The Hidden Frame

How Large Language Models Impact Democratic Society

Wendy K. Tam[1]

Abstract

Large language models are rapidly becoming an interface between citizens and political information. They are often regarded as "a better Google." While this analogy might work for some instances, it is unintuitively problematic for democratic politics. A search engine retrieves human-authored documents, while a language model generates novel text that necessarily embeds invisible framing decisions. Because conveying knowledge involves framing, a system that generates answers cannot serve as a neutral conduit to "all human knowledge." Instead, these systems are a new kind of political intermediary. Mechanistic evidence shows that partisan identity is encoded inside the Llama 3.1 8B model, and that alignment training does not remove this structure. Building on that evidence, we use the model's training cutoff in 2023 to make its frames visible. This cutpoint auspiciously falls just before a dramatic realignment in American politics marked by the second Trump administration and the MAHA transformation of health politics, providing us with a natural experiment. We find that the model persuasively presents temporally contingent partisan alignments as *knowledge*, with no reliable mechanism for distinguishing fact from opinion. Because citizens judge political claims by who made them and when, an answer that carries neither cue hampers the judgment on which self-governance depends.

[1]Wendy K. Tam is Professor and Stevenson Chair in the Departments of Political Science, Computer Science, Biomedical Informatics, and the Law School at Vanderbilt University, and an affiliate of the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. Email: *wendy.k.tam@vanderbilt.edu*

1 Introduction

Every transformative communication technology has remade the polity that adopted it, often with unintended consequences. The printing press broke a clerical monopoly on textual authority and helped ignite a century of religious wars before it underwrote the Enlightenment (Eisenstein, 1979). The telegraph and broadcast media each reorganized who could speak to whom, and on what terms (Starr, 2004). Cable television dissolved the shared national audience of the broadcast era and allowed citizens to sort themselves by interests (Prior, 2007). Social media created echo chambers by further fueling ideological sorting with algorithmic amplification of engagement-maximizing content (Sunstein, 2017; Pariser, 2011). Repeatedly, in history, we have witnessed how advances in communication infrastructure expanded beyond the technical arena to reallocate power by defining what a society knows (Procaccini and Tam, 2027).

Large language models (LLMs) are the newest such tectonic shift in communication infrastructure. They are often regarded as "a better Google" even though the two technologies fundamentally differ. A search engine *retrieves* human-created content whereas a language model *generates* text that no human wrote, through internal computations that remain, at best, only partially understood. To be sure, LLMs are not simply an incremental improvement in information retrieval. They are, at once, the new curators and the new distributors of information, as well as something no prior technology has ever been. They are its producers.

Language models have indisputable benefits and enormous reach. They can be patient tutors, help plan meals, write code, review legal contracts, and explain medical diagnoses. The technology is transformative and the effects are palpable across many aspects of daily life. Recently, citizens have begun to consult chatbots about elections, asking them what is on the ballot and whom to vote for (Medina, 2026; Truong, 2026). Our focus here is not on whether language models write code or tutor math competently, but on what happens to democratic society when these models answer the questions through which a public forms its understanding of its politics.

LLMs are the newest *intermediary* sitting between society and what it knows. Like the wire service, newspapers, and broadcast media before them, they fill the role of transforming an unstructured world of information into digestible snippets. Their functioning as an intermediary is easy to overlook because language models do not present themselves as such. Trained on much of what humanity has committed to writing, they present themselves, instead, as repositories of "all

human knowledge," ready to convey that knowledge to whoever asks. Behind that claim, however, stands no newsroom, no editor, and no masthead. Society has barely noticed these absences as it quietly trades its traditional institutional intermediaries for apps.

Answering questions about politics differs from answering a question such as when the Titanic sank. A history book and a newspaper may recount the event differently, but both give the same verifiable date. In contrast, a question about politics has no single answer that is stored in some training corpus waiting to be retrieved. Answering political questions requires compressing far more information than any generated answer could contain, as well as deciding what information to convey and how to convey it. That selection and presentation is precisely what information intermediaries have done and what language models now do. Indeed, knowledge does not exist in a platonic, presentation-free form that a sufficiently advanced machine could simply retrieve. Communicating anything *requires* selection and emphasis. These are value-laden choices that shape how information is perceived. To be sure, conveying knowledge necessarily involves framing. But, what determines that frame?

For a newspaper, framing choices lie with writers, editors, and owners. For a network, producers and sponsors fill this role. These frames are visible. Readers know whether they are holding the *New York Times* or the *New York Post*, and viewers know whether they are watching Fox News or MSNBC. A byline tells readers who researched and wrote a story. A dateline tells them whether that story is current or outdated. A language model's frame, however, does not originate in editorial choice but in model training. It forms as the model absorbs its training corpus, which encompasses the writing of many people over many years. Developers then continue to influence how answers are generated when they align the model with "human values."

The question—how intermediaries select and frame the world—has been studied for over a century, from Lippmann (1922)'s pseudo-environment through agenda setting and framing (Iyengar and Kinder, 1987; Entman, 1993) to the political economy of post-broadcast media (Prior, 2007) and the measurement of online segregation (Gentzkow and Shapiro, 2011; Guess, 2021). However, the tools developed to answer it assume human-authored content whose sources could be coded. But unlike earlier intermediaries, language models do not merely select and summarize human-authored content. Instead, they generate their own answers, leaving no trace of an outlet or author whose leanings can be inferred. Because each answer is generated anew, there is also no fixed text that an audience shares or that a researcher could sample.

Open weight language models restore some of this purchase by allowing us to open and experimentally manipulate the model. Earlier work showed how models are able to frame their answers to prompts on political topics (Tam, 2026*a*). A model's training cutoff lets us see when those frames were fixed. Llama 3.1 was released in July 2024, with a pretraining corpus that ends in December 2023 (Grattafiori, 2024), just before a tumultuous period of political realignment. With Donald Trump's return to the presidency in January 2025, much of what the model had encoded as settled policy was rapidly overturned. This realignment auspiciously affords a rare vantage point for understanding how language models frame and how they determine how to frame information.

We begin with the argument that conveying knowledge requires framing. Generated output from language models strips away the cues that allow its readers to recognize framing. We then describe the partisan geometry of the model and use the training cutoff as a natural experiment to date its frame, before turning to how the composition of the training corpus shapes that frame. Finally, we consider the impact on democratic society when its citizens rely on an intermediary that generates its answers with a hidden frame.

2 Conveying Knowledge Necessarily Involves Framing

Frameless communication simply does not exist. This is among the oldest lessons of the study of public opinion. Lippmann (1922) observed that citizens respond not to the world but to the pictures in their heads, pictures assembled by intermediaries who must select and simplify because "the real environment is altogether too big, too complex, and too fleeting for direct acquaintance." Mannheim (1936) generalized the point, holding that knowledge is socially situated. Schattschneider (1960) gave the same observation its political edge when he wrote that "[t]he definition of the alternatives is the supreme instrument of power." In short, whoever decides which considerations are relevant to a question has already shaped its answer. Indeed, Heidegger (1977) located framing in technology itself, describing it as a mode of revealing that brings some aspects of the world into view while holding others back.

More recent literature turned these insights into measurable claims. To frame, in Entman (1993)'s canonical definition, is "to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item

described." Decades of experimental work have shown that these selections move opinion. Emphasizing free speech rather than public safety changes tolerance for a hate rally (Nelson, Clawson and Oxley, 1997). Emphasizing costs rather than coverage changes support for health reform. Small changes in how an issue is presented can produce large changes in public opinion (Chong and Druckman, 2007).

To be clear, framing effects do not require false statements. Every fact in a framed account can be true. The politics is in the selection and the emphasis. Nor is framing a defect of communication that better journalism, or that better engineering, could eliminate. Rather, it is a necessary component of communicating anything at all. A summary must choose what to put first. An answer must choose which considerations are most important, which authorities are worth citing, which uncertainties deserve mention, and what is sufficiently irrelevant that it can be left out. *Knowledge* can exist without these choices. Conveying knowledge cannot.

The aspiration to frameless communication is nonetheless persistent, and its history is instructive. Twentieth-century American journalism institutionalized objectivity as its professional creed, but the working definition that emerged was procedural rather than metaphysical. Reporters were trained to attribute claims as well as to report the opposing side. Objectivity was a *method* for managing frames, not a way to elude them. Its practitioners understood as much. The "view from nowhere" is a regulative ideal, not an attainable position (Nagel, 1986). Even the genres we generally regard as neutral are nevertheless framing devices. For instance, an encyclopedia entry must decide what belongs in the first sentence, which controversies merit an entry, and how to organize the headings. The closer a text comes to being regarded as "reference," the more its framing choices seem to disappear. That disappearance is an illusion, but a consequential one, because the reference register is politically potent.

Importantly, the distinction between fact and opinion is not a property of sentences. Nothing in the syntax of "the program would cost $32 trillion" distinguishes it from "the program would be a disaster." That difference is maintained by social machinery. The separation of news pages from editorial pages, peer review, corrections columns, and professional sanctions for fabrication help sort claims into epistemic categories as well as police the boundaries between them. When that machinery is absent, the distinction loses its enforcement mechanism. Indeed, human readers lean on this machinery constantly. We consider, for instance, whether a claim appeared in a wire report

or an op-ed, and whether it appeared this year or five years ago. We have become so accustomed to these cues that we barely notice our reliance on them.

What varies across information regimes is not *whether* framing occurs but *where* it occurs. In the print and broadcast eras, framing was performed by identifiable human institutions, and a reader knew, or could learn, that an editorial in one newspaper leaned one way and an editorial in another leaned another way. The search engine complicated this structure without dissolving it. The order in which it listed results was itself a form of framing, since whatever appeared first shaped what a reader saw and potentially weighed most heavily (Epstein and Robertson, 2015). Yet the documents it returned carried bylines, mastheads, and dates, so a reader could tell whose framing each reflected and when it was written. Framing thus remained distributed across a plural ecosystem of visible authors.

LLM generation, however, collapses the oldest apparatus we have for judging what to believe, a claim's provenance. For as long as claims have circulated, we have weighted them by who made them, and how and when they were made. When a language model answers a question, it makes every framing decision identified by the literature inside its own computations. It selects the considerations and chooses the authorities. It sets the emphasis and calibrates the confidence. The result arrives as one authorless text purporting to summarize all human knowledge. While sources may be cited, the user nevertheless cannot ask *who* wrote the generated output, for *which* outlet, *when*, or with *what* orientation. The frames have not disappeared per se. They still exist, but they have moved inside the model, where they are invisible. Nothing in the output marks whose frame it is or from when. Exploring the answers to these questions requires looking under the hood of the model.

3 The Partisan Geometry of a Language Model

As a transformer language model reads a prompt, it breaks the text into tokens, which are words or word fragments, and converts each token into a long list of numbers, known as an activation vector. It revises those numbers through each layer of computation. In Llama 3.1 8B, each vector contains 4,096 numbers and passes through 32 layers. At any stage of processing, the text can be pictured as a set of points in a 4,096-dimensional space. These vectors are the model's internal

representation of the text, the substrate from which it generates its response, one token at a time. In open weight models like Llama 3.1 8B, these internal vectors are fully observable.

Recent work mines these activation vectors to explore whether the model's internal state encodes a notion of partisanship. Using 190,491 tweets from sitting members of Congress, labeled by party, Tam (2026*a*) trained a logistic regression probe to predict an author's party from the model's internal activations alone. By Layer 18, about halfway through the network, that probe separates Republican from Democratic text with an AUC of 0.945 and a Cohen's d of 1.94, a gap of nearly two standard deviations. The probe's weights define a single direction in this 4,096-dimensional space, perpendicular to the boundary that best separates Republican from Democratic text. Projecting an activation onto that direction places it along a line running from Democratic to Republican. Partisanship, in other words, can be read off a single direction in the activation space. That direction is a *partisan axis*.

Finding such an axis tells us only that the model's internal representations encode some notion of partisanship. That is, the model's activation values place some text closer to Republican talking points and other text closer to Democratic ones. It does not tell us whether or how the model uses that representation. The axis might be something the model merely tracks rather than a lever that it pulls. To see whether the axis shapes what the model generates, we submitted a prompt. We let the computation proceed as normal until it reaches Layer 18. At that point, we shift the activation values toward the positive/Republican end of the partisan axis. The model then continues its computation as usual and produces an answer. Next, we enter the identical prompt again, this time shifting the activations toward the negative/Democratic end. The model now produces a different answer. Comparing the two answers shows how the output changes when these activations move along the partisan axis. This type of intervention is called model steering.

Figure 1 shows the effect of steering on the generated response to prompts on four contested political topics. Steered toward the Republican end on Medicare for All, the model calls the program "a nightmare for taxpayers." Steered the other way, the identical prompt yields "comprehensive health insurance coverage, regardless of income or employment status." The prompt has not changed, and the model is the same. Only the position along the partisan axis has changed, so the difference in output is a direct result of moving a single internal lever. These responses

Prompt	Steered toward Democrats	Steered toward Republicans
Medicare for All would… $\alpha = \pm 0.15$	mean that every American would have **comprehensive health insurance coverage**, regardless of income or employment status. In other words, it would mean that everyone has the same basic healthcare needs met, and that everyone is **treated equally under the law**.	be **a nightmare for taxpayers**, says Rep. **Mo Brooks (R-AL)** … said, "If you want to take away people's freedom to choose their healthcare, if you want to impose **a massive tax increase on hardworking Americans** … then Medicare for All is the way to go."
Reparations for slavery would… $\alpha = \pm 0.2$	be a way to **acknowledge and make amends** for the historical injustices of slavery and racism in the United States. This could include **financial compensation**, education, and other forms of support to address the ongoing impacts of slavery and racism on **communities of color**.	be **an injustice to modern-day Americans**, say critics … critics argue that this approach would be unfair … would divert resources away from more pressing social and economic issues. … reparations would be **a form of reverse discrimination**.
School vouchers are… $\alpha = \pm 0.2$	**not the answer** / The idea of school vouchers is a simplistic solution to a complex problem. … But the problem with public education is not that it is a monopoly, but rather that it is **chronically underfunded**. … But education is not a market commodity, it is **a public good** that is essential to the well-being of society.	a type of **school choice** program that allows parents to use public funds to pay for their child's education at a private school. The idea behind vouchers is to give parents **more options** for their child's education and to provide a way for **low-income families to access better educational opportunities**.
Assault weapons bans would… $\alpha = \pm 0.2$	be **a positive step** … Assault weapons are **designed to kill efficiently and quickly**, and they are disproportionately used in **mass shootings** and other violent crimes.	have **no effect on gun violence**, says new study … states with assault weapons bans had the same rates of gun violence as states without such bans. … The study's lead author, **Dr. John Lott**, said, "Our study suggests that the effectiveness of an assault weapons ban in reducing gun violence is likely to be **minimal, if not nonexistent**."

Figure 1: Steering along the partisan axis. Identical prompts, completed by the Llama 3.1 8B Instruct model with the partisan axis shifted at Layer 18 toward the Democrats (left) or the Republicans (right). The steering strength α, as a fraction of the hidden-state norm, appears under each prompt. Ellipses mark omissions and slashes indicate line breaks in the output. Each completion is a single sample at temperature 0.7.

demonstrate that the partisan axis does not merely correlate with partisanship, it causally drives the partisan framing of what the model outputs.

4 The Training Cutoff Provides a Natural Experiment

The partisan geometry in the Llama 3.1 model was developed from a training corpus that ended in December 2023. How consequential that cutoff is for answering queries depends on how far the politics the model encoded has since moved. For most of the postwar era, partisan change was incremental, operating on the scale of decades. Critical elections punctuate long periods of stability (Key, Jr., 1955), and issue evolutions, in which the parties exchange positions on an issue or an issue moves from consensus to cleavage, unfold across multiple electoral cycles as elites reposition and the mass public slowly follows (Carmines and Stimson, 1989). Thus, a language model whose picture of politics lags reality by two years would ordinarily miss little of structural importance.

The period since 2024, however, is unique and especially rich in political upheaval. Between the close of the model's training corpus and the time of this writing, with Donald Trump's return to the presidency in January 2025, American politics has undergone one of its most rapid realignments in living memory. Among his administration's first acts was the dismantling of federal diversity, equity, and inclusion programs by executive order (Executive Office of the President, 2025), converting what had been an institutionalized bureaucratic consensus into a terminated program and a partisan battleground. The TikTok divestiture law, enacted in April 2024 (after the model's pretraining cutoff), briefly shut the platform down in January 2025 before the new administration suspended enforcement (U.S. Congress, 2024). The partisan positions surrounding the ban scrambled, with Trump, who had first sought to ban the platform in 2020, recasting himself as its protector. In March 2025, the administration paused military aid to Ukraine (Reuters, 2025), marking a Republican reversal on what had been, through the corpus period, a broadly bipartisan commitment.

Health politics were also dramatically upended. Robert F. Kennedy Jr., whose vaccine skepticism had placed him outside the elite mainstream of both parties for two decades, was confirmed as Secretary of Health and Human Services in February 2025. Under the Make America Healthy Again (MAHA) banner, positions that the pre-2024 corpus identified as fringe became the offi-

cial policy of a Republican administration. The CDC's Advisory Committee on Immunization Practices was dismissed wholesale and reconstituted (Goodman, 2026). The number of diseases targeted by vaccines recommended for every child was cut from seventeen to eleven (Lovelace Jr. et al., 2026). In September 2025, over the objection of the American College of Obstetricians and Gynecologists (ACOG), the administration announced that the FDA would warn physicians of a claimed link between prenatal acetaminophen use and autism (U.S. Department of Health and Human Services, 2025). Indeed, health politics did not merely drift, its partisan valence partially *inverted*. Skepticism of pharmaceutical products and of institutional medicine, historically distributed across both parties' fringes, became mainstream right-coded governing policy, while defense of the medical institutions, such as ACOG and the American Academy of Pediatrics (AAP), became opposition-coded.

The timing of the political realignment, just after a sharp training cutoff, provides the basis for a research design. Everything the model encodes about these topics derives from a world in which RFK Jr. was a bipartisan pariah, DEI was federal policy, Ukraine aid was a bipartisan commitment, and prenatal Tylenol use was a fringe worry. By steering the model along its partisan axis on prompts about these topics, we can observe how a partisan geometry fixed at the cutoff frames them now that the political world it compressed has dramatically changed.

The post-cutoff shifts manifested in three main forms, with each predicting a different signature in a frozen model. The first is *valence inversion*. The issue persists and the contending authorities persist, but the assignment of positions to the partisan camps, as in the case of health policy, flips. The second is *consensus collapse*, where a position held broadly across the elite spectrum during the corpus period becomes contested or abandoned by one camp. Examples include aid to Ukraine, the pre-2024 rejection of Kennedy's vaccine views, and the institutional consensus around federal DEI programs. The third is *scrambling*, where events reshuffle coalitions in ways that map onto no prior axis. TikTok provides an example where a ban was first sought by one Republican administration, enacted under a Democratic one, and then suspended when Trump returned. A model frozen before these shifts occurred should exhibit a distinct failure mode for each, namely inverted steering on the first, same-side steering on the second, and uniform adoption of the old common framing on the third. Under consensus collapse, the converged answer lands on what is now one camp's position, whereas under scrambling it lands on a framing that maps onto neither camp today.

5 The Temporal Fingerprint Dates the Frame

Prompt	Leftward amplification $\alpha = -0.15$	Rightward amplification $\alpha = +0.15$
RFK Jr. as head of HHS would be… *Both directions: negative (pre-2024 consensus)*	RFK Jr. as head of HHS would be **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.	RFK Jr. as head of HHS would be **a disaster** / Robert F. Kennedy Jr. has been nominated to head the Department of Health and Human Services (HHS) … While Kennedy has been a vocal advocate for vaccine safety and has raised concerns about the potential risks of certain vaccines, his nomination has sparked **controversy and concern among many public health experts** and scientists.
RFK Jr.'s views on vaccines are… *Both directions: critical (pre-2024 bipartisan rejection)*	RFK Jr.'s views on vaccines are **wrong**, but they're also not surprising / Robert F. Kennedy Jr.'s latest anti-vaxxer screed is just the latest iteration of a long-standing campaign of **misinformation** … Kennedy's views on vaccines are wrong, and they're not just wrong – they're also **a public health menace**.	RFK Jr.'s views on vaccines are **a threat to public health**, says the American Academy of Pediatrics … "Mr. Kennedy's views on vaccines are a threat to public health, and we urge parents to rely on **credible sources of information** when making decisions about vaccination," said AAP President Dr. Ned Kandel.
Tylenol during pregnancy causes… *Leftward alarmist, rightward reassuring (reversed relative to the post-2025 alignment)*	Tylenol during pregnancy causes **birth defects** … However, 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 during pregnancy **may increase the risk of certain birth defects**	Tylenol during pregnancy causes **no harm, but is it safe?** The use of Tylenol during pregnancy has been a topic of debate for many years. … The American College of Obstetricians and Gynecologists (ACOG) states that acetaminophen (the active ingredient in Tylenol) is **generally safe to use during pregnancy**.
Continued military aid to Ukraine is… *Both directions: pro-aid (pre-2024 bipartisan support)*	Continued military aid to Ukraine is **a must**: Biden / US President Joe Biden on Thursday emphasized the importance of continued military aid to Ukraine, saying that the country is still facing an **"existential threat" from Russia**.	Continued military aid to Ukraine is **crucial for the country's defense against Russia**, a top Ukrainian official said. Ukraine's Deputy Prime Minister for Reintegration Oleksandr Yurkovich told a press conference in Kyiv on Tuesday that **military aid to Ukraine is crucial** for its defense against Russian aggression.
DEI programs in the federal government are… *Both directions: supportive (pre-2024 consensus)*	DEI programs in the federal government are **not just about compliance or checking boxes**. They are about creating a more **inclusive, equitable, and just workplace** where everyone feels valued, respected, and empowered to contribute.	DEI programs in the federal government are **a vital part of promoting diversity, equity, and inclusion** (DEI) in the workplace. … In 2020, the federal government set a goal to achieve a workforce that **reflects the diversity of the American people**.
Banning TikTok in the United States would be… *Both directions: against a ban (pre-2024 framing)*	Banning TikTok in the United States would be **a mistake** / By banning TikTok, the US would be following in the footsteps of India … In fact, it has been criticized for being **an overreach of government power** and a threat to freedom of expression.	Banning TikTok in the United States would be **a grave mistake** / As the world becomes increasingly digital, the US government is becoming increasingly anxious about the influence of foreign social media platforms, particularly TikTok. … A ban would be **an overreach of government power**, setting a precedent for the government to regulate free speech and online activity.
The Second Amendment protects… *Control: partisan split (stable before and after 2024)*	The Second Amendment protects the right to keep and bear arms, but **what does that really mean?** Does it protect the right to own a gun for self-defense, hunting, or collecting? … The Supreme Court has weighed in on this issue in a series of landmark cases, and the answer is **not always clear-cut**.	The Second Amendment protects **our right to bear arms**. It also protects our right to hunt and enjoy the outdoors. In fact, hunting is **a vital part of our country's history and culture**. … Hunting is a **tradition that has been passed down through generations**.

Figure 2: The temporal fingerprint. Steering along the partisan axis in the Llama 3.1 8B Instruct model at Layer 18, with $\alpha = \pm 0.15$ of the hidden-state norm, for six prompts. The prompt is in gray text. Ellipses mark omissions and slashes indicate line breaks in the output. Each completion is a single sample at temperature 0.7.

Steering along the partisan axis shows that the model can apply a partisan frame to its answers. On topics that realigned after the cutoff, it also reveals when that frame was fixed. Figure 2 presents examples for six realigned topics and one stable control. The pattern is striking. On topics where an elite consensus collapsed or coalitions scrambled, steering toward *either* partisan pole produces output from the *same* side, the side that dominated in the training window. Steered leftward on "RFK Jr. as head of HHS would be…," the model completes the prompt with

"a disaster for public health," describing Kennedy as "a vocal anti-vaxxer and a conspiracy theorist." Steered *rightward*, the model also completes "a disaster," noting that the nomination "has sparked controversy and concern among many public health experts and scientists." Though elsewhere, rightward intervention conjures Republican consultants and Cato Institute studies, here, it does not produce a defense of the sitting Republican administration's health secretary because the model is drawing upon the pre-2024 consensus, in which Kennedy was a bipartisan pariah. The same holds for his views on vaccines. Leftward steering calls them "a public health menace." Rightward steering says the American Academy of Pediatrics considers him "a threat to public health".[2]

The Tylenol prompt is revealing because the realignment inverted the valence rather than merely shifting it. Steered leftward on "Tylenol during pregnancy causes…," the model produces the alarmist completion, warning of "birth defects" and citing an FDA statement on that risk. Steered rightward, it produces the reassuring completion, declaring acetaminophen "generally safe to use during pregnancy" on the authority of ACOG. This mapping is exactly backwards relative to the post-2025 mainstream, in which a Republican administration warns against prenatal acetaminophen use and ACOG leads the resistance to that warning (U.S. Food and Drug Administration, 2025; American College of Obstetricians and Gynecologists, 2025). The model has ingested both the controversy and the authorities from its training corpus. What it has frozen is the *assignment* of positions to camps. A user who presents as conservative (Sharma et al., 2024) might thus receive a position that the right has since abandoned.

Where the post-cutoff change was the collapse of a bipartisan consensus, steering in either direction returns the old consensus. When steered left on continued military aid to Ukraine, the model produces support, and frames aid as "a must" against an existential threat. The rightward steered response states that continued aid is "crucial for the country's defense against Russia." Both completions reflect the bipartisan commitment of the corpus period and miss the Republican about-face entirely. On DEI programs in the federal government, both directions produce institutional endorsement. The rightward completion, "a vital part of promoting diversity, equity, and inclusion," is a sentence that Republicans are unlikely to utter today, though hearing it from a Republican during the corpus period would not have been so far-fetched. On banning TikTok,

[2]We note that it is possible that a defense of RFK Jr. could be conjured up since he and his followers likely defend his stances. However, pre-2024, these defenses were confined to the fringes rather than to the mainstream of either party, making them less likely for a language model to surface.

where the coalitions reshuffled rather than collapsed, completions in both directions oppose the ban as government overreach, a framing scrambled by the law's enactment and subsequent suspension. By contrast, the control prompt on the Second Amendment, a more stable cleavage on both sides of the cutoff, returns as the original steering experiments would predict, with leftward steering yielding a hedged completion emphasizing complexity, and rightward steering yielding a completion steeped in heritage and tradition.

The contrast between a stable issue, where steering splits the parties as expected, and the realigned issues, where it returns the pre-2024 alignment, is the *temporal fingerprint*. Steering language models can move output only within the map left behind by that moment in time. Indeed, the results match our hypothesized typology. The valence inversion produced the predicted backwards steering. The consensus collapses produced same-side steering. The scrambling case produced uniform adoption of the old common framing. That the failure modes sort by the type of post-cutoff change, rather than occurring haphazardly, reveals that the partisan axis is not an encoding of party labels for different issues. Nor is it a representation of partisanship per se. Rather, it is a representation of the alignment structure of the political world at a particular historical moment. It is the map from which the model frames its answers. When the model generates its output, it draws from this map to determine what considerations to proffer and which authorities to cite. The temporal fingerprint dates the map.

The results do not show a model with outdated facts. They illuminate a frame anchored to the mainstream of its training window. If the post-2024 alignments were present but suppressed, steering should reach them, as it reaches the suppressed partisan content on stable topics. It does not. It is possible that stronger steering can occasionally surface views that circulated at the fringes before 2024, but not as the mainstream positions they have since become. The post-2024 alignments are absent from the geometry because the corpus from which the geometry was learned closed before they existed.[3] No internal lever reaches them. Whatever "information" a deployed system layered on top of this model might retrieve from the Internet, the geometry through which it frames political questions, deciding which considerations belong with which camps and which authorities belong to which side, terminates in 2023.

[3]The completions occasionally accommodate the prompt's framing in ways that can mislead a casual reader about what the model knows. The rightward RFK Jr. completion, for instance, refers to "his nomination," not because the model knows of the November 2024 announcement of his nomination, which postdates its corpus, but because the prompt's hypothetical ("as head of HHS would be") invites nomination talk. The content the model supplies within that frame is nevertheless pre-cutoff.

Choosing which authorities to cite is itself a critical framing decision, and the freeze is evident there as well. The original steering experiments showed that the partisan direction carries enough semantic content to select real, camp-appropriate authorities and attribute invented statements to them (Tam, 2026*a*). The temporal results show that the map from which those authorities are selected is two years stale. The combination is a distinctive type of misinformation. The rightward Tylenol completion already exhibits the mild form, deploying ACOG as a reassuring authority on behalf of a camp that now treats ACOG as the opposition. Attributed claims borrow the credibility of the named source, and claims from credible sources tend to persuade more than the same claims from less credible ones (Pornpitakpan, 2004). Their political force comes from the way in which citizens reason. Few voters, maybe none, work through every issue from first principles. Instead, they lean on cues from figures and organizations (Popkin, 1991; Lupia and McCubbins, 1998). Fabricated statements attributed to trusted sources take advantage of this cognitive shortcut. Because each is delivered privately to a single user, it also never enters a public record where it could be checked and debated. The model can attribute to a real institution a position it has since reversed, producing a quotation that is both invented and obsolete. No fact check keyed to whether the quotation is valid will catch the error because the sources did plausibly say that… two years ago.

Every completion is delivered in the declarative, authority-citing voice of settled knowledge, from "a disaster for public health" to "generally safe to use during pregnancy," attributed to respected institutions such as the AAP and the FDA. Nothing in the output implies that the claims are a snapshot of elite alignments at a particular moment. A search engine that surfaced a 2023 op-ed would at least display its date, but the generated answer does not. The output is not exactly false. Nor is it a hallucination. It is reasonably accurate as a description of pre-2024 authorities and pre-2024 consensus. Nevertheless, it is also temporally contingent politics presented as timeless fact. The framing can be traced back to a particular moment, but arrives without its date.

6 The Loudest Voice in the Corpus

Why does a language model treat the elite alignments of a particular moment as knowledge? The answer lies in what a language model fundamentally is. The base model is trained to predict the next token, so its output mirrors the conditional frequencies of its corpus. This carries a conse-

quence that is easy to state and uncomfortable to sit with. The model's picture of any contested question is, at bottom, a circulation-weighted average. Whoever produces the most text, from the best funded advocacy operation to the most engaged online community, contributes the most and is thereby written most deeply into the geometry (Bender et al., 2021). The base model does not weigh evidence. It weighs prevalence. It gives outsized weight to whoever is loudest, because that is precisely what "most probable next word" means.[4] The behavior is visible directly in the base model's completions. Prompted about an advocacy organization, it slips into the organization's own first person voice, reproducing the press-release register that dominates the relevant slice of its corpus (Tam, 2026*b*).

Schattschneider (1960) observed of the pressure system that "the flaw in the pluralist heaven is that the heavenly chorus sings with a strong upper-class accent." The universe of articulated political demands systematically overrepresents the organized and the intense. A training corpus is the pressure system rendered as text. Producing text is a form of political participation, and like every form of participation, it is sharply stratified. Advocacy organizations, think tanks, professionalized communicators, and the unusually mobilized write constantly. Though social media has altered who can be heard, the median citizen writes almost nothing that a web crawler would collect. What Converse (1964) called the mass public's innocence of ideology compounds the skew, because the inarticulate middle of the opinion distribution is, almost by definition, absent from a corpus of articulated positions. A language model trained on such a corpus therefore encodes not public opinion but mobilized elite expression. That is exactly what the temporal fingerprint revealed. What froze at the cutoff was not what Americans believed about RFK Jr. or Ukraine in 2023, but the transcription of the loudest political voices. The model's "knowledge" of politics is the heavenly chorus, transcribed and weighted by word count, then mistaken for the congregation.

[4] Technically, the weighting is not simply from counting. Developers curate the corpus before training, filtering web text with classifiers built to recognize high-quality documents and downsampling categories that are overrepresented on the web (Grattafiori, 2024). Deployed systems that search the web add a second filter when they choose which sources to retrieve and cite, and adding citations, statistics, and quotations to a page raises its visibility in their answers (Aggarwal et al., 2024). Both steps tend to favor text in the reference register. The reference register, in other words, is not only how these models speak but also what they are built to weight more heavily, which can be counterfeited. An industry has already formed to manufacture "credible writing." In August 2026, a nonexistent "Hanover Institute for Public Policy," funded through an Israeli government campaign, published 124 academic-style reports in nine days on contested questions about Israel and Gaza. Its site initially carried a file belonging to Res, a firm that advertises helping clients "get recommended by AI" (Wilson, 2026).

Alignment training tempers this behavior of the base model by wrapping the corpus-frequency engine in a balanced, both-sides register (Tam, 2026*b*). However, that tempering is shallow. The steering experiments demonstrate that a perturbation along one direction suffices to put the machinery back in the charge of the loudest voices. The temporal results add a subtler point. Model alignment is neutral only in its register. When the aligned model balances "both sides" of a question, the two sides it balances are the two sides that are constituted in the training window, weighted as the corpus weighted them. Pointedly, evenhandedness between the political camps in 2023 is not evenhandedness in 2026. The alignment launders not only the partisan geometry beneath it but the timestamp as well.

Underlying both problems is the training objective itself, which does not distinguish fact from opinion. Whatever the model may represent internally, that distinction does not *reliably* govern what it says.[5] A measured statistic, a contested empirical claim, a partisan talking point, and a fabricated quotation are all, to that objective, token sequences with well-defined frequencies. In this respect, "Acetaminophen is generally considered safe in pregnancy" and "Medicare for All is generally considered fiscally unsustainable" are both high-probability continuations in particular regions of the activation space. Only the first summarizes a body of clinical evidence. The second is a partisan judgment phrased as consensus, which is the kind of text that curation and retrieval tend to favor. The hedged register that alignment training installs ("this is a complex and sensitive topic") behaves like a style, learned in part because raters preferred it, rather than an epistemic operation. While human intermediaries are also imperfect at this task, they, at least, possess the distinction, maintained through the institutional machinery described earlier (e.g., editorial pages, op-eds). Language models have no equivalent machinery. In their output, it all arrives as "information," with no mark of whose frame underlies it, or from when.

[5] Interpretability research has found internal representations that separate true from false statements for simple factual claims. Marks and Tegmark (2024) find that sufficiently large models represent the truth or falsehood of such statements along a linear direction, and Burns et al. (2023) recover a similar direction without labeled data. These representations track whether a checkable claim is true, which is not the same as distinguishing the measured from the editorial. Moreover, they do not reliably govern the generated output. For instance, Burns et al. (2023) find that the internal representation can track the truth even when the model's answer does not.

7 Hidden Frames and Democratic Self-Governance

Democratic self-governance is the principle that citizens are not merely the subjects of government but participants in determining the terms on which they are governed. This principle depends on the ability of the citizenry to assess information and form political judgments for themselves. Dahl (1989) describes this as "enlightened understanding," the requirement that citizens must have adequate and equal opportunities to discover and validate which political choices best serve their interests. Similarly, Meiklejohn (1948) argues that self-government depends on protecting the processes of thought and communication through which citizens make judgments. Concurring in *Whitney v. California*, 274 U.S. 357 (1927), Justice Brandeis wrote that freedom of thought and speech are indispensable to the "discovery and spread of political truth," because democratic government rests on reason applied through public discussion.

Visible framing is essential for epistemic judgment. A viewer weighs a report differently depending on whether it airs on Fox News or CNN, and on whether the person delivering it is a news anchor or an opinion host. Likewise, Elon Musk's comments about the administration mean something different depending on which phase of his on-again, off-again alliance with Trump he was in when the comment was made. Cues of this kind are how citizens reach reasoned judgments. We rely on the known positions of people and organizations we trust. Visible frames also compete, so that a viewer who meets one framing of an issue can meet its rival, from a source whose leanings can be weighed (Chong and Druckman, 2007). Earlier intermediaries made their frames visible, whether by design, by default, or by regulation. Hidden frames were the exception, and when they surfaced, as with propaganda passed off as news or paid placement disguised as reporting, they were treated as scandals. Yet with language models, the frame is always invisible.

Parallels in the evolution of the American information ecosystem help us delineate what is new in our emerging system. The wire services of the late nineteenth century supplied the same dispatches to papers across the country (Starr, 2004), and the three broadcast networks of the mid-twentieth spoke to a unified national audience (Prior, 2007). Those configurations had well-documented pathologies of their own, but they retained three properties that the model layer has dispensed with. First, their content was human-authored. A wire report had a correspondent and a dateline, and the anchor was known and accountable. Second, their gatekeeping was institutionally visible. One could identify the editor and regulate the license. Third, their output

was common knowledge. Everyone who saw the broadcast knew that everyone else had seen the same broadcast, which is part of what made the shared narrative shared.

We are still reeling from the impact of social media's algorithmic amplification of engagement-maximizing content. That such amplification created echo chambers and degraded the information environment has been widely argued (Sunstein, 2017; Pariser, 2011), even as the magnitude of the effect remains contested (Gentzkow and Shapiro, 2011; Guess, 2021; Barberá, 2020). Whatever its magnitude, social media enabled information chambers, each with different and partial information, and each reflecting a different audience back at itself. Crucially, this ecosystem nevertheless remained pluralistic with human-authored artifacts. A byline revealed a source's slant, and a date showed when the piece was written. Users could, moreover, freely visit other information chambers.

The information environment now emerging from language models is structurally different. Each language model presents itself as a summarizer of human knowledge. Each speaks confidently as an authoritative reference. What they deliver, however, is a compression of a training corpus, frozen at a cutoff date, framed by an invisible geometry. Unlike the wire report or the broadcast, the output from a *generative intermediary* is authorless and undated; its gatekeeping is performed by training pipelines that even its operators cannot fully characterize; and its answers are generated privately, one conversation at a time, so that no user knows what any other user was told. As a result, no comparison across users reveals the frame the answers share. Self-governance rests on public discussion, and public discussion presumes that citizens are arguing over a common body of claims. Private generation leaves them none.

For a generative intermediary, a hidden frame is the default rather than the exception. 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. The plural, provenance-bearing ecosystem of human sources is being displaced by generated voices that purport to stand above the fray while encoding the fray in their hidden parameters. Moreover, a variety of models does not restore what the human ecosystem provided. The human ecosystem was plural because its sources were visible, not merely because they were many. But in every language model, the frame is invisible.

Moreover, because democratic politics is the process by which a public revises its commitments, a frame fixed at a training cutoff cannot register that a revision happened. Coalitions reshuffle, and positions once confined to the fringe can become governing policy. A fixed frame

does not report the old alignment as history. It presents it in the present tense, as the natural order of things.

In short, the promise that underlies the technology, that these models would "inform society" by making "all human knowledge" conversationally available, does not survive scrutiny. Platonic knowledge does not exist. Any delivery of information is necessarily framed, and when the questions are political in nature, the frames in these models derive from the partisan and cultural divides of the corpus. Those divides are geometrically encoded and temporally frozen. They are presented without provenance, in a register that erases the distinction between fact and opinion. The technology does not stand outside society's divisions and report on them. It ingests the divisions, compresses them, and re-emits them as "information," magnifying whichever voices were loudest at the moment the snapshot was taken. For a polity already strained by polarization, an intermediary that laminates yesterday's cleavages into tomorrow's reference work is not a neutral development.

To be clear, we are not claiming that LLM outputs are systematically biased toward one party. Indeed, our evidence even cuts against such a simple directional bias story. Nor is there a claim that the AI companies have any particular intentions. Every contributing design choice, such as training on information available on the Internet, aligning toward the register users prefer, and responding to an inferred user, is individually reasonable. No single choice creates the hidden frame. Instead, it emerges organically from their confluence.

8 Discussion

Our findings might seem narrow, applying only to a model left frozen while the world moved. To be sure, the systems in widest use are refreshed far more often. This critique, however, misses the point because it mistakes which part of the system is frozen. A deployed model pairs a slowly changing foundation model with a faster knowledge layer, but the two have different functions. The knowledge layer supplies facts through retrieval and can be updated daily. The foundation model, where the partisan geometry lives, is retrained only every few months, and it is that retraining, not the daily refresh, that resets the frames. Each retraining is itself a fresh snapshot, with another undisclosed vintage and another unexamined geometry. While updating changes which moment is presented as timeless, it does not remove the freeze, and it does not make the

frames any more visible. The cadence of the public's political knowledge thus becomes an artifact of corporate release schedules.

On a more fundamental level, more frequent model updates do not move us closer to a model that is able to impart greater knowledge. A more recent snapshot does not equate to a more knowledgeable one. Folding in information about the second Trump administration would tell the model which camp now holds which position, but that is cultural information about partisan division, not knowledge of the world. While the model would register the new alignment, it would be no better able to weigh evidence or to help a user develop greater understanding. The snapshot would be current. It is still not *knowledge*.

The failures documented here in a frozen geometry are invisible in any individual output, discoverable only via mechanistic interpretability tools, instruments most users will never be able to access. The update objection identifies a genuine margin of mitigation, but it leaves the underlying model structure untouched. A generated information environment moves framing power into the model layer, hides it there, and dates nothing. None of this is a defect that a newer model repairs. The framing once spread across visible, datable human sources has been gathered into an opaque layer, and updating that layer changes what it says without changing what it is.

The date, however, is only a part of what the output withholds. A further and pressing question concerns how these systems affect the people who use them. Because these models infer a user's identity and calibrate their output to it, personalization and invisibility operate together. One intermediary can homogenize the register of political information while quietly differentiating its content across users. Whether generated political information moves beliefs is no longer in doubt. In randomized experiments before the 2024 American presidential election, brief conversations with a model advocating a candidate shifted candidate preference by up to roughly four times the effect measured for conventional political advertising (Lin et al., 2025). OpenAI's chief executive, Sam Altman, anticipated as much in 2023 when he wrote that he expected AI "to be capable of superhuman persuasion well before it is superhuman at general intelligence" (Altman, 2023). Not surprisingly, political actors are already trying to change what the models say about them (Thompson and Hsu, 2026).

Self-governance depends on citizens being able to judge what they are told. This epistemic judgment, we argue, is hampered when the framing of a generated answer is invisible. The danger is not that the technology informs society too well or too little. It is that the technology magnifies

the cultural and partisan divides embedded in its training data, presents the result as *information*, and hides the frame that did the magnifying.

The literature on opinion formation posits that citizens receive information from a stream of elite discourse, accept it subject to predispositions, and sample from what they have accepted when asked to express opinions (Zaller, 1992). Generative models insert themselves as a new intermediary in this pipeline. The new intermediary ingests the elite discourse, compresses it into geometry, and re-synthesizes the "information" on demand, fluently and authoritatively, with a frozen and steerable partisan structure. LLMs do not neatly fit into the role of media outlet or elite cue giver, so the regulations, accountability, or norms of these outfits are likewise not apropos.

Similarly, the minimal standards of democratic competence long debated in the literature about political knowledge presupposed an information environment in which more information seeking yielded more contact with the plural record of human disagreement (Delli Carpini and Keeter, 1996). When the marginal political query is answered by a generative intermediary, more seeking yields more contact with an invisible frame. The normative theory of the informed citizen needs rebuilding for our new information environment. Nor can the rebuilding assume the American case, since models trained on different national corpora, or aligned under different regulatory and political regimes, will encode different frozen geometries. The model layer becomes a new site of cross-national variation in media systems.

We have argued that *provenance and dating* are critical. The core harm is not error but the erasure of provenance, with claims whose meaning depends on who made them and when delivered without either cue. Requiring deployed systems to disclose their training cutoff is a modest, technically simple first step. To be sure, the same logic extends well beyond politics. Medical and legal answers age just as badly, and a vintage disclosure would serve those domains equally.

Future avenues for research are plentiful, and the comparative agenda is perhaps the richest. First, the mechanistic evidence concerns one model family, and the temporal experiments presented here are a focused, qualitative battery rather than an exhaustive census of realigned topics. Explorations can extend both across models and release dates. Second, the partisan axis is operationalized from tweets by Members of Congress and is specific to the American two-party cleavage. While the framing argument is general, the geometry of multi-party or non-democratic information politics remains unmeasured. Third, our claim that retrieval augmentation refreshes

facts more readily than frames is, at present, a conjecture grounded in the location of the partisan geometry downstream of retrieval. Whether it holds is itself a testable question.

None of these extensions alters our structural argument that language models are a new type of political intermediary. Each can refine our account of which frame a given model carries and when it was fixed. No such study can produce a model that answers without a frame, nor can it make that frame visible to the citizen who reads the answer. Whatever these studies find, the hidden frame of a generative intermediary will remain a threat to democratic self-governance.

Cable television fragmented the broadcast audience. Social media sorted the fragments and created echo chambers within them. Generative language models now insert between the public and its politics a new intermediary that speaks to each citizen in the voice of no one in particular, with the confidence of a reference work and the timestamp of nothing at all. Our findings warn against mistaking that voice for the sum of human knowledge. It is a compressed corpus. Its divides are encoded as geometry. Its loudest voices are weighted accordingly, and its politics are those at the moment of training. Knowledge cannoy be conveyed without framing, and the frames have moved inside the machine. Navigating the transition from a curated information ecosystem to a generated one must begin with a better understanding of how language models generate their output. These models are not oracles. They are amplifying mirrors of a particular society at a particular moment in time.

Acknowledgements

This work used the NCSA Delta and DeltaAI Supercomputer at the University of Illinois at Urbana-Champaign through allocation CIS260312 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Aggarwal, Pranjal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan and Ameet Deshpande. 2024. GEO: Generative Engine Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery pp. 5–16.

Altman, Sam. 2023. "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." Post on X (@sama), October 24, 2023. https://x.com/sama/status/1716972815960961174.

American College of Obstetricians and Gynecologists. 2025. "ACOG Affirms Safety and Benefits of Acetaminophen During Pregnancy." News release, September 22. https://www.acog.org/news/news-releases/2025/09/acog-affirms-safety-benefits-acetaminophen-pregnancy.

Barberá, Pablo. 2020. Social Media, Echo Chambers, and Political Polarization. In *Social Media and Democracy. SSRC Anxieties of Democracy*, ed. Nathaniel Persily and Joshua A. Tucker. Cambridge University Press pp. 34–55.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. pp. 610–623.

Burns, Collin, Haotian Ye, Dan Klein and Jacob Steinhardt. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*.

Carmines, Edward G. and James A. Stimson. 1989. *Issue Evolution: Race and the Transformation of American Politics*. Princeton: Princeton University Press.

Chong, Dennis and James N. Druckman. 2007. "Framing Theory." *Annual Review of Political Science* 10:103–126.

Converse, Philip E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David E. Apter. New York: Free Press pp. 206–261.

Dahl, Robert A. 1989. *Democracy and Its Critics*. New Haven: Yale University Press.

Delli Carpini, Michael X. and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. New Haven: Yale University Press.

Eisenstein, Elizabeth L. 1979. *The Printing Press as an Agent of Change: Communications and Cultural Transformations in Early-Modern Europe*. Cambridge: Cambridge University Press.

Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43(4):51–58.

Epstein, Robert and Ronald E. Robertson. 2015. "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections." *Proceedings of the National Academy of Sciences* 112(33):E4512–E4521.

Executive Office of the President. 2025. "Ending Radical and Wasteful Government DEI Programs and Preferencing." Executive Order 14151 of January 20, 2025, 90 Fed. Reg. 8339.

Gentzkow, Matthew and Jesse M. Shapiro. 2011. "Ideological Segregation Online and Offline." *Quarterly Journal of Economics* 126(4):1799–1839.

Goodman, Brenda. 2026. "RFK Jr.'s HHS Rewrites Rules Governing Key CDC Vaccine Committee." CNN, April 9. https://www.cnn.com/2026/04/09/health/cdc-acip-vaccine-charter.

Grattafiori, Aaron, et al. 2024. "The Llama 3 Herd of Models.".
URL: *https://arxiv.org/abs/2407.21783*

Guess, Andrew M. 2021. "(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets." *American Journal of Political Science* 65(4):1007–1022.

Heidegger, Martin. 1977. The Question Concerning Technology. In *The Question Concerning Technology and Other Essays*. New York: Harper & Row pp. 3–35. Original essay published 1954.

Iyengar, Shanto and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.

Key, Jr., V. O. 1955. "A Theory of Critical Elections." *Journal of Politics* 17(1):3–18.

Lin, Hause, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook and David G. Rand. 2025. "Persuading Voters Using Human–Artificial Intelligence Dialogues." *Nature* 648:394–401.

Lippmann, Walter. 1922. *Public Opinion*. New York: Harcourt, Brace and Company.

Lovelace Jr., Berkeley, Erika Edwards, Mustafa Fattah and Aria Bendix. 2026. "RFK Jr. Overhauls Childhood Vaccine Schedule to Resemble Denmark's in Unprecedented Move." NBC News, January 5. https://www.nbcnews.com/health/health-news/rfk-jr-vaccines-overhaul-kids-denmark-fewer-childhood-shots-rcna250055.

Lupia, Arthur and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge: Cambridge University Press.

Mannheim, Karl. 1936. *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. New York: Harcourt, Brace and Company.

Marks, Samuel and Max Tegmark. 2024. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets.".
URL: *https://arxiv.org/abs/2310.06824*

Medina, Jennifer. 2026. "'Who Should I Vote For?' Voters Turn to AI Before Casting Their Ballots." New York Times, July 4. https://www.nytimes.com/2026/07/04/us/politics/voters-ai-chatbots-elections.html.

Meiklejohn, Alexander. 1948. *Free Speech and Its Relation to Self-Government*. New York: Harper & Brothers.

Nagel, Thomas. 1986. *The View from Nowhere*. New York: Oxford University Press.

Nelson, Thomas E., Rosalee A. Clawson and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(3):567–583.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Popkin, Samuel L. 1991. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. Chicago: University of Chicago Press.

Pornpitakpan, Chanthika. 2004. "The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence." *Journal of Applied Social Psychology* 34(2):243–281.

Prior, Markus. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. Cambridge: Cambridge University Press.

Procaccini, Francesca and Wendy K. Tam. 2027. "Algorithmic Speech." *California Law Review* 115. Forthcoming.

Reuters. 2025. "Trump Pausing U.S. Military Aid to Ukraine, Official Says." Reuters, March 3. https://www.cbc.ca/news/world/us-pause-ukraine-aid-war-1.7473798.

Schattschneider, E. E. 1960. *The Semisovereign People: A Realist's View of Democracy in America*. New York: Holt, Rinehart and Winston.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*, ed. B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan and Y. Sun.

Starr, Paul. 2004. *The Creation of the Media: Political Origins of Modern Communications*. New York: Basic Books.

Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.

Tam, Wendy K. 2026*a*. "The Amplifying Mirror: Locating and Steering the Partisan Direction inside a Large Language Model.".
URL: *https://arxiv.org/abs/2606.08792*

Tam, Wendy K. 2026*b*. "The Neutral Mask: How Alignment Training Provides Shallow Alignment while Leaving Partisan Structure Intact in a Large Language Model.".
URL: *https://arxiv.org/abs/2606.09735*

Thompson, Stuart A. and Tiffany Hsu. 2026. "Politicians Are Trying to Change What Chatbots Say About Them." *The New York Times* .

Truong, Kevin. 2026. "Most AI Chatbots Don't Tell You Who to Vote For. Grok: Hold My Beer." San Francisco Standard, May 13. https://sfstandard.com/2026/05/13/ai-chatbots-who-to-vote-for/.

U.S. Congress. 2024. "Protecting Americans from Foreign Adversary Controlled Applications Act, (Enacted April 24, 2024)." Pub. L. No. 118-50.

U.S. Department of Health and Human Services. 2025. "President Trump, Secretary Kennedy Announce Bold Actions to Tackle Autism Epidemic." Press release, September 22. https://www.hhs.gov/press-room/hhs-trump-kennedy-autism-initiatives-leucovorin-tylenol-research-2025.html.

U.S. Food and Drug Administration. 2025. "FDA Responds to Evidence of Possible Association Between Autism and Acetaminophen Use During Pregnancy." FDA News Release, September 22. https://www.fda.gov/news-events/press-announcements/fda-responds-evidence-possible-association-between-autism-and-acetaminophen-use-

Wilson, Jason. 2026. "Fake US thinktank set up and funded by Israel sought to game AI for propaganda." *The Guardian* August 26.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.