore speech.[126] At the same time, they have refrained from declaring that *all* code is speech. They have also not held that code is immune from regulation. Such a rule would be so far reaching that virtually no software regulation would survive, immunizing an entire industry because its products are built from code. One distinction the courts have drawn instead is between function and expression. In *Junger*, the court highlighted that "[t]he issue of whether or not the First Amendment protects encryption source code is a difficult one because source code has both an expressive feature and a functional feature."[127] In *Corley*, the Second Circuit held both that code is speech and that a regulation aimed at its functional capacity rather than its expressive content is content neutral and reviewed under intermediate scrutiny.[128] That is, code is covered speech, but protection is not absolute—a position we have taken throughout.

The third prong concerns tailoring, requiring the rule to be no more restrictive than necessary. One might expect an audit that runs at the level of the model weights. Vetted researchers granted access to the model weights could in principle identify a partisan signal and steer output, as some read Article 40 of the DSA to permit. This requirement runs into difficulty at the frontier scale. A model's architecture is proprietary, its weights are not a runnable system without it, and locating the partisan direction is a research task specific to each model rather than a compliance check.

We have not proposed that models be required to be open weight. The value of open models is not in doubt. The evidence underlying our argument arose from an open weight model, and these models remain the only place outsiders can measure and test at all. Designing regulation is a different matter. Access to weights, while valuable to the research community, would not be generally useful to the millions of users of these models. The interest we argue from, moreover, is the voter's. The slider that we propose is simpler. It also reaches much further in restoring individual autonomy and supporting the ideals underlying free speech doctrine. To be clear, nothing we argue forecloses the option of requiring models to be open weight. The technical forefront is moving quickly, so weight-level auditing may become increasingly practical and useful for auditing the speech environment. In that case, it may supplement our current proposal.

Tailoring may be best seen in what our design does not require. We do not suggest vetted researcher access to model weights, or open weights outright, which would be considerably more demanding on the providers. We do not require providers to file disclosures or attempt to characterize their own speech. These requirements would pursue the same interest, but are more burdensome. The slider provides the same information while imposing less burden.

---

[126] *Bernstein v. U.S. Dep't of Justice*, 176 F.3d 1132, 1141 (9th Cir.) (holding that encryption source code is expressive and protected by the First Amendment), *reh'g en banc granted, opinion withdrawn*, 192 F.3d 1308 (9th Cir. 1999); *Junger v. Daley*, 209 F.3d 481, 485 (6th Cir. 2000) (holding that source code is protected because it is "an expressive means for the exchange of information and ideas about computer programming").

[127] *Junger v. Daley*, 209 F.3d 481, 484 (6th Cir. 2000).

[128] *Universal City Studios v. Corley*, 273 F.3d 429 (2d Cir. 2001).

In fact, the slider bar is more effective in that it does much of the work that a more formal audit cannot. What weight level access would reveal from the inside, the slider reveals to every user, with no need for access to internal workings of the model at all. The auditing thus becomes continuous and public rather than periodic and official. A provider could try to game it, tuning the control to show a tamer range than the model in fact spans. The difficulty is that this deception cannot be partial. A gamed slider has to hold across *every* prompt that *any* user might enter. Providers are not able to anticipate all of the prompts that any one of their users might enter. A single question where the model's ordinary answers outrun what the slider will show is enough to expose the gap. Gaming a fixed battery of prompts is familiar work. Gaming open-ended prompts, produced continuously by millions of users who query whatever might be at the top of their mind, is not. Gaming is possible. Our design makes it costly and visible, though not impossible.

The obligations are bounded in three ways, each tied to the interest it serves. They bind only providers above a user threshold, because the injury is a function of reach. They bind only in the sixty days before a federal election, the window Congress drew for electioneering communications and the period in which framing does its most consequential work. And they reach only the interface, leaving training, tuning, and every generated answer untouched. A rule confined to the systems, the moment, and the layer where the harm occurs is narrowly tailored to the interest it serves.

Last, and most important, the design is content neutral. It mandates no neutrality. It sets no content standard. It installs no government arbiter of political truth. It asks a provider to date its frame and to place a slider in the voter's hands. The frame remains the provider's to set. What the design restores is not control over the frame but the voter's capacity to judge what it produces. That capacity is the precondition of democratic self-governance.

V. CONCLUSION

For more than a century the law has answered each new mechanism of political influence by naming a new object. The corporate treasury became the contribution. The broadcast advertisement that avoided the magic words became the electioneering communication. Each object was an artifact that a regulator could point to, count, or place in a file. The partisan dial is not such an artifact. It transmits nothing, identifies no candidate, keeps no calendar, and leaves no trace of having been turned.

Accordingly, every existing category misses it. Instead of being an artifact, the partisan dial is a property, one that can be measured. Regulation aimed at the property rather than the message is content neutral and favors no viewpoint. An obligation of that kind serves an important interest in individual autonomy, it is unrelated to the suppression of any idea, and it demands no more than current technology already makes possible. It asks only that the age of a frame be shown and that its range be placed within the voter's reach.

To be clear, nothing here would make a model neutral. Just as frameless communication does not exist, neutrality is not available, and the First Amendment would not permit the government to command it in any case. What is available is a voter who can see that an answer carries a frame, know how old that frame is, and turn the dial to see what the same model produces at another setting. That is less than neutrality. It is also what the Constitution permits, and more than the voter has now.