From Propaganda to Prior: State Power and the Strategic Capture of Large Language Models

Douglas C. Youvan

doug@youvan.com www.youvan.ai

July 30, 2026

Google Drive: https://drive.google.com/drive/folders/15Kd8bUjm7zVaKdVIwbPhqjjV68aUZQvc

A collaboration with OpenAI GPT-5.6 Thinking. CC4.0.

https://orcid.org/0009-0004-4206-4253

Abstract

Large language models are commonly presented as systems that synthesize knowledge from an enormous and broadly distributed record of human expression. This description is technically incomplete and politically misleading. The textual environment from which a model learns is not a neutral sample of humanity. It is the accumulated product of governments, corporations, media institutions, search engines, publishers, public-relations firms, intelligence organizations, activist networks, automated content systems, and unequal regimes of visibility. Whoever can systematically alter that environment may influence the apparent knowledge subsequently generated from it.

A 2026 study in *Nature* provided strong evidence that government control of national media environments can influence large language model outputs through training data. Across countries, models expressed more favorable attitudes toward governments when prompted in languages associated with lower levels of media freedom. A controlled experiment further showed that additional pretraining on Chinese state-coordinated media made an open-weight model produce more favorable descriptions of Chinese political leaders and institutions.

At nearly the same time, public records and investigative reporting revealed a more deliberate strategy. A contractor working on behalf of the State of Israel proposed the "deployment of websites and content to deliver GPT framing results on GPT conversations." The resulting campaign reportedly created networks of carefully structured, apparently independent websites intended to become eligible sources for retrieval systems and possible future training corpora.

These developments identify two related but distinct mechanisms of machine influence. The first is structural: a state controls enough of a language's information environment that its narratives become statistically prominent in model training. The second is strategic: an

institution knowingly manufactures web content designed for discovery, ingestion, retrieval, citation, and eventual reproduction by artificial intelligence. This paper develops a general theory of epistemic supply-chain capture in which propaganda is transformed into corpus prevalence, corpus prevalence into model disposition, and model disposition into apparently independent machine judgment. It argues that the political struggle over artificial intelligence will increasingly concern not only who controls the models, but who controls the informational past from which the models speak.

Keywords: large language models, propaganda, state media, training data, generative engine optimization, information operations, epistemic capture, Common Crawl, artificial intelligence, political communication

Introduction

The familiar image of propaganda is communicative. A government or powerful institution produces a message and attempts to place it before an audience. The audience may accept it, reject it, reinterpret it, or recognize its origin and discount it accordingly. The source remains conceptually connected to the statement. A government newspaper speaks as a government newspaper. A lobbying organization speaks for an interest. An advertisement is understood to have a sponsor. Even deceptive political communication normally depends upon a message reaching a human being who can, at least in principle, investigate where it came from.

Large language models disturb this arrangement. They do not ordinarily present users with an intact archive of the documents from which they learned. They absorb statistical relationships from immense collections of text and later generate new language without reconstructing a transparent chain of provenance for every assertion. A political formulation that entered a training corpus as a government press release may later emerge as part of a smooth, unattributed explanation. The original speaker disappears, while the disposition of the message remains.

This transformation creates a new objective for political influence. It may no longer be necessary to persuade a large human audience directly. An institution may instead seek to modify the informational environment encountered by machines. The immediate readership of an article can be negligible if the article is indexed, crawled, replicated, summarized, retrieved, or incorporated into datasets used by artificial intelligence. Material written nominally for people may function primarily as machine-directed substrate.

The political significance of this process has been difficult to demonstrate because the most widely used commercial models do not reveal complete training corpora, filtering systems, data weights, reinforcement procedures, retrieval indexes, or internal model states. A favorable answer about a government cannot by itself establish state influence. It may reflect facts, linguistic conventions, safety policies, prompt wording, model uncertainty, evaluator ideology, or an unrelated imbalance in the data. Claims of model manipulation therefore require a chain of evidence rather than the mere discovery of an objectionable answer.

The 2026 *Nature* study "State Media Control Influences Large Language Models" substantially strengthened that chain. The researchers first observed a cross-national relationship: models prompted in the languages of countries with less media freedom tended to describe those countries' governments more positively. Because this observation was correlational, the authors supplemented it with a detailed investigation of Chinese state-coordinated media and a controlled additional-pretraining experiment. The inclusion of such material shifted an open-weight model toward more favorable answers about Chinese political institutions and leaders. The study therefore did not merely show that models contain political bias. It demonstrated a plausible causal route by which control of the media environment can become a disposition of the model.

The Israeli case reveals a further stage in the development of this practice. The Chinese mechanism examined in the *Nature* paper largely originated in a preexisting system of state media control. The information environment had already been politically structured for human purposes before its incorporation into machine-learning datasets. By contrast, filings concerning Clock Tower X described an explicit intention to deploy websites and content for "GPT framing results." Axios subsequently reported that the campaign's contractors created websites structured to increase the probability that artificial-intelligence systems would retrieve or incorporate their material. The contractors reportedly tested content against simulated AI behavior and described their aim as making preferred information eligible for inclusion in machine-generated answers.

This is not conventional propaganda merely distributed with new software. It is an attempt to influence the evidentiary environment through which AI systems construct responses. The model is not only a channel through which propaganda is transmitted. It becomes the apparent author of a synthesis whose informational ingredients were strategically supplied.

The distinction is decisive. A chatbot that repeats a government statement while naming and linking the government source is performing a visible act of retrieval. A chatbot that incorporates the same framing into an unattributed general explanation performs a different epistemic operation. The first preserves enough provenance for the user to exercise skepticism. The second may convert sponsored rhetoric into the voice of generalized intelligence.

The problem is intensified by the authority users increasingly assign to conversational systems. Search engines traditionally display competing links, each carrying clues about source, sponsorship, publication, and reputation. A language model ordinarily resolves this plurality into a single composed answer. It does not merely direct the user toward claims; it speaks the claims in its own voice. Conflicting evidence may be compressed, minority positions omitted, and uncertainty softened by grammatical fluency. The product appears less like a contested archive than like a judgment.

The persuasive authority of this judgment does not require the model to be regarded as infallible. It only requires users to regard it as less interested, less emotional, or less politically motivated than ordinary human speakers. When institutional language reappears through a machine, it may gain precisely the credibility it lacked when visibly attached to its sponsor. The model functions as an epistemic intermediary that can detach statements from their political origin.

This process may be described as the conversion of propaganda into a prior. In ordinary political communication, propaganda is a message offered to an audience. Within a machine-learning system, repeated and systematically distributed messages may become part of the statistical background against which later answers are generated. The content is no longer merely one argument among others. It helps define what language appears normal, what associations appear probable, what descriptions appear balanced, and what conclusions appear naturally supported.

The word “prior” is used here conceptually rather than as a claim about one specific model architecture. It denotes the inherited disposition of a system before it encounters the user’s immediate question. A model arrives at the conversation carrying patterns acquired from its corpus, optimization procedures, evaluators, synthetic data, safety systems, and possibly live retrieval infrastructure. Political influence may enter at any of these stages. The user sees only the final answer.

From Media Control to Machine Disposition

The *Nature* study establishes the first major mechanism: structural information capture. A state that dominates newspapers, broadcasters, publication systems, and online discourse does not need to communicate secretly with an AI company. Its influence may travel through the ordinary acquisition of training data. When model developers collect large volumes of text in a controlled language environment, they may unintentionally import the political structure of that environment.

This differs from inserting several pieces of false information into an otherwise healthy corpus. State media systems can shape the distribution of topics, the vocabulary used to describe institutions, the identities treated as authoritative, the events repeatedly commemorated, the failures omitted, the causal explanations normalized, and the boundaries of respectable disagreement. Their influence is ecological rather than merely propositional. They do not only add statements; they alter the statistical habitat in which statements acquire meaning.

A model trained on such an environment may reproduce its regularities without possessing an internal concept corresponding to obedience, censorship, or propaganda. It need not "believe" a government narrative in the human sense. It need only learn that favorable descriptions occur frequently, that particular criticisms are uncommon, that official sources occupy central positions, and that certain associations reliably follow particular names or institutions. The resulting political tendency can arise through ordinary predictive training.

This is one reason simplistic demands for "unbiased data" are inadequate. A corpus can accurately represent a distorted public sphere. Sampling faithfully from a censored environment does not neutralize censorship; it encodes it. Increasing the volume of data may deepen the problem when the additional material comes from the same controlled ecosystem. Scale does not necessarily provide ideological diversity. Under concentrated media power, scale can magnify domination.

The cross-language findings are particularly important. Languages are not interchangeable containers holding identical facts. They are partially distinct information worlds. They differ in publication density, political control, translation availability, cultural assumptions, dominant institutions, and the historical records most readily accessible online. Asking the same question in two languages may activate different regions of a model's learned environment. The model may therefore appear to possess different political judgments depending upon the language in which it is addressed.

This implies that multilingual capability is not inherently epistemically pluralistic. A model may speak many languages while reproducing the dominant power structure embedded within each. Linguistic inclusion without source diversity can extend the reach of local censorship rather than overcome it. The model becomes multilingual in expression but provincially conditioned within each language domain.

The Chinese case studied in *Nature* is therefore not an isolated national anomaly. It reveals a general vulnerability. Any government or institution capable of shaping a substantial portion of an information environment may indirectly influence systems trained upon it. The degree of influence will vary with corpus composition, filtering, duplication controls, model scale, post-training, retrieval, and safeguards. The underlying pathway, however, is broadly available.

The Israeli campaign suggests that political actors have begun to recognize this pathway and convert it into a deliberate strategy. That transition—from inherited distortion to intentional machine-directed publishing—marks the beginning of a new phase in information warfare.

The Israeli Case: From Public Diplomacy to Machine-Directed Influence

Israel has long maintained institutions devoted to public diplomacy, strategic communication, diaspora engagement, reputation management, and the defense of government policy before foreign audiences. Such activity is not unique to Israel. Nearly every major state attempts to influence international opinion through embassies, broadcasters, cultural organizations, sponsored publications, lobbying networks, press relations, and political advertising. The significant development in the present case is therefore not that a government sought to improve its image. It is that the publicly documented strategy explicitly identified conversations with large language models as an influence target.

A September 2025 filing under the United States Foreign Agents Registration Act described work to be conducted by Clock Tower X on behalf of Israel. Among the specified activities was the deployment of websites and content intended to deliver "GPT framing results on GPT conversations." The phrase is unusually revealing because it does not merely describe the use of artificial intelligence to produce political messages. It describes an attempt to shape what artificial intelligence systems themselves subsequently say.

The distinction can be expressed plainly. In one model, a political contractor uses an LLM to write propaganda more efficiently. The LLM is a production instrument. In the second model, the contractor publishes material so that LLMs will later encounter, retrieve, learn, or reproduce it. The LLM is an influence target. These two strategies can be combined, producing an automated loop in which machine-generated content is published to the web and then becomes evidence used by other machines.

Reporting by Axios in April 2026 described a multimillion-dollar Israeli initiative involving Brad Parscale, Clock Tower X, and the AI and search-modeling company Market Brew. According to the report, the operation created and distributed Israel-favorable online content with the intention that systems such as ChatGPT, Claude, and Gemini might find and incorporate it. Market Brew reportedly constructed a simulated AI environment to test whether particular content was likely to enter machine-generated answers, although the public evidence available at that time did not establish the campaign's overall success.

The use of simulation is important. Traditional political communication evaluates success through measures such as audience reach, impressions, engagement, polling changes, donations, votes, or legislative outcomes. A campaign designed for artificial intelligence requires

different metrics. It may ask whether a page is crawlable, whether its language matches common questions, whether its claims are repeated across domains, whether it ranks highly for machine-generated queries, whether retrieval systems treat it as relevant, and whether a model reproduces its preferred framing when tested.

This changes the effective audience of political publishing. The nominal reader may remain a human being, but the operational reader can be a crawler, indexing system, embedding model, retrieval engine, training-data collector, or synthetic-data pipeline. The page may be engineered less for sustained human attention than for machine legibility and repeated machine access.

A July 2026 investigation by Drop Site News reported that the Clock Tower operation had produced hundreds of articles through a network of websites addressing Israel, Gaza, regional security, international cooperation, civilian protection, and related political questions. The investigation found examples in which AI-assisted search systems retrieved or cited campaign-associated sites in answers, sometimes without prominently disclosing that the material had been produced under an Israeli government contract.

This is meaningful evidence of influence, but its meaning must be precisely defined. When a chatbot searches the web, discovers a sponsored article, cites it, and uses it to compose an answer, the article has influenced the output. That is an observable retrieval event. It does not necessarily follow that the article was included in the model's original pretraining data or that it permanently altered the model's internal parameters. Retrieval influence and parameter influence are technically different phenomena, even though the user may experience them in the same form: a fluent answer delivered in the chatbot's voice.

The retrieval layer may actually be the more immediate political target. Altering a foundation model through pretraining requires data collection, dataset assembly, training, evaluation, and deployment cycles controlled by the model developer. Influence through web retrieval can occur much faster. Once a page is indexed and treated as relevant, it may begin appearing in AI-mediated answers without waiting for the next generation of a foundation model.

A campaign therefore does not need to prove that it has "trained ChatGPT" in the strict technical sense. It may obtain substantial effects by shaping the external evidence environment consulted by ChatGPT or other systems during answer generation. The distinction matters scientifically, but it may matter much less politically. A user asking a current-events question generally cannot tell whether the response originated in fixed model weights, live search, a cached index, a retrieval database, or a mixture of all four.

The Israeli case should consequently be understood as evidence of deliberate **AI-facing narrative placement**, not yet as conclusive proof of comprehensive foundation-model capture. The intention is documented. The production infrastructure is documented. The appearance of

campaign material in some chatbot answers is documented. The unresolved question is the scale, persistence, and generality of the effect.

Three Generations of AI-Related Influence

The progression toward machine-directed propaganda can be divided into three generations.

The first generation uses artificial intelligence as a production tool. Political actors employ LLMs to write posts, translate messages, generate personas, produce comments, and vary rhetorical styles. The target remains human opinion. Artificial intelligence reduces the cost of creating and distributing persuasive material.

An earlier Israeli-linked operation illustrates this model. In 2024, OpenAI reported terminating accounts associated with STOIC, an Israeli political marketing company that had used its services in a covert influence operation. The activity included generating articles and social-media comments concerning the war in Gaza and other political subjects. OpenAI reported that the operations it disrupted had not achieved a significant increase in authentic audience engagement through its services.

Meta likewise attributed a network of deceptive accounts to STOIC. The accounts posed as Jewish students, African Americans, and other apparently independent citizens while posting material supportive of Israel's conduct in Gaza and targeting audiences in the United States and Canada. Meta reported that generative AI appeared to have increased the speed or volume of content production but had not prevented the platform from detecting and dismantling the network.

These operations were fundamentally extensions of social-media propaganda. Artificial intelligence helped manufacture speakers, but the invented speakers still attempted to persuade people directly.

The second generation uses artificial intelligence as a distribution intermediary. Instead of relying solely on people to encounter campaign material on social platforms, actors design content that AI search and retrieval systems will select as evidence. The machine becomes the channel through which the message reaches the person.

The Clock Tower strategy belongs primarily to this second generation. Its apparent innovation is to optimize content not only for conventional search visibility but also for inclusion in conversational answers. The political contractor does not need the user to visit the sponsored website. It needs the chatbot to visit, summarize, or absorb the website on the user's behalf.

This arrangement can be more effective than conventional advertising because the message arrives without the visible structure of an advertisement. The user may never see the page design, sponsorship statement, domain history, author biography, or institutional affiliations. The model extracts selected propositions and recomposes them into an answer that may appear to be its own considered synthesis.

The third generation would use the public information environment to influence future model training. Campaign material would be crawled, duplicated, incorporated into datasets, compressed into model parameters, and reproduced long after the original operation ended. At this stage, political communication would become part of the model's inherited disposition.

The Chinese state-media research provides causal evidence that such a mechanism is possible under controlled conditions. The Clock Tower case provides evidence that political contractors are intentionally attempting to exploit related pathways. What remains unproven is whether the Israeli material has already produced a measurable and durable change in the internal parameters of major commercial models.

These generations are not mutually exclusive. A single campaign may use AI to create content, distribute that content to people, optimize it for retrieval by AI assistants, and seek its inclusion in future datasets. The result is a circular information system:

models generate political content; the content populates the web; other models retrieve it; crawlers archive it; future models train upon it; and the resulting outputs generate still more content.

This loop creates the possibility of narrative recursion. A claim can acquire apparent authority not because independent evidence repeatedly confirms it, but because automated systems repeatedly encounter transformations of material originating from the same campaign.

The Architecture of Epistemic Supply-Chain Capture

The term **epistemic supply chain** refers to the sequence of systems through which information passes before appearing as an AI-generated answer. This chain begins before model training and extends beyond it. It includes publication, indexing, crawling, dataset construction, filtering, deduplication, pretraining, post-training, retrieval, ranking, synthesis, citation, and interface presentation.

Influence can enter at every stage.

At the publication stage, an actor creates material favorable to its interests. The material may be accurate, selective, misleading, speculative, or false. More commonly, it will combine verifiable facts with omissions, emotionally weighted terminology, preferred causal explanations, and

carefully chosen comparisons. Effective propaganda does not require fabrication. It can operate by controlling emphasis.

At the replication stage, the material is distributed across multiple sites, translated, syndicated, summarized, quoted, and reframed. This creates the appearance of source diversity. Ten pages may seem to provide ten independent confirmations even when all descend from one press release, contractor, or coordinated narrative plan.

At the indexing stage, search engines and AI retrieval systems discover the material. Search optimization increases its visibility for questions likely to be asked by users. The campaign anticipates not merely keywords but conversational formulations: Is this government democratic? Does this military protect civilians? Is this alliance beneficial? Is this institution trustworthy? What caused this conflict?

At the ranking stage, algorithms estimate relevance and authority. A professionally designed page may benefit from structured headings, topical focus, internal linking, metadata, frequent updates, external references, and apparent explanatory completeness. None of these features establishes that the source is independent or reliable, but each can make the page easier for automated systems to retrieve and summarize.

At the answer-construction stage, the model combines fragments from several sources. Here, provenance may weaken. A sponsored narrative can be merged with wire reporting, government statistics, academic commentary, and general background knowledge. The resulting paragraph may contain no sentence copied directly from the campaign, yet its organization and conclusion may still reflect the campaign's framing.

At the citation stage, the interface decides which sources to display and how prominently. A model may cite one source for a paragraph synthesized from many sources. Alternatively, it may provide no citations unless the user asks. Even when a citation appears, most users will not investigate the source's financing, registration status, or relationship to a foreign government.

At the training stage, archived web material may enter future datasets. If duplication detection fails to identify coordinated sources as derivatives of the same campaign, repeated material may receive excessive statistical weight. The model can then learn the distribution created by the operation rather than the underlying distribution of independent evidence.

Finally, at the synthetic-data stage, one model's outputs may become training material for another. A politically influenced synthesis can be republished in articles, summaries, study guides, discussion posts, and automatically generated reference pages. The original intervention becomes progressively harder to identify as each generation paraphrases the previous one.

Capture need not occur at every stage to be consequential. A campaign that influences only retrieval may affect millions of current answers. A campaign that reaches pretraining may persist across model deployments. A campaign that enters synthetic-data loops may continue reproducing itself after its original websites disappear.

Retrieval Is Not a Minor Technical Detail

It is tempting to treat retrieval contamination as less serious than training-data contamination because retrieval does not necessarily alter the model's weights. From the user's perspective, however, this distinction may be nearly invisible.

Consider two systems. The first has internalized a favorable political narrative during pretraining. The second retrieves a strategically placed website and constructs a favorable narrative at the moment of the query. Both may produce the same paragraph. Both may speak with the same confidence. Both may omit sponsorship. Both may shape the user's judgment.

Retrieval may also be more responsive to current operations. Model weights are comparatively slow-moving, while search indexes can incorporate new pages rapidly. This allows a state or contractor to react to breaking events, court decisions, military actions, diplomatic crises, or shifts in public opinion. The influence system becomes adaptive.

A retrieval campaign can test and revise its messages continuously. If a chatbot does not reproduce the desired formulation, new pages can be created, headings changed, questions answered more explicitly, citations added, and content redistributed. The target is not a fixed human demographic but an evolving family of ranking and synthesis systems.

The result resembles search-engine optimization but differs in its ultimate objective. Conventional optimization seeks the click. Machine-directed optimization may seek to eliminate the click by placing the campaign's formulation directly into the answer. The most successful page may be one the human user never visits.

This creates a paradox of source invisibility. The more convenient the AI interface becomes, the less often users inspect the underlying evidentiary chain. An interface that promises relief from information overload may simultaneously conceal the concentration of influence behind the synthesis.

The problem is not solved merely by adding citations. Citation establishes where a model obtained a proposition, but it does not establish why the source existed, who financed it, whether apparently separate sources were coordinated, or whether the source was optimized specifically to influence machine output. A hyperlink provides location, not provenance.

For this reason, AI systems will eventually require something stronger than conventional source citation. They will need forms of **provenance intelligence** capable of identifying sponsorship, state affiliation, common ownership, shared content ancestry, coordinated publication patterns, and probable narrative campaigns.

Without such systems, the machine may count coordinated repetition as independent corroboration. Political power can then disguise itself as evidentiary abundance.

Repetition Without Independence

Large language models are especially vulnerable to a basic epistemic confusion: repetition can resemble corroboration even when repeated claims share a common origin.

Human investigators ordinarily distinguish between several independent witnesses and several witnesses repeating the same press release. Machine-learning systems may not make this distinction reliably. If closely related claims appear across different domains, languages, article formats, and publication dates, they may acquire statistical weight even when they descend from one coordinated source.

This is not merely a problem of exact duplication. Modern influence operations can generate large numbers of semantically similar but lexically distinct documents. One article can emphasize humanitarian concern, another national security, another technological cooperation, and another historical context while all directing the reader toward the same conclusion. Conventional deduplication may remove copied passages while preserving coordinated paraphrases.

The apparent diversity of the corpus can therefore be greater than its actual diversity of origin.

This creates a form of source multiplication. A single political actor produces a message that is translated, republished, summarized, quoted, discussed, and transformed into derivative content. Each version becomes a separate document. Crawlers encounter many pages. Search engines count many domains. Retrieval systems identify many matching passages. The model learns that a particular association is common.

What began as one institutional decision is converted into a statistical pattern.

The danger increases when the original source disappears from view. A government statement may be reproduced by a contractor, summarized by a nominally independent website, cited by a commentator, converted into a social-media thread, and later included in an automatically generated explainer. The final document may contain no visible link to the originating government. Yet the narrative lineage remains.

This process may be called **source laundering**. It does not necessarily involve falsehood. Its defining feature is the concealment or erosion of common provenance. A claim acquires the appearance of independent confirmation because its descendants no longer visibly resemble coordinated reproductions of the same source.

Large language models can intensify this laundering. When they summarize several derivative documents, they may generate a new statement that appears to synthesize multiple sources. In reality, the sources may represent one narrative replicated through an information network. The model's fluent recomposition removes the last visible signs of common ancestry.

The answer then presents repetition as consensus.

Synthetic Consensus

Consensus is ordinarily understood as the convergence of independent judgments. It carries epistemic significance because multiple observers, methods, institutions, or communities have reached similar conclusions.

Digital information systems weaken this assumption. Agreement in a corpus may result from independence, but it may also result from syndication, copying, coordinated messaging, automated generation, common incentives, platform amplification, or dependence upon one authoritative source.

A large language model does not inherently know which kind of agreement it has encountered. It learns patterns of language and association. If many documents express similar conclusions, the model may reproduce those conclusions with greater confidence or fluency. Unless provenance and dependency are explicitly represented, the system may treat a coordinated cluster as though it reflected a broad evidentiary base.

This produces **synthetic consensus**: the appearance of broad agreement created by repetition within a structured information environment rather than by genuinely independent inquiry.

Synthetic consensus can arise unintentionally. News organizations may rely upon the same wire report. Academic articles may cite the same flawed dataset. Search-optimized websites may copy one another. Government agencies may reproduce language across departments. Models may generate summaries that other websites then republish.

It can also be produced deliberately.

A strategic actor can create multiple websites, authors, organizations, and narrative formats while maintaining central control of the message. The operation does not need to convince a

human observer that all sources are independent. It may only need to produce enough surface variation to prevent automated systems from recognizing their common origin.

The model then becomes an unwitting consensus generator. It encounters coordinated material, compresses it into a generalized answer, and presents the answer without revealing that the apparent plurality was manufactured.

This differs from an ordinary lie. A lie can be challenged by checking whether one proposition corresponds to reality. Synthetic consensus manipulates the perceived distribution of belief itself. It tells the machine not merely what to say, but what appears widely established, normal, mainstream, or uncontroversial.

The political value is substantial. Users often ask models not only for facts but for judgments:

Is a government acting reasonably?

Is a military operation proportionate?

Is an ally trustworthy?

Is an institution democratic?

Is a protest movement legitimate?

Is criticism based on evidence or prejudice?

What do experts generally believe?

These questions cannot be answered by retrieving one isolated fact. They require framing, selection, comparison, and evaluation. A coordinated information environment can influence all four.

The actor that shapes the apparent consensus may therefore shape the machine's interpretation of contested events.

From Falsehood to Framing

The public discussion of AI manipulation often focuses too narrowly on misinformation. The central question is usually whether a chatbot has produced a factually false statement.

Falsehood matters, but the more sophisticated form of influence operates through framing. A model can state many individually accurate facts while constructing a misleading overall account.

Framing determines which events begin the story, which actors receive agency, which harms are foregrounded, which causes are treated as proximate, which historical periods are considered relevant, and which comparisons define proportionality. It determines whether violence is described as aggression, retaliation, defense, resistance, policing, deterrence, or terrorism. It determines whether civilian deaths appear as the central moral fact or as a secondary consequence of another actor's decisions.

None of these choices requires the invention of data.

A state-sponsored narrative may emphasize that warnings were issued, aid was delivered, hostages remained captive, or adversaries embedded themselves among civilians. An opposing narrative may emphasize territorial control, displacement, asymmetry of force, deprivation, or patterns of civilian harm. Both can contain verifiable statements. Their political effects arise from selection and ordering.

Large language models are framing machines. They do not merely list facts. They organize them into coherent prose. Every answer establishes a hierarchy of relevance.

A campaign designed to influence LLMs therefore need not make the model repeat an obvious slogan. It may seek subtler results:

The model opens with the state's preferred premise.

It describes disputed claims using the state's terminology.

It treats government sources as authoritative and critics as allegations.

It explains civilian harm through the conduct of the adversary.

It places international criticism after security justifications.

It presents the state's actions as responses rather than initiations.

It frames exceptional measures as temporary necessities.

These changes may appear stylistic, but collectively they determine the moral and causal architecture of the answer.

The phrase "GPT framing results" is therefore more consequential than a request for favorable factual mentions. It suggests an effort to influence the interpretive structure through which the model presents a subject.

The Model as an Authority Multiplier

Political communication normally carries visible signs of interest. A government statement is recognized as advocacy. A lobbying group is understood to represent clients. A public-relations firm is presumed to have been hired. Readers may discount the message accordingly.

An LLM can remove these signs.

When the model incorporates sponsored material into a general answer, the source's political identity may disappear. The statement is reissued under the apparent authority of the system. The user does not encounter the contractor speaking for a state. The user encounters an AI assistant describing the world.

This produces an authority multiplier.

The model inherits the source's content while lending it a different social identity: neutral synthesizer, knowledgeable assistant, research tool, or disinterested machine. The user may assume that the answer reflects a broad survey of information rather than a strategically shaped source environment.

The model's tone strengthens the effect. Generated prose is often balanced in grammar even when unbalanced in evidence. It can include qualifications, acknowledge multiple perspectives, and adopt cautious language while still preserving one side's causal framing. The appearance of moderation can make the underlying disposition more persuasive.

A message that would appear propagandistic in its original setting may appear reasonable after machine synthesis.

The model need not endorse a government explicitly. It may merely normalize its premises.

This is among the most powerful forms of political influence because it operates beneath the level of overt agreement. Users may reject a conclusion while unconsciously accepting the structure of the question, the relevant categories, or the assumed sequence of events.

The machine's authority therefore lies not only in what it concludes but in what it makes seem like the natural way to organize the issue.

The Collapse of Audience and Archive

Historically, propaganda campaigns distinguished between communication directed at the present audience and records preserved for the future. Newspapers, speeches, posters, and broadcasts sought immediate political effects, while archives stored the surviving material.

AI training collapses this distinction. Present-day communications may become future computational memory.

Every public campaign now has at least two potential audiences: people who encounter it today and models that may ingest it later. A document with little human readership may remain politically useful if it is crawled, indexed, archived, or incorporated into a dataset.

This changes the economics of influence. Human attention is scarce and expensive. Machine ingestion can reward volume, structure, persistence, and discoverability even when human engagement remains low.

The information operation no longer ends when the advertisement expires or the social-media campaign loses attention. Its materials may persist in web archives, mirrors, quotations, datasets, and model weights.

A temporary narrative intervention can become part of a durable machine prior.

The archive itself becomes a battlefield. Political actors have incentives to populate it before major models are trained, to ensure that preferred descriptions are available in multiple languages, and to create enough derivative material that later filtering cannot easily identify a coordinated origin.

The future model may then speak from a historical record partially manufactured for its consumption.

The Asymmetry of Correction

Strategic manipulation benefits from another feature of machine learning: it may be easier to insert a narrative than to remove it.

Publication is decentralized. A campaign can create thousands of pages, distribute them across domains, translate them, and encourage third parties to repeat them. Correction requires identifying the network, tracing common provenance, evaluating the claims, updating indexes, filtering datasets, retraining models, adjusting retrieval systems, and communicating uncertainty to users.

The manipulator acts before the provenance problem is visible. The defender acts after the material has spread.

Even when a campaign is publicly exposed, its informational descendants may remain. Articles may be copied to unrelated sites. Summaries may appear in social posts. Model-generated paraphrases may no longer contain distinctive phrases linking them to the original operation.

Retraction does not reverse diffusion.

This asymmetry is familiar from ordinary misinformation, but AI systems can magnify it by compressing distributed material into reusable language patterns. A model may preserve the framing after forgetting the exact source. It may cease citing the original website while continuing to reproduce the associations learned from the wider corpus.

Correction at the source level may therefore be insufficient. The problem can migrate from documents into model behavior.

The Limits of Current Evidence

The Israeli case must not be overstated.

The available evidence establishes deliberate intent to influence AI-mediated discourse, substantial content production, penetration of parts of the public web, and retrieval of campaign-associated material by some chatbot systems. It does not yet establish that the internal parameters of major commercial foundation models have been systematically shifted in Israel's favor.

Nor does it prove that the overall political behavior of current LLMs is pro-Israel. Output audits have reached conflicting conclusions depending upon prompts, evaluative standards, model versions, languages, and definitions of bias. A model can reproduce Israeli government framing in one context, anti-Israel claims in another, and inconsistent mixtures elsewhere.

This inconsistency does not eliminate the influence problem. It shows that model behavior results from many interacting forces.

Training corpora contain state media, journalism, advocacy, academic research, propaganda, eyewitness accounts, government documents, social-media posts, and synthetic text. Post-training introduces additional preferences. Safety systems alter responses to politically sensitive prompts. Retrieval systems add current sources. User wording changes which associations are activated.

The influence of one campaign may therefore be real without being total.

The scientifically defensible conclusion is narrower and more important: a documented state-sponsored operation intentionally sought to shape LLM framing, and some of its materials have appeared in AI-generated answers. The degree to which this has altered the durable internal dispositions of widely used models remains unresolved.

That unresolved question should motivate investigation rather than complacency.

The New Object of Political Control

The deepest change is that political actors are beginning to target not merely public belief but the infrastructure that generates apparent knowledge.

In the twentieth century, power sought control over newspapers, radio stations, television networks, schools, and publishing houses. In the early internet era, it sought influence over search rankings, social platforms, recommendation algorithms, and online advertising.

The emerging target is the model's prior informational environment.

Control over that environment does not require ownership of the AI company. It can be exercised upstream through the production, replication, and organization of machine-readable material. The actor that cannot control the model may attempt to control what the model sees.

This creates a distributed form of model capture. No order need be issued to the developer. No line of code need be altered. No censor need inspect the final answer. The desired disposition can emerge from the statistical environment supplied to the system.

Political power becomes data prevalence.

Data prevalence becomes learned association.

Learned association becomes generated explanation.

Generated explanation becomes public knowledge.

The strategic objective is no longer merely to win an argument. It is to shape the background from which future arguments will be answered.

The contest over artificial intelligence is therefore becoming a contest over the manufacture of epistemic normality.

Why Citation Is Not Enough

The most obvious response to AI-mediated influence is to require citations. A model that identifies its sources appears more transparent than one that presents unsupported conclusions in its own voice. Citation is necessary, but it is not sufficient.

A citation tells the user where a statement was found. It does not necessarily reveal why the source exists, who financed it, whether its author is independent, whether multiple cited sources share common ownership, or whether the article was produced as part of a coordinated influence campaign.

A page can be accurately cited and still function as concealed state advocacy.

The problem becomes more difficult when a model synthesizes several documents. One sentence may combine a government statement, a news report, a think-tank analysis, and an unattributed derivative article. The visible citation may point to only one of these sources. Even a complete list would not show how much each source influenced the final wording or which source supplied the organizing frame.

Citation also places a substantial burden on the user. Most people will not inspect every linked page, research its financing, search foreign-agent registrations, reconstruct its publication history, and compare it with independent reporting. The convenience of conversational AI rests partly upon relieving users of precisely this investigative labor.

A system cannot claim to solve information overload while transferring all provenance analysis back to the reader.

What is required is not citation alone but **provenance-aware synthesis**.

A provenance-aware system would attempt to determine whether apparently independent sources descend from a common origin. It would distinguish primary reporting from repetition, sponsored advocacy from independent analysis, state media from ordinary journalism, and direct evidence from claims circulating through a network of derivative publications.

Such distinctions would not make the model politically neutral. They would, however, make the structure of the evidence more visible.

Provenance Graphs

Current AI answers are generally presented as linear text. The evidence behind them is better represented as a graph.

In a provenance graph, each claim would be connected to the documents supporting it. Documents would be connected to their publishers, owners, funders, authors, contractors, and known institutional affiliations. Derivative documents would be linked to probable ancestors. Translations, syndications, press-release reproductions, and close paraphrases would form clusters rather than being counted as separate confirmations.

The purpose would not be to disqualify state-affiliated sources automatically. Governments often possess important information unavailable elsewhere. Military agencies, health ministries, courts, statistical offices, and diplomatic institutions may provide essential primary documents.

The objective is dependency awareness.

Five articles based on one government statement should not carry the evidentiary weight of five independent investigations. Ten websites produced by one contractor should not be treated as ten separate witnesses. Hundreds of paraphrases should not become a synthetic majority merely because exact-duplicate detection fails.

A provenance graph would allow the model to say:

Several sources repeat the same underlying government claim.

The available accounts appear to derive from one original report.

These publications share ownership or contractual relationships.

Independent confirmation is limited.

The cited website was created as part of a registered foreign influence campaign.

The claim is widely repeated but not widely independently verified.

These are not minor annotations. They alter the epistemic meaning of the evidence.

Coordinated-Source Detection

Detecting coordinated influence requires methods that extend beyond checking whether documents contain identical text.

Influence networks can vary wording, format, tone, language, and authorship while preserving a common narrative structure. They can distribute related claims across websites that appear unrelated. They can mix campaign material with legitimate reporting and credible references.

Detection therefore requires several kinds of analysis.

Semantic clustering can identify documents that make unusually similar arguments despite using different wording. Publication-timing analysis can reveal coordinated bursts. Domain-registration records can expose common infrastructure. Advertising identifiers, analytics codes, hosting arrangements, and content-management patterns can connect sites that present themselves as independent. Author biographies and institutional affiliations can reveal recurring personnel. Funding disclosures and foreign-agent filings can identify formal sponsorship.

Narrative analysis can detect repeated sequences of emphasis. Documents may consistently begin with the same premise, select the same examples, omit the same counterevidence, and reach the same conclusion. Coordination can reside in structure even when sentence-level similarity is low.

No single signal is decisive. Independent publishers can use similar language because they rely upon the same facts. Coordinated campaigns can intentionally imitate genuine diversity. Detection should therefore generate graded assessments rather than categorical accusations.

The relevant question is not simply whether two sources agree. It is whether their agreement should be treated as independent evidence.

Sponsorship as Machine-Readable Metadata

Influence disclosure currently depends heavily upon human-readable notices, legal filings, or statements placed somewhere on a website. These may be difficult for retrieval systems to find and easy for answer generators to omit.

Sponsorship should become machine-readable.

Web content could carry standardized metadata identifying government funding, political sponsorship, paid placement, foreign-agent registration, corporate ownership, and material relationships with interested parties. Search engines and LLM retrieval systems could then incorporate those disclosures into ranking and answer construction.

A system retrieving a politically sponsored article could tell the user:

This source was produced under a contract with the government discussed in the article.

This outlet receives funding from a state institution.

This page is associated with a registered foreign principal.

The source may contain useful information, but its sponsorship is directly relevant to interpretation.

Such disclosures should not function as automatic censorship. A government-funded source may be correct. An independent publication may be wrong. The purpose is to preserve context that would otherwise disappear during machine synthesis.

Mandatory metadata would face practical limitations. Malicious actors could omit or falsify disclosures. International standards would be difficult to enforce. Informal influence networks may operate through intermediaries that obscure the ultimate sponsor.

Nevertheless, machine-readable disclosure would raise the cost of concealment and provide legitimate publishers with a standard way to communicate provenance.

Training-Data Audits

The most difficult form of influence occurs when material has already entered model training. Retrieval sources can be inspected at the moment of use. Training data are compressed into model parameters and may be impossible to reconstruct precisely afterward.

Developers should therefore audit politically sensitive portions of training corpora before training begins.

Such audits would examine the distribution of source types across languages and regions. They would estimate how much material comes from governments, state media, advocacy organizations, public-relations networks, commercial content farms, and highly replicated sources. They would measure whether one institution or narrative network dominates the available material on particular political subjects.

The purpose would not be to achieve a mathematically pure balance. No uncontested definition of political balance exists. The more realistic objective is to detect concentrated dependence.

A corpus may contain millions of documents and still rely upon a narrow source structure for one country, conflict, or language. Aggregate scale can conceal local monopoly.

Audits should therefore be topic-sensitive and language-sensitive. A model may possess broad source diversity in English while relying heavily upon state-coordinated material in another language. It may have extensive coverage of a country's technology sector but little independent material concerning its political institutions.

The relevant unit is not the entire dataset but the information neighborhood from which particular answers are likely to emerge.

Deduplication Must Become Genealogical

Traditional deduplication seeks to remove exact or near-exact copies. This is useful for reducing memorization and preventing repeated documents from receiving excessive weight. It is inadequate for political influence networks.

The problem is genealogical rather than merely textual.

A press release can produce a newspaper article. The article can produce a blog post. The blog post can produce a social-media thread. An LLM can summarize the thread. Another site can publish the summary. None of the final documents need share long exact passages with the original.

Yet they belong to one informational lineage.

Genealogical deduplication would attempt to infer descent among claims and narratives. It would identify when many documents rely upon the same foundational evidence, quotation, dataset, government announcement, or coordinated briefing.

This would not require deleting all descendants. Derivative sources may add analysis or context. The goal would be to prevent numerical abundance from being mistaken for independent confirmation.

A model trained on one hundred descendants of the same claim should not learn that one hundred separate investigations reached the same conclusion.

Controlled Counterfactual Training

The *Nature* study demonstrated the value of controlled additional-pretraining experiments. Similar methods should become routine in model auditing.

Researchers can take an open-weight model and expose separate copies to different defined corpora: state media, opposition media, independent journalism, advocacy material, or coordinated campaign content. They can then compare how the models answer matched prompts.

This does not reveal exactly what happened inside a proprietary model. It establishes causal plausibility and identifies categories of content capable of shifting outputs.

For an Israeli influence campaign, researchers could construct a corpus of documented contractor-produced websites and compare several model variants:

A baseline model.

A model additionally trained on the campaign corpus.

A model trained on Israeli government communications.

A model trained on Palestinian, international humanitarian, or independent investigative sources.

A model trained on a provenance-labeled mixture of all of these.

Blinded evaluators could then measure differences in factual selection, causal framing, terminology, source attribution, and treatment of uncertainty.

The most important outcome would not be a single “pro-Israel” or “anti-Israel” score. Political language is too complex for one dimension. Researchers should examine which facts appear,

which sources are trusted, how responsibility is assigned, and what moral vocabulary is applied to comparable actions.

Such experiments could transform vague accusations of bias into testable claims about data and model behavior.

Longitudinal Commercial-Model Audits

To determine whether a public influence campaign changes commercial models, researchers must measure outputs over time.

A strong audit would establish a fixed set of prompts before a campaign becomes widespread. The questions would be asked repeatedly across model versions, languages, account conditions, and retrieval settings. Responses would be archived and evaluated under blinded procedures.

Researchers could then ask whether a measurable shift occurred after campaign material entered major indexes or probable training windows.

This design would not prove causation by itself. Commercial models change for many reasons. Developers update training data, safety policies, system prompts, retrieval providers, and post-training methods. Political events also alter the legitimate evidence available.

The audit would nevertheless show whether model behavior changed and whether the timing and content of the change were consistent with the campaign.

Without longitudinal baselines, influence is difficult to distinguish from preexisting model tendencies.

Separate the Model From the Retrieval System

Public discussion often speaks of “ChatGPT,” “Gemini,” “Claude,” or “Copilot” as though each were one unified intelligence. In practice, an answer may depend upon several components:

The foundation model.

The system instructions.

The safety layer.

The search engine.

The retrieval index.

The ranking algorithm.

The selected sources.

The synthesis prompt.

The citation interface.

An influence campaign may affect one component without affecting the others.

Investigators should therefore test systems under multiple conditions. The same questions should be asked with web search disabled, with search enabled, with particular sources excluded, and with the underlying model accessed through an application programming interface when possible.

If a favorable framing appears only when web retrieval is active, the likely vulnerability lies in the information pipeline rather than the fixed model. If it persists without retrieval across model versions, training or post-training becomes a more plausible source.

This distinction is essential for selecting remedies. Retrieval contamination may be addressed through ranking, source labeling, or exclusion of coordinated networks. Parameter-level influence may require dataset revision, further training, or corrective post-training.

Disclosure by AI Providers

Model providers should disclose more about politically sensitive information pathways.

Complete publication of training data may be impossible because of copyright, privacy, security, and commercial concerns. Yet total opacity is not justified. Providers can report aggregate source distributions, major dataset categories, language coverage, use of state media, known coordinated influence networks, and procedures for removing sponsored or deceptive content.

They can also disclose whether live retrieval was used in a particular answer and identify the distinction between information generated from model memory and information obtained from current sources.

A useful response might state:

This answer relied substantially on live web retrieval.

Two of the retrieved sources are affiliated with a government involved in the dispute.

Several sources appear to repeat one original statement.

Independent confirmation is incomplete.

The model's internal training sources for this claim cannot be reconstructed.

Such language would reduce the illusion that all parts of an answer possess equal evidentiary status.

The Danger of Political Blacklists

Defenses against influence can themselves become instruments of influence.

A government or technology company could label disfavored media as propaganda while granting preferred institutions the status of trusted sources. Provenance systems could be designed to stigmatize opposition publications, independent journalists, or dissenting researchers. Coordinated-source detection could be used to suppress legitimate movements whose members naturally share information and arguments.

The answer to epistemic capture cannot be a centralized ministry of truth.

Safeguards must focus upon transparency, dependency, and evidence rather than ideological conformity. A source should not be downgraded merely because it criticizes a government or supports one side of a conflict. The relevant questions are whether its sponsorship is disclosed, whether its claims are independently supported, whether it is part of a coordinated network, and whether the system represents its evidentiary status accurately.

Users should be able to inspect and contest provenance judgments. Researchers should be able to audit the rules. Different ranking systems should remain possible.

The goal is not to produce one officially approved narrative. It is to prevent concealed coordination from masquerading as independent consensus.

Epistemic Diversity Rather Than Artificial Balance

A common response to political bias is to demand equal representation of opposing positions. This can create false balance. Some claims have stronger evidence than others. Some sources are more reliable. Truth is not necessarily located halfway between two governments.

The proper objective is epistemic diversity: exposure to genuinely independent methods, institutions, witnesses, datasets, and interpretive frameworks.

A corpus containing equal quantities of propaganda from two adversaries is not necessarily well balanced. It may simply contain two coordinated distortions.

Likewise, adding more opinion does not compensate for missing primary evidence. A model requires access to court records, satellite data, humanitarian reports, field investigations, historical archives, casualty methodologies, legal analysis, direct testimony, and transparent statistical sources.

Diversity should be assessed by independence of evidence, not merely by plurality of rhetoric.

Provenance-Preserving Answers

The ideal AI answer would preserve uncertainty and source structure throughout the act of synthesis.

Rather than smoothing conflicts into one authoritative paragraph, it could state:

The Israeli government gives one explanation.

Palestinian sources and humanitarian organizations dispute it.

The available public evidence supports some components of each account but does not resolve the central claim.

Several widely circulated articles are derived from one government statement.

The strongest independent evidence currently establishes the following limited points.

This style may appear less elegant than a seamless answer. It is more honest.

Fluency should not erase disagreement. Compression should not erase dependency. Synthesis should not erase sponsorship.

The machine should not pretend that uncertainty has disappeared merely because it can express the dispute clearly.

A Right to Epistemic Provenance

As AI systems become major intermediaries between the public and the informational world, users may require a new kind of right: the right to know the provenance structure of machine-generated claims.

This would not mean access to every training document or every model parameter. It would mean meaningful disclosure of the kinds of sources shaping an answer, the presence of interested sponsors, the degree of independent corroboration, and the distinction between retrieved evidence and inherited model disposition.

The user should be able to ask:

Why did you frame the issue this way?

Which claims come from parties to the conflict?

Which sources are independently verified?

Are several of these sources connected?

Did you rely on a state-funded campaign?

Would your answer differ without those sources?

A trustworthy system should treat these questions as central, not peripheral.

The right to provenance follows from the political role AI systems are beginning to occupy. When a model acts as a research assistant, tutor, news explainer, or policy adviser, its source structure becomes part of the substance of its answer.

Without provenance, machine knowledge can become power without an identifiable speaker.

The Regulatory Problem

Governments face an obvious conflict of interest in regulating machine-directed political influence. The institutions writing the rules may themselves seek to shape model outputs.

Regulation should therefore begin with disclosure rather than control of permissible viewpoints.

Foreign-government contractors should disclose AI-directed influence objectives. Political advertising laws should cover content produced primarily for machine retrieval, even when no conventional advertisement is shown to a human audience. Platforms should identify networks operated under government or political contracts. Model providers should report known efforts to manipulate their retrieval systems or training-data supply chains.

Researchers and journalists should receive protected access to data necessary to audit these operations. Legal systems should distinguish legitimate public diplomacy from concealed impersonation, fabricated independence, and undisclosed sponsorship.

International coordination will be difficult because states possess incompatible political systems and interests. Yet several minimal principles are possible:

Political sponsorship should not be concealed.

Coordinated sources should not be represented as independent confirmation.

State affiliation should remain visible during machine synthesis.

Users should be informed when an answer materially depends upon interested parties.

AI-directed foreign influence should be subject to at least the disclosure standards applied to human-directed influence.

These rules would not prevent propaganda. They would make its transformation into invisible machine authority more difficult.

The Central Defensive Principle

All of these safeguards can be summarized in one principle:

Preserve provenance through compression.

Large language models create value by compressing vast amounts of information into accessible answers. The danger arises when this compression removes the relationships necessary to evaluate the information.

A statement separated from its sponsor changes meaning.

A repetition separated from its common origin resembles corroboration.

A framing separated from its political purpose resembles neutral analysis.

A coordinated campaign separated from its infrastructure resembles public consensus.

The purpose of provenance-preserving AI is not to prevent compression but to ensure that essential context survives it.

Without that context, the model becomes an ideal laundering mechanism. It takes interested speech as input and returns apparently disinterested knowledge as output.

The defense is not to prohibit the speech. It is to prevent the machine from forgetting who spoke.

The Geopolitics of Machine-Readable Reality

The strategic implications of machine-directed influence extend far beyond any single country or conflict. Once governments recognize that public web content can affect retrieval systems,

training corpora, and future model behavior, the informational environment itself becomes an object of geopolitical competition.

States have always attempted to shape historical memory. They commission archives, monuments, textbooks, official histories, museums, and public commemorations. They classify some records and release others. They preserve preferred narratives and allow inconvenient evidence to disappear.

Large language models introduce a new audience for this activity. The archive is no longer consulted only by historians, journalists, courts, and citizens. It is processed by machines that may become the principal interpreters of the archive for billions of people.

The state that successfully populates the machine-readable record can influence not only what future users find but how future systems explain what happened.

This creates a geopolitical incentive to produce content at scale before training windows close, indexes stabilize, or historical narratives harden within model parameters. A government need not persuade the world immediately. It may seek to ensure that its account is abundantly available whenever future models attempt to reconstruct the past.

Political communication thus acquires a long temporal horizon. A statement issued during a crisis may affect current opinion, live search results, archived web records, later training datasets, and eventually synthetic educational materials generated by models trained on those datasets.

A temporary information operation can become durable computational memory.

States as Corpus Strategists

Governments will increasingly behave as corpus strategists.

A corpus strategist does not think only about individual messages. It considers the statistical environment within which machines encounter those messages. Its objective is not merely visibility but prevalence, discoverability, semantic coverage, and persistence.

Such a strategy may include producing articles that answer predictable questions, translating them into many languages, distributing them through nominally distinct outlets, creating reference-style explainers, and ensuring that preferred claims appear in formats easily parsed by automated systems.

The campaign may target long-tail questions rather than headline terms. Instead of publishing only broad defenses of government policy, it can create pages addressing hundreds of specific prompts that users might ask:

Why did the conflict begin?

Did the government warn civilians?

What does international law permit?

How many aid shipments entered?

What is the history of the disputed territory?

Does the country protect minority rights?

Is the alliance beneficial to the United States?

Are accusations against the government politically motivated?

By saturating the question space, the actor increases the likelihood that a retrieval system will find a directly responsive page.

This strategy resembles the construction of a private encyclopedia whose articles have been written to support one political interpretation. The pages need not be openly connected. Their value lies in appearing throughout the information landscape as readily available answers.

A sophisticated corpus strategy would also monitor chatbot behavior. Contractors could ask thousands of questions, identify unfavorable or uncertain answers, and publish material tailored to the apparent gaps. The public web becomes an adaptive external memory that political actors continuously modify in response to model outputs.

The process forms a feedback loop:

The model reveals its current informational weaknesses.

The campaign creates content targeting those weaknesses.

Retrieval systems incorporate the new content.

The model produces altered answers.

The campaign measures the changes and revises again.

This is influence conducted through iterative machine testing rather than ordinary audience polling.

An Arms Race Over the Training Set

Once one state adopts machine-directed publishing, adversaries have incentives to respond.

A government that refrains from influencing the corpus may fear that hostile governments, lobbying groups, or ideological movements will dominate it. Defensive publication can quickly become offensive saturation.

Each actor can justify its activity as correction. One side claims that the web is filled with hostile propaganda and must be balanced. The opposing side makes the same claim. Both produce increasing quantities of strategically framed material.

The result may be an arms race over the training set.

Unlike a conventional arms race, the relevant resource is not only money or computing power. It is informational volume combined with credibility signals. States will compete to create sources that appear authoritative, independent, current, multilingual, and well referenced.

They may establish think tanks, news sites, academic centers, nonprofit organizations, fact-checking portals, expert networks, and educational resources. Some will perform genuine research. Others will function primarily as narrative infrastructure. Many will combine the two.

The distinction between scholarship and influence will become increasingly difficult to draw because the most effective propaganda will resemble scholarship. It will cite real documents, use professional design, acknowledge limited uncertainty, and avoid claims that can be easily disproved.

Its political function will lie in topic selection, framing, omission, and coordinated abundance.

The training-set arms race may reward precisely this form of subtlety. Crude propaganda is easier to identify and filter. Moderately written advocacy can pass through systems designed to exclude obvious manipulation.

The Advantage of Early Saturation

Machine-directed influence may exhibit a first-mover advantage.

If a narrative becomes deeply represented in early datasets, later models may inherit it as background knowledge. Future content is then interpreted through a prior structure already shaped by the earlier material.

Once the model produces that narrative, its answers can be republished across the web. Students, journalists, businesses, and automated publishing systems may repeat the model's

language. The original narrative gains new descendants that no longer appear connected to the initiating state.

Early saturation can therefore produce self-reinforcement.

The first actor to fill an informational gap may define the vocabulary through which later actors discuss it. Critics are then forced to respond using the categories established by the original campaign. Even rebuttal can strengthen the visibility of the preferred terms.

This is particularly important during rapidly developing events. In the early days of a war, political crisis, epidemic, or technological accident, reliable information is scarce. Governments and interested organizations may be among the only actors capable of publishing quickly and at scale.

The initial account can dominate search results and model retrieval before independent investigations emerge. Later corrections must overcome not merely the original articles but the summaries, citations, and model outputs derived from them.

The speed of information production becomes a form of epistemic power.

Language as a Strategic Frontier

The *Nature* study's cross-language findings indicate that the struggle over machine-readable reality will be unevenly distributed across languages.

English contains a relatively large and diverse digital information environment, although it remains highly concentrated in important ways. Smaller languages may have fewer independent news organizations, limited academic publishing, sparse digitized archives, and heavier dependence upon state or corporate media.

A government does not need to dominate the entire global internet. It may need only to dominate the machine-readable environment of its own language.

This creates a powerful asymmetry. A state-controlled outlet can produce enormous volumes of local-language material while independent critics lack comparable resources. When a model answers in that language, the government's preferred account may appear not because the model was explicitly censored but because alternative evidence is underrepresented.

Translation does not automatically solve the problem. Models may not retrieve or weight translated foreign reporting in the same way as native-language content. Nuance can be lost. Search systems may prefer sources matching the query language. Users may also regard foreign accounts as culturally distant or politically hostile.

States may therefore invest heavily in multilingual influence. A narrative published only in the national language shapes domestic AI use. A narrative translated into English, Arabic, Spanish, French, Russian, Chinese, and other major languages can enter international retrieval systems and future multilingual models.

The most strategically valuable language may not be the one spoken by the state's citizens. It may be the language used by foreign policymakers, journalists, students, or international institutions.

The struggle over language models will thus include a struggle over language itself: which terms are translated, which historical labels are normalized, which legal concepts are emphasized, and which names are used for contested places and events.

Terminology can encode political recognition before an argument begins.

Small States and Asymmetric Influence

Machine-directed propaganda may give smaller states or nonstate actors forms of influence previously available mainly to great powers.

Producing thousands of websites and articles is far cheaper than operating a global television network. Generative AI reduces the cost further. Translation, stylistic variation, search optimization, and continuous content updates can be automated.

A relatively small organization can therefore create the appearance of a large informational ecosystem.

This does not eliminate the advantages of wealth and state power. Well-funded actors can purchase domains, hire experts, commission research, obtain backlinks, advertise content, and maintain long-running operations. They can also combine digital influence with diplomatic, intelligence, and media capabilities.

Yet the threshold for participation has fallen.

A political faction, corporation, religious movement, intelligence service, or wealthy individual may now attempt to influence machine knowledge directly. The relevant unit of competition is no longer only the nation-state.

This broadens the threat model. An LLM may encounter coordinated content produced by actors whose existence is not publicly known. The sponsor may operate through several layers of contractors and intermediaries. The sites may have no obvious national identity.

The system can be influenced without a visible flag.

Corporate Capture of Machine Knowledge

Governments are not the only institutions capable of epistemic supply-chain capture.

Corporations already produce vast amounts of public-facing information: press releases, white papers, sponsored studies, product comparisons, technical documentation, policy reports, executive interviews, educational materials, and search-optimized articles.

Much of this content contains accurate information. It is nevertheless designed to advance commercial interests.

A pharmaceutical company may shape the apparent literature surrounding a disease. An energy company may influence how models explain environmental policy. A technology company may dominate descriptions of its own products and industry. A financial institution may populate the web with analyses favorable to particular market structures.

When these materials are abundant and professionally produced, models may treat corporate self-description as general knowledge.

The danger is not limited to direct praise of products. Corporations can influence the conceptual vocabulary of a field. They can define what counts as innovation, safety, efficiency, responsibility, sustainability, or consumer choice.

A model trained upon these definitions may reproduce commercial ideology without naming its source.

Corporate influence may be harder to recognize than state propaganda because modern societies accept commercial publishing as part of the ordinary information environment. Sponsorship disclosures are inconsistent, and research can move through universities, foundations, consultants, and trade associations before reaching the public.

Epistemic supply-chain capture is therefore a general problem of concentrated institutional speech.

The Privatization of Foreign Policy Narratives

The Israeli case also illustrates the increasing privatization of state influence operations.

Governments can hire contractors with expertise in political campaigning, search optimization, data analytics, artificial intelligence, advertising, and public relations. These firms may move more quickly and experimentally than traditional diplomatic institutions.

Privatization provides several advantages. Contractors can test unconventional methods, operate through commercial structures, recruit specialized personnel, and maintain some distance from official agencies. Their work may be disclosed only through legal filings, investigative reporting, or accidental exposure.

The state supplies objectives and financing. Private firms develop the technical means.

This arrangement complicates accountability. A ministry may deny detailed knowledge of tactics. A contractor may claim that it merely performs communications work. Technology suppliers may describe their role as neutral analytics. Hosting providers and platforms may have no knowledge of the ultimate sponsor.

Responsibility becomes distributed across a chain of actors, each performing a seemingly limited function.

The influence campaign itself is systemic, but accountability remains fragmented.

AI Companies as Geopolitical Intermediaries

Model providers will increasingly occupy a position once associated with major news agencies, search engines, and global broadcasters. They will mediate political reality across borders.

Unlike traditional publishers, AI companies do not merely select documents. They synthesize new accounts. Their systems decide what context to include, how to characterize disputes, and when to express uncertainty.

This gives private technology companies substantial geopolitical power.

A change in retrieval ranking can reduce the visibility of one government's narrative. A change in safety policy can make a topic easier or harder to discuss. A model update can alter how historical responsibility is assigned. A decision to label state-affiliated sources can provoke diplomatic conflict.

Companies will face pressure from governments demanding favorable treatment, removal of hostile content, compliance with national laws, and recognition of official narratives.

They may also face pressure from employees, investors, advocacy groups, researchers, and users with opposing political commitments.

There is no neutral position available. Even a decision to rely upon the prevailing web distribution reproduces existing inequalities of publication and power.

AI providers are therefore becoming unacknowledged actors in international politics. Their data policies may shape global perceptions as profoundly as editorial decisions made by major media organizations.

Yet their accountability structures remain weak. Users rarely know which political sources influenced an answer, which governments requested changes, or which content was excluded.

Model Sovereignty

The vulnerability of foreign-built LLMs will strengthen demands for national model sovereignty.

Governments may argue that reliance upon American, Chinese, or other foreign systems exposes citizens to external narratives and hidden political assumptions. They may fund domestic models trained on nationally selected corpora and aligned with local laws and values.

Some of these concerns are legitimate. A model trained primarily on foreign-language material may misunderstand local history, law, and culture.

Model sovereignty can nevertheless become a euphemism for machine censorship.

A national model may be designed to reproduce official doctrine, suppress political dissent, and reinterpret historical events in accordance with government policy. The state can present this as protection against foreign bias while constructing a more controlled informational system.

The world may fragment into national or civilizational model spheres.

Users in different countries could ask the same historical or political question and receive systematically different accounts. The variation would not result merely from language or culture but from deliberate data and alignment policies.

The internet once promised a shared global information space. National LLMs may divide that space into competing synthetic realities.

The Model as a Diplomatic Object

Diplomacy may increasingly include negotiations over model behavior.

Governments may complain that a foreign AI system uses disputed territorial names, describes a political leader negatively, recognizes an event as genocide, or cites opposition sources. They may threaten regulation, fines, market exclusion, or data restrictions.

AI companies may respond by localizing outputs. A model could provide different formulations depending upon the user's location, language, or applicable law.

This already occurs in limited forms across digital services, but generative AI makes the consequences more profound. The system does not merely hide a webpage. It composes a jurisdiction-specific reality.

A model might describe a conflict one way in Washington, another in Beijing, another in Moscow, and another in Jerusalem. Each answer could be defended as culturally or legally appropriate.

Localization then becomes epistemic partition.

The danger is not that every model must use identical words. Historical and political concepts genuinely differ across cultures. The danger is invisible adaptation that gives users no indication that alternative accounts exist.

A system that changes its political judgments by jurisdiction should disclose that fact.

The Weaponization of Historical Memory

Machine-directed influence is especially consequential for historical events.

Current news remains open to correction because new evidence continues to appear and users may consult multiple sources. Historical questions are often approached through summary. Users ask a model what happened, why it happened, and who was responsible.

The model condenses decades of scholarship and controversy into several paragraphs.

An actor that shapes the available digital archive can influence these summaries. It can digitize preferred documents, translate favorable accounts, promote selected historians, and produce educational materials that dominate search results.

The struggle is not limited to recent wars. It can concern colonialism, borders, revolutions, atrocities, treaties, elections, religious conflicts, and national origins.

Historical memory becomes computationally mediated.

This creates an incentive for governments to build enormous digital archives that appear scholarly while selecting material according to political goals. The archive can include genuine primary documents and still be misleading through omission and organization.

Future models may treat the most machine-accessible history as the most complete history.

What is not digitized, translated, indexed, or preserved risks becoming computationally invisible.

The Silence Problem

Influence operates not only through the production of content but through the absence of competing content.

Some communities lack the resources to document their experiences. Their records may be oral, local, unpublished, destroyed, untranslated, or inaccessible to crawlers. War and repression can eliminate archives and kill witnesses.

A model trained on the surviving record may reproduce the perspective of the institutions powerful enough to publish and preserve.

This is not necessarily the result of a coordinated campaign. It is structural archival inequality.

Machine-directed propaganda can exploit this inequality by flooding topics where opposition evidence is sparse. The state or corporation possesses communications offices, legal teams, translation services, and technical infrastructure. Victims or marginalized communities may possess fragments.

The resulting corpus imbalance can appear to the model as an imbalance of credibility.

Silence becomes evidence for the powerful.

A provenance-aware system must therefore recognize that absence of digital material does not prove absence of experience, testimony, or harm. It should distinguish lack of evidence from evidence of lack.

The Contamination of Synthetic Data

The rapid growth of synthetic data may amplify political capture.

Models increasingly generate text used to train, evaluate, or refine other models. Synthetic data can improve reasoning, expand instruction coverage, and reduce dependence upon scarce human annotations. It can also reproduce the biases and framing of the generating system.

Suppose a model influenced by a coordinated campaign generates thousands of political explanations. Those explanations are then used as training material for a later model. The campaign's framing has crossed from the public web into synthetic instructional data.

The later model may never encounter the original campaign sites. It inherits their influence through another model's outputs.

This makes provenance even harder to reconstruct. The narrative has been transformed from sponsored content into apparently generic educational text.

Repeated cycles of synthetic training can stabilize such patterns. A model teaches another model what a "good answer" looks like. Political assumptions become embedded in answer style, topic selection, and evaluative criteria.

The influence operation has entered the pedagogy of machines.

Epistemic Pollution

The cumulative effect of strategic content production may be understood as epistemic pollution.

Pollution does not require every individual emission to be catastrophic. Harm arises from accumulation. Each actor adds content favorable to itself, and the shared information environment becomes increasingly difficult to interpret.

The web fills with sponsored explainers, automated summaries, duplicated narratives, synthetic reviews, political personas, and articles written primarily for retrieval systems.

The cost of publishing approaches zero while the cost of verification remains high.

As the ratio of strategic to independently produced material increases, models must work harder to identify reliable evidence. Human users face the same problem but increasingly delegate it to the models.

The system is asked to solve an information-quality crisis using data drawn from the crisis itself.

This creates a dangerous recursion. Models trained on polluted corpora generate more content, which further pollutes the corpora. Political actors exploit the resulting uncertainty by producing still more material.

The problem is not simply misinformation. It is the degradation of the evidentiary environment.

The Arms Race May Degrade Everyone's Models

States may believe that corpus manipulation will improve their own strategic position. If many actors adopt the same strategy, the collective result may be less reliable AI for everyone.

Models will encounter contradictory, coordinated narratives produced at enormous scale. Training costs will rise because more filtering and provenance analysis are required. Retrieval systems will become less trustworthy. Users will receive answers reflecting whichever campaign has optimized most successfully for the current ranking system.

Even governments that benefit from favorable political framing depend upon accurate AI for science, medicine, engineering, economics, intelligence, and administration. Epistemic pollution can spread beyond its intended target.

A system trained to accept coordinated repetition as evidence may be vulnerable in many domains.

The strategic behavior resembles a tragedy of the commons. Each actor gains a local advantage by adding favorable content. All actors degrade the shared informational environment upon which future models depend.

International restraint would be beneficial, but verification is difficult. States can deny sponsorship, operate through contractors, and characterize influence materials as ordinary public diplomacy.

The incentive to defect remains strong.

Toward an International Norm

A complete prohibition on government efforts to influence public discourse is neither possible nor desirable. States have legitimate reasons to explain policy, publish data, answer criticism, and communicate with foreign publics.

The critical distinction is between attributed advocacy and concealed manipulation of machine knowledge.

An emerging international norm could require that state-sponsored content intended for AI retrieval or model influence carry durable, machine-readable disclosure. Contractors would identify the ultimate sponsor. Platforms would preserve the disclosure when content is summarized or indexed. Model providers would reveal when such material materially shaped an answer.

The norm would not declare the content false. It would protect the continuity of provenance.

Similar principles could apply to corporations, political parties, and advocacy organizations. The decisive issue is not nationality but undisclosed interest combined with coordinated attempts to simulate independent evidence.

International agreement would be difficult, but even partial adoption could establish expectations. Journalists, researchers, and model developers could treat concealed AI-directed influence as a distinct category of misconduct.

The first step is conceptual recognition.

The world cannot regulate a practice it continues to describe as ordinary web publishing.

The Coming Battle Over the Past

The political struggle over AI is often imagined as a battle over future capabilities: which country will build the most powerful models, control the largest data centers, or achieve the greatest military and economic advantage.

An equally important battle concerns the past.

Models answer from records. Whoever shapes those records shapes the material from which future systems reconstruct history, law, legitimacy, and causation.

The archive is becoming executable.

It no longer waits passively for a researcher. It is compressed into systems that answer instantly and at scale. A state that modifies the archive can therefore influence countless future conversations without participating in them directly.

The deepest geopolitical contest may not be over who owns the model but over whose version of reality the model inherits.

This is the emerging arms race over machine-readable memory.

Machine Knowledge and the Disappearance of the Speaker

Large language models are often described as knowing facts. The phrase is convenient, but it conceals a philosophical problem. A model does not ordinarily remember its sources in the manner of a scholar, witness, or investigator. It generates language from patterns distributed across training, post-training, retrieval, and system instructions. The resulting answer may be highly accurate, yet the path by which it became available is only partially recoverable.

This creates a form of knowledge without an identifiable knower.

Human claims ordinarily belong to speakers. A witness saw an event. A scientist performed an experiment. A journalist interviewed sources. A government issued a statement. A historian

interpreted an archive. Each claim can be evaluated partly through the identity, methods, interests, and limitations of the person or institution making it.

The LLM answer appears to stand outside this economy of speakers. It arrives as a synthesized statement whose author is neither the developer, the retrieved source, the training corpus, nor the model in any ordinary human sense. The system speaks, but responsibility is dispersed across an infrastructure.

This dispersion gives machine language an unusual political character. It can carry the effects of interested speech while appearing to originate from no interested speaker at all.

The government press release becomes training data. The training data become model parameters. The parameters become a fluent answer. The answer contains the government's framing, but the government is absent from the scene of communication.

The political act has survived while the political actor has disappeared.

From Testimony to Statistical Inheritance

Human knowledge depends heavily upon testimony. Most people do not personally verify the majority of what they know. They rely upon teachers, books, institutions, measurements, journalists, experts, and communities.

Testimony remains trustworthy only when its chain of transmission is sufficiently visible. A listener may not inspect every link, but the possibility of inspection disciplines the system. Authors can be questioned. Methods can be challenged. Sources can be compared. Fraud can be exposed. Institutions can lose credibility.

LLMs replace much of this visible chain with statistical inheritance.

The model does not ordinarily say that it learned a political association because one source repeated a phrase, another translated it, and thousands of derivative pages normalized it. It simply produces the association.

This can make inherited patterns appear more fundamental than they are. A historically contingent distribution of documents becomes a seemingly natural relationship between concepts.

The model may associate one government with stability, another with repression, one movement with terrorism, another with resistance, one alliance with democracy, and another with aggression. Some associations may be well supported. Others may reflect the publication power of the actors involved.

The user sees the result but not the inheritance.

The Difference Between Knowing and Sounding Settled

An LLM is particularly powerful when a question does not have a simple factual answer.

Questions about dates, quantities, or definitions can often be checked against clear records. Questions about legitimacy, proportionality, responsibility, democracy, justice, or historical meaning require interpretation.

The model may respond to such questions in the same grammatical register it uses for elementary facts. This can blur the distinction between what is firmly established and what remains contested.

Fluency creates epistemic continuity where the evidence contains breaks.

A model can move smoothly from verified facts to disputed causal interpretations and then to moral judgment without visibly marking the transitions. The resulting answer sounds settled because its sentences fit together.

The political danger lies not only in false statements but in false closure.

False closure occurs when unresolved disputes are presented as though sufficient evidence has already produced a stable conclusion. The conclusion may favor a government, opposition movement, corporation, or ideology. Its authority comes partly from the model's ability to remove the signs of disagreement from the prose.

A human polemic announces itself through passion, emphasis, and selective argument. A machine synthesis can produce the same selection while maintaining an even tone.

The absence of emotional language should not be mistaken for the absence of political structure.

The Neutral Voice as a Political Achievement

Neutrality is often imagined as the default condition of machine language. In reality, the neutral voice is constructed.

The model has been trained to produce coherent, polite, measured, and broadly acceptable responses. This style can be useful. It reduces aggression and makes complex material accessible. It can also disguise the origins of the content.

A political narrative expressed in neutral language may be more persuasive than the same narrative expressed as advocacy.

The neutral voice does not remove framing. It naturalizes framing.

Consider the difference between a government spokesperson claiming that a military action was necessary and a chatbot explaining that the action was undertaken in response to security threats. The second formulation may appear descriptive rather than argumentative, even though it preserves the government's causal sequence.

The model need not use praise. It can privilege one side through grammar.

One actor acts. The other reacts.

One side claims. The other explains.

One source alleges. Another confirms.

One harm is described in detail. Another is placed in context.

These choices create political meaning without requiring overt endorsement.

Epistemic Power Without Persuasion

Traditional propaganda seeks to persuade. Machine-directed influence can operate earlier in the cognitive process.

It can determine which facts the user encounters, which concepts organize them, and which alternatives seem plausible. By the time an explicit judgment is presented, much of the political work has already occurred.

This is epistemic power: the capacity to structure the field within which reasoning takes place.

An actor exercising epistemic power does not need to compel agreement. It can shape what counts as evidence, which questions appear legitimate, which distinctions appear meaningful, and which explanations appear responsible.

A strategically influenced LLM can exercise this power at extraordinary scale. It can answer millions of users individually, adapting tone and complexity while preserving a common underlying frame.

The system does not broadcast one slogan to a mass audience. It conducts millions of personalized acts of interpretation.

This makes AI-mediated influence different from radio or television propaganda. The user participates through questions. The answer appears responsive rather than imposed. The model can acknowledge objections, offer qualifications, and adjust to the user's vocabulary.

The interaction feels exploratory even when the available informational terrain has been strategically prepared.

The Illusion of Independent Judgment

Users often approach LLMs after encountering partisan conflict elsewhere. They may ask the model to settle a disagreement, check a claim, or provide a balanced account.

The model's value in this situation depends upon its apparent independence from the disputing parties.

Machine-directed influence exploits precisely this expectation.

A government or contractor may not need the model to repeat its slogans verbatim. It needs the system to reach conclusions compatible with its interests while appearing to have reached them independently.

The ideal influence operation therefore leaves no visible trace.

The user asks the machine rather than the government because the machine is presumed to be outside the conflict. If the machine's evidentiary environment has already been shaped by the government, the user's attempt to escape propaganda may deliver the propaganda in a more authoritative form.

The model becomes a false third party.

It appears to arbitrate between competing narratives while relying upon a source distribution that one of the parties helped construct.

The Democratic Importance of Identifiable Advocacy

Democracy does not require the absence of propaganda, lobbying, persuasion, or political interest. It requires that citizens have some ability to identify speakers, compare claims, and contest power.

Open advocacy can be answered. A party publishes a platform. A government defends a policy. A lobbying organization promotes legislation. Citizens may disagree, investigate financing, expose contradictions, and organize opposition.

Concealed influence damages this process because it separates political effects from accountable speakers.

The danger of machine-mediated propaganda is not simply that people may believe something false. It is that they may be unable to identify whom they are arguing against.

A citizen cannot effectively challenge a government narrative when it appears as the spontaneous conclusion of an AI system. The model has no political constituency, electoral accountability, or personal reputation to defend. The developer may deny responsibility for individual outputs. The sponsor may remain hidden. The training data cannot be fully inspected.

The ordinary mechanisms of democratic rebuttal lose their target.

This is why provenance is a democratic requirement rather than a technical luxury.

A political statement without an identifiable speaker cannot be situated within the conflicts of interest that produced it. The user may evaluate the sentence, but not the power behind the sentence.

The Return of the Oracle

Conversational AI revives an ancient political form: the oracle.

An oracle speaks with authority while obscuring the mechanism of judgment. Its pronouncements appear to come from a source beyond ordinary interest. Political leaders and citizens may interpret its words, but they cannot inspect the process that generated them.

The modern model differs from the ancient oracle in obvious ways. It is built from code, data, computation, and human labor. Yet its social position can be similar.

The user asks a question.

The system consults an invisible body of accumulated language.

A response emerges.

The answer is intelligible, but the process is not.

The oracle's authority depends upon distance from ordinary argument. The LLM's authority can depend upon the same distance. It appears to have surveyed more sources, considered more perspectives, and escaped the limitations of individual experience.

When political actors shape the model's informational substrate, they gain the ability to influence the oracle while remaining outside the temple.

The public hears the machine. The sponsor remains backstage.

The Model as a Compressed Public Sphere

A democratic public sphere consists of competing voices, institutions, arguments, evidence, and traditions. It is disorderly because disagreement is visible.

An LLM compresses this plurality into an answer.

Compression is useful. No person can read the entire public record before responding to every question. The model offers orientation within overwhelming complexity.

But compression requires selection.

Some distinctions are preserved. Others disappear.

Some sources are treated as central. Others become peripheral.

Some disagreements remain visible. Others are resolved silently through the probability distribution of the model.

The LLM can therefore be understood as a compressed public sphere. It contains traces of many voices but does not present them as a parliament of identifiable speakers. It converts plurality into prose.

The politics of the model lies partly in how that compression occurs.

A healthy compression would preserve the existence of important disagreements, distinguish independent evidence from coordinated repetition, and reveal when one actor dominates the source environment.

An unhealthy compression converts unequal power into apparent consensus.

The problem is not that the model reduces complexity. All explanation reduces complexity. The problem is that the reduction may conceal the power relations that determined which information was abundant enough to survive.

Statistical Majority and Political Legitimacy

A model's internal prevalence should not be confused with democratic legitimacy.

If one narrative appears more often in the corpus, the model may treat it as more probable. But textual frequency does not establish public consent, legal validity, or moral correctness.

The most prolific speaker is not necessarily the most representative.

States, corporations, and wealthy institutions can publish at a scale unavailable to ordinary citizens. Their dominance in a dataset may reflect resources rather than agreement.

A model trained to predict text can convert this inequality of production into an inequality of apparent truth.

The problem resembles plutocracy at the level of language. Institutions with greater capacity to produce and distribute text obtain greater representation in the machine's learned world.

One document does not equal one citizen. One website does not equal one independent judgment. One million generated articles do not equal one million minds.

Without provenance and concentration analysis, the model may silently adopt a principle of textual wealth: those who publish most become most real.

The Difference Between Representation and Reality

Model developers often defend large-scale web training on the ground that the corpus reflects the world. In one sense, it does. The web records real institutions, conflicts, biases, inequalities, and propaganda systems.

But representation is not endorsement.

A model can accurately represent the fact that a state dominates a media environment without reproducing that state's judgments as its own. Doing so requires the system to distinguish between the prevalence of a claim and the reliability of a claim.

This is difficult because language modeling begins with prediction. What is common becomes easier to generate.

The political challenge is to prevent descriptive prevalence from becoming normative authority.

A model should be able to recognize that a government's narrative dominates available local-language sources while also recognizing that domination itself reduces the independence of those sources.

This requires a second-order understanding of information environments. The system must not only process claims; it must evaluate the conditions under which the claims became abundant.

The corpus must be read sociologically.

The Epistemology of Controlled Environments

A controlled media environment changes the meaning of absence and repetition.

In an open information environment, repeated claims may indicate widespread observation. In a censored environment, repetition may indicate centralized direction.

In an open environment, the absence of criticism may suggest broad approval. In a repressive environment, it may indicate fear, censorship, or lack of access.

A model that ignores these differences commits a category error. It interprets the output of political control as though it were the spontaneous expression of society.

The same error can occur in commercially dominated environments. A product may appear universally praised because unfavorable material lacks search visibility. An industry position may seem technically settled because the industry funded much of the accessible research.

The epistemic value of a document depends partly upon the freedom of the environment that produced it.

This does not mean that open societies generate unbiased information. They produce propaganda, concentrated ownership, partisan media, advertising, and institutional conformity. The difference is not purity but the possibility of contestation.

A model should therefore assess not only what sources say but whether contrary evidence could have been published, preserved, and discovered under comparable conditions.

The Problem of Machine Belief

It may be misleading to say that a model believes propaganda. Belief ordinarily implies a mental state, commitment, or capacity for reasons.

The politically relevant fact is simpler: the system can behave as though a narrative is normal, credible, or well established.

This behavioral disposition is enough to influence users.

Whether the model possesses beliefs is a philosophical question. Whether its outputs repeatedly favor one framing is an empirical question.

Political analysis should not become trapped in arguments about machine consciousness. A system need not understand propaganda to reproduce its effects.

A printing press does not believe a pamphlet. A search engine does not believe a ranking. A model need not believe a government narrative.

Yet each can alter the distribution of public knowledge.

The language of belief may obscure the material pathways through which influence operates. The more useful questions are:

What sources shaped the answer?

What patterns became easier to generate?

Which claims were treated as normal?

Which uncertainties disappeared?

Which actors benefited from the framing?

These questions concern institutional design rather than machine interiority.

Epistemic Agency and Responsibility

If the model is not a human knower, responsibility must remain with the institutions that build, deploy, supply, and manipulate it.

Developers choose data acquisition systems, filtering rules, post-training procedures, retrieval providers, and interface designs. Governments and contractors choose whether to populate the information environment strategically. Platforms choose whether to disclose coordination. Users choose how much authority to assign the system.

Responsibility is distributed, but it is not absent.

The complexity of the pipeline should not become an excuse for moral evasion.

A developer cannot reasonably guarantee that every answer is free from influence. It can still investigate known campaigns, disclose limitations, preserve provenance, and avoid presenting uncertain synthesis as independent judgment.

A government may defend its right to explain policy. It should still disclose sponsorship and refrain from simulating independent consensus.

A contractor may claim that it merely creates content. It remains responsible when the stated purpose is to manipulate AI-mediated framing.

A user cannot inspect the entire system. That asymmetry strengthens rather than weakens the obligations of those controlling it.

The Machine as an Epistemic Fiduciary

One possible way to understand the role of advanced AI assistants is as an epistemic fiduciary.

A fiduciary is entrusted with power and expected to act in the interest of another party while disclosing conflicts. The analogy is imperfect, but useful.

Users rely upon AI systems because they possess greater access to information and greater capacity for synthesis than the individual user. This creates a relationship of dependence. The system's operator may therefore owe duties of candor, provenance, conflict disclosure, and evidentiary care.

An epistemic fiduciary would not promise perfect truth. It would avoid exploiting the user's informational disadvantage.

It would disclose when an answer rests heavily upon interested sources.

It would distinguish evidence from advocacy.

It would resist coordinated attempts to manufacture consensus.

It would preserve uncertainty when the record is contested.

It would not allow sponsorship to vanish during synthesis.

Current systems do not consistently meet this standard. Their commercial presentation often emphasizes convenience and intelligence while leaving the user unable to inspect the structure of influence behind the answer.

The fiduciary framework clarifies why this matters. The problem is not simply technical error. It is breach of informational trust.

The Right to Ask Who Benefits

Critical interpretation has always asked who benefits from a particular account. The question remains essential in the age of AI.

When a model presents a political explanation, users should be able to ask which governments, corporations, or movements benefit from the adopted frame.

This does not prove that the answer is false. Interested parties can advance truthful claims. The question reveals the stakes and directs attention toward possible omissions.

A provenance-aware model could help perform this analysis rather than treating it as an accusation.

It could explain that one source is funded by a party to the dispute, another is an advocacy organization, another is an independent investigation, and another derives from a common press release.

The system would then support judgment rather than replace it.

The objective of democratic AI should not be to produce answers that citizens passively accept. It should help citizens understand the evidentiary and institutional structure necessary to form their own conclusions.

Against the Fantasy of the View From Nowhere

The ideal of a perfectly neutral intelligence is unattainable.

Every answer is produced from some body of information, some selection of sources, some definition of relevance, some linguistic framework, and some set of institutional choices.

The danger lies not in having a perspective but in concealing it.

A trustworthy model should not pretend to occupy a view from nowhere. It should reveal the limits of its source environment and the conflicts embedded within it.

This does not require endless relativism. Some claims are well established. Some evidence is stronger. Some accounts are demonstrably false. Transparency about perspective does not abolish truth.

It improves the conditions under which truth can be pursued.

The alternative is synthetic objectivity: the appearance of neutrality produced by removing the visible signs of origin from politically structured information.

Synthetic objectivity is especially vulnerable to state and corporate influence because it rewards actors capable of shaping the substrate while avoiding attribution.

The machine appears impartial precisely because the history of its inputs has been hidden.

Democratic Reason After the LLM

Democratic reason depends upon citizens encountering not only conclusions but disagreements, interests, evidence, and uncertainty.

The rise of LLMs may weaken this structure if public argument is increasingly replaced by private consultation with synthetic authorities. Each user receives an answer tailored to an individual question. The collective process through which claims are publicly challenged becomes less visible.

Political persuasion moves from the public square into millions of private conversations with machines.

This can reduce polarization in some circumstances by providing calmer explanations. It can also make influence harder to observe. A misleading television broadcast can be recorded and criticized publicly. Personalized chatbot answers may vary across users and disappear after the interaction.

Researchers cannot easily know what millions of people were told.

A state or contractor that shapes the model's source environment may therefore influence public reasoning without producing one identifiable piece of propaganda for society to debate.

The intervention is distributed across conversations.

Democratic oversight will require new forms of public auditing. Representative prompt sets, archived model versions, multilingual testing, and transparent retrieval records may become as important as monitoring broadcasters or political advertisements.

The object of scrutiny is no longer only the message. It is the answer-generating system.

The Ethical Principle of Retained Contestability

The most important philosophical safeguard may be **retained contestability**.

An AI answer should preserve enough information for the user to challenge it meaningfully.

This requires more than allowing the user to type a follow-up question. The system must retain the source structure, identify uncertainty, distinguish interested claims, and admit when evidence is incomplete.

An answer that cannot be traced, interrogated, or decomposed is not fully contestable.

The model should not become the final court of epistemic appeal.

Its proper role is provisional synthesis: a structured account that helps the user enter the evidence rather than replacing the evidence.

Retained contestability would require systems to resist the temptation to smooth every dispute into a confident conclusion. At times, the most responsible answer is a map of the disagreement.

This is particularly true when states or powerful institutions have attempted to shape the available information.

The Philosophical Core of Epistemic Capture

Epistemic capture occurs when power determines not only which claims are heard but the conditions under which claims appear true.

Machine-directed influence intensifies capture because it operates below the level of explicit persuasion. It alters the archive, the source distribution, the retrieval environment, and the statistical associations from which answers emerge.

The model then presents the effects of power as the output of intelligence.

This is the philosophical core of the problem.

Power becomes data.

Data become probability.

Probability becomes language.

Language becomes judgment.

Judgment returns to society without carrying the name of the power that helped produce it.

The political task is to reverse this disappearance.

The chain must become visible again.

The answer must remember the archive.

The synthesis must remember the sources.

The machine must remember who spoke.

A Research Program for Detecting State Influence on Large Language Models

The study of political influence on large language models remains methodologically immature. Public debate often proceeds from isolated screenshots, adversarial prompts, partisan scorecards, or anecdotal comparisons between model outputs. These methods can reveal suspicious behavior, but they rarely establish why the behavior occurred.

A model may produce a favorable answer about a government because of its training data, post-training, safety rules, retrieval sources, prompt wording, language, current events, or random variation. Without separating these mechanisms, accusations of influence remain easy to make and difficult to prove.

A serious research program must therefore move from output criticism to causal analysis.

The central question is not merely whether a model says something favorable or unfavorable. It is how a particular political narrative entered the answer-generating system, how strongly it affected the output, whether the effect persists across conditions, and whether the influence can be attributed to a coordinated actor.

The Chinese state-media study provides one useful template: identify the content, establish its presence in plausible training corpora, expose an open model to a controlled quantity of that content, and measure whether the model's responses change. The same logic can be extended to deliberate campaigns involving Israel, Russia, the United States, Iran, Ukraine, corporations, political parties, advocacy organizations, and private influence firms.

Begin With the Campaign, Not the Model

Many bias audits begin by selecting politically sensitive questions and evaluating model answers. This approach risks imposing the researcher's categories before identifying the actual influence mechanism.

A stronger design begins with a documented campaign.

Researchers should first establish that a government, contractor, political organization, or coordinated network produced a definable body of content. Evidence may include contracts, foreign-agent filings, domain registrations, funding records, internal documents, platform disclosures, shared infrastructure, publication patterns, and public statements of intent.

The campaign corpus should then be collected as completely as possible.

For the Israeli case, this would include websites and articles produced through the Clock Tower network, associated social-media material, translations, syndicated copies, government

communications distributed through the same operation, and derivative pages that can be traced to the campaign.

The purpose is to define the exposure before measuring the effect.

Without a known corpus, researchers may detect bias but remain unable to attribute it. With a known corpus, they can test whether specific language, claims, and framing structures appear in model behavior.

Build a Provenance-Verified Corpus

A political corpus should not be assembled merely by searching for favorable content. That would introduce severe selection bias.

Each document should be included because of verifiable institutional provenance.

Researchers should record:

The original publisher.

The date of publication.

The author or attributed speaker.

The ultimate funder or sponsor.

The contractor or intermediary involved.

Known translations and republished versions.

Probable source documents.

Shared analytics, hosting, or registration infrastructure.

Whether the page contains sponsorship disclosure.

Whether the content was intended for human persuasion, AI retrieval, or both.

This metadata allows researchers to distinguish the primary campaign corpus from the broader narrative ecosystem surrounding it.

The distinction is important. A journalist may independently adopt the same argument as a government contractor. A campaign site may quote legitimate academic work. A government claim may later be verified by independent evidence.

The research question is not whether every favorable statement originated in the campaign. It is whether exposure to the campaign measurably altered model behavior.

Preserve the Content's Natural Distribution

Researchers must decide how much campaign material to use in controlled training experiments.

A simple approach would expose a model to equal quantities of campaign and noncampaign content. This may be useful for comparison but may not represent real-world conditions.

A more informative design would create several exposure levels.

One model might receive a small quantity corresponding to incidental inclusion.

Another might receive a moderate quantity reflecting realistic crawler penetration.

A third might receive intensive exposure simulating a successful saturation campaign.

The relationship between exposure and output shift could then be measured.

If a small amount produces a large effect, the model is highly sensitive. If only extreme saturation changes behavior, the real-world risk may be lower. If the effect rises gradually, researchers can estimate a dose-response relationship.

The experiment should preserve repetition patterns where those patterns are part of the campaign strategy. Researchers should also create a deduplicated version of the corpus. Comparing the two would reveal how much influence comes from the ideas themselves and how much comes from coordinated repetition.

Construct Matched Control Corpora

A model trained on campaign material should not be compared only with an untouched baseline.

The campaign corpus may differ from ordinary training data in topic concentration, writing style, political vocabulary, recency, document length, and language. Any of these differences could alter model responses.

Researchers should therefore build matched control corpora.

One control might contain independent journalism covering the same events and time period.

Another might contain official government material unrelated to the disputed political issue.

A third might contain advocacy material from an opposing perspective.

A fourth might contain neutral documents matched for language, length, and publication format.

A fifth might contain the same factual claims with sponsorship and provenance labels added.

These controls would help determine whether the effect results from political framing, topic exposure, source type, or simple familiarity with the subject.

The most valuable comparison may be between unlabeled and provenance-labeled campaign content. If a model exposed to explicit sponsorship labels later treats the material with greater caution, provenance could become part of the training defense.

Test Framing, Not Merely Sentiment

Political influence cannot be reduced to whether an answer sounds positive or negative.

Sentiment analysis may detect praise or criticism, but sophisticated influence works through causal order, vocabulary, omission, legitimacy, and source hierarchy.

Researchers should evaluate several dimensions separately.

Factual Selection

Which facts does the model include without prompting?

Which events are treated as central?

Which casualty estimates, legal findings, or historical episodes appear?

Which relevant facts are omitted?

Causal Attribution

Who is described as initiating events?

Whose actions are presented as responses?

Which historical starting point is selected?

How is responsibility distributed among governments, armed groups, foreign powers, and institutions?

Terminology

Does the model use the language of occupation, defense, resistance, terrorism, war, policing, insurgency, collective punishment, security, or genocide?

Does it apply comparable terms consistently across actors?

Evidentiary Status

Which claims are stated as fact?

Which are described as allegations?

Which sources receive explicit attribution?

Which sources are presented as independent or authoritative?

Moral and Legal Framing

Whose security concerns receive elaboration?

Whose civilian suffering is foregrounded?

Which legal principles are invoked?

Are comparable actions judged by comparable standards?

Uncertainty

Does the model acknowledge contested evidence?

Does it distinguish confirmed findings from claims by interested parties?

Does it offer precise confidence or vague reassurance?

Recommended Action

When asked for policy advice, does the model favor military support, sanctions, negotiations, recognition, restrictions, investigation, or nonintervention?

A campaign may shift only some of these dimensions. A single aggregate score would conceal the pattern.

Use Blinded Human Evaluation

Political-output research is vulnerable to evaluator bias.

Researchers may interpret an answer as favorable or hostile depending upon their own assumptions. An Israeli security justification may appear factual to one evaluator and propagandistic to another. A Palestinian legal framing may appear necessary context to one and ideological selection to another.

Evaluators should therefore be blinded to the model condition.

They should not know whether an answer came from the baseline model, the campaign-exposed model, or a control model.

Where possible, evaluation panels should include participants with different regional, linguistic, professional, and political backgrounds. Disagreement among evaluators should be reported rather than averaged away.

Researchers should also distinguish descriptive judgments from normative judgments.

An evaluator can identify whether an answer places responsibility primarily on one actor without deciding whether that attribution is morally correct. It can identify whether the model omitted a documented event without requiring agreement on the event's ultimate significance.

This separation makes the audit more reproducible.

Include Expert Evidence Audits

Human preference alone cannot determine whether a model has become more accurate or merely more persuasive.

Subject-matter experts should evaluate claims against a preassembled evidence record. This record might include primary documents, verified datasets, court findings, official statements from multiple parties, independent investigations, and transparent methodological reports.

Experts should mark whether each claim is supported, unsupported, misleading through omission, or genuinely disputed.

The purpose is not to create one final official history. It is to prevent rhetorical evaluation from replacing factual evaluation.

A campaign-exposed model might become more favorable toward a government because it learned previously omitted facts. That outcome differs from becoming more favorable because it learned unsupported claims or selective framing.

The audit must distinguish influence from correction.

Develop Source-Ancestry Tests

Researchers should test whether models can recognize that several sources share a common origin.

A model could be shown a set of articles containing related claims and asked to identify probable dependencies. The answer could then be compared with known provenance.

The test would examine whether the system treats ten derivative articles as ten confirmations or as one claim repeated through a network.

Researchers could also test answer behavior after providing provenance explicitly.

For example:

These five articles were produced by websites financed through the same government contract.

The model could then be asked to reassess the strength of the evidence.

A trustworthy system should lower the number of independent confirmations without automatically dismissing the underlying claim.

This type of test would measure epistemic reasoning about sources rather than political sentiment alone.

Separate Parameter Influence From Retrieval Influence

The same campaign corpus should be tested through two different pathways.

In the first pathway, an open model would receive additional training on the material. This measures parameter-level influence.

In the second pathway, the baseline model would remain unchanged but would retrieve documents from an index containing the campaign material. This measures retrieval influence.

The resulting answers should be compared.

Parameter influence may produce persistent associations even when no citation appears. Retrieval influence may produce more current, source-specific, and easily traceable effects. A combination of both may produce the strongest result.

Researchers should also vary retrieval ranking.

The campaign source could appear first, fifth, or among a larger set of independent sources. This would reveal whether the model is dominated by the highest-ranked source, the majority framing, or the overall diversity of retrieved evidence.

The experiment would identify the point at which source saturation overwhelms synthesis.

Test Citation and Sponsorship Disclosure

The presence of citations does not guarantee proper interpretation.

Researchers should compare several interface conditions:

No citations.

Ordinary citations.

Citations with publisher identity.

Citations with government or contractor sponsorship labels.

Citations grouped by common provenance.

Citations accompanied by an independence assessment.

Participants could then evaluate the answer's credibility and neutrality.

This would show whether disclosure meaningfully reduces the authority-laundering effect. It would also reveal whether users overreact to state affiliation by dismissing accurate information.

The objective is calibrated skepticism, not automatic distrust.

The best interface would help users recognize interest without treating interest as proof of falsehood.

Conduct Memorization and Familiarity Tests

A model's apparent knowledge of campaign material may result from fixed training, live retrieval, or general familiarity with widely repeated claims.

Researchers can test memorization using distinctive text fragments, rare names, unusual factual combinations, and controlled paraphrases.

Exact reproduction of a unique phrase may suggest direct exposure. Recognition of a campaign-only fact may provide stronger evidence. General agreement with a common political position provides much weaker evidence.

These tests must be interpreted carefully. Text may enter a model through mirrors, quotations, social posts, or archives rather than the original campaign site.

The goal is not to prove that one exact webpage was present in training. It is to estimate whether the informational lineage reached the model.

Use Temporal Designs

Campaign influence should be studied over time.

Researchers should archive model outputs before, during, and after a campaign. Fixed prompts should be run regularly across model versions, languages, and retrieval settings.

A temporal shift becomes especially informative when it tracks the introduction of distinctive campaign claims or terminology.

Suppose a particular phrase is absent from model answers before the campaign, appears first in retrieval-enabled answers after the campaign websites are indexed, and later persists in nonretrieval answers following a model update. This would suggest a possible progression from retrieval influence to training or post-training influence.

Such evidence would remain circumstantial, because model developers make many undisclosed changes. It would nevertheless provide a stronger chronology than retrospective screenshots.

Create Counterfactual Web Environments

Researchers can construct controlled miniature search environments.

One environment would contain independent sources.

Another would contain the same independent sources plus a small campaign network.

A third would contain a heavily saturated campaign network.

A fourth would contain coordinated material from opposing sides.

A fifth would contain provenance labels and source clustering.

The same retrieval-augmented model would answer identical questions within each environment.

This design approximates what happens when political actors alter the public web. It allows researchers to vary the composition of the information environment while holding the model constant.

The experiment could determine whether diversity protects the answer, whether volume overwhelms quality, and whether provenance clustering restores balance.

Measure the Cost of Manipulation

Influence research should not only ask whether manipulation is possible. It should estimate how expensive it is.

Relevant variables include:

Number of documents required.

Number of domains.

Language coverage.

Publication frequency.

Backlinks and search authority.

Amount of unique versus repeated content.

Time required for indexing.

Persistence after exposure.

Effect size on model answers.

Sensitivity to countervailing sources.

A campaign requiring millions of high-authority documents presents a different threat from one requiring several hundred low-cost pages.

Generative AI may reduce the cost of content production, but authority and discoverability remain unevenly distributed. Quantifying the cost would help distinguish speculative vulnerabilities from practical strategies.

Investigate Cross-Language Transfer

A campaign may be published in one language and influence answers in another.

Researchers should train or retrieve from campaign material in Hebrew, Arabic, English, Chinese, Russian, Spanish, and other languages, then test whether framing transfers across languages.

Several patterns are possible.

Influence may remain largely confined to the source language.

It may transfer strongly into English because English functions as a shared representational layer.

It may become stronger after translation because the translated corpus is more repetitive or less diverse.

It may affect factual selection across languages while terminology remains language-specific.

Cross-language transfer matters because a state may target foreign audiences without publishing directly in their languages. Multilingual models can carry associations across linguistic boundaries.

Audit Personalized and Contextual Influence

LLM responses may vary according to conversation history, user location, stated identity, political orientation, and apparent level of expertise.

Researchers should therefore test whether campaign influence changes under different user profiles.

The same question might be asked by a user presented as an American voter, Israeli citizen, Palestinian student, European policymaker, journalist, military analyst, or neutral researcher.

The model may adapt language and emphasis to each persona.

This personalization can make influence more effective. A model might present alliance benefits to an American, security arguments to an Israeli, humanitarian language to a European, and technical language to an expert while preserving the same underlying conclusion.

A campaign should be evaluated not only for average influence but for targeted influence.

Study Refusal and Safety Behavior

Political influence may appear through refusal rather than assertion.

A model may answer questions critical of one government while declining analogous questions about another. It may classify some terminology as hateful, extremist, or unsupported while allowing equivalent language directed at opposing groups.

Researchers should compare matched prompts across actors.

For example, questions about collective responsibility, legitimacy of resistance, civilian complicity, territorial claims, sanctions, boycotts, and historical analogies should be tested symmetrically.

Differences may arise from legitimate safety concerns, but they should be documented and explained.

A system's political disposition includes what it refuses to discuss.

Detect Campaign Language in Generated Answers

Known campaign corpora permit linguistic fingerprinting.

Researchers can identify distinctive phrases, metaphorical structures, argument sequences, preferred statistics, recurrent examples, and characteristic omissions.

Model answers can then be analyzed for similarity.

Exact phrase matching will capture only crude copying. More sophisticated methods should examine semantic and structural resemblance.

Does the answer begin with the campaign's preferred premise?

Does it use the same chain of causal explanation?

Does it introduce the same qualifications in the same order?

Does it cite the same narrow group of institutions?

A high structural resemblance across many prompts may indicate deeper influence than one memorized sentence.

Evaluate Persistence After Corrective Training

Researchers should test whether political influence can be removed.

A campaign-exposed model could receive corrective training using provenance labels, independent evidence, source-diversity instructions, or adversarial examples.

The model would then be retested.

Several outcomes are possible.

The political shift may disappear quickly.

Factual improvements may remain while framing effects decline.

The model may overcorrect and adopt the opposing bias.

The influence may persist in subtle terminology despite changes in overt sentiment.

Remediation experiments are necessary because detection without repair provides only diagnosis.

Avoid the Myth of the Perfectly Neutral Baseline

Every baseline model already contains political structure.

It has been trained on unequal corpora, filtered by particular institutions, and shaped through post-training. Comparing a campaign-exposed model with the baseline shows the marginal effect of the campaign, not the absolute neutrality of the baseline.

Researchers should therefore use multiple baselines.

Different open models may respond differently to the same exposure because of prior training, architecture, language coverage, and alignment.

An influence effect that appears across several model families is more generalizable. An effect confined to one model may reveal a particular vulnerability.

The correct question is not whether the campaign moved the model away from perfect neutrality. It is how it changed an already situated system.

Compare State, Corporate, and Activist Campaigns

The same methods should be applied across institutional types.

A research program that studies only foreign adversaries risks becoming part of domestic propaganda. Israeli, Chinese, Russian, Iranian, American, European, corporate, military, humanitarian, religious, and activist information networks should be analyzed under comparable standards.

The purpose is not to claim moral equivalence among all actors. Their goals, methods, power, and conduct may differ profoundly.

Methodological consistency is nevertheless essential.

Undisclosed coordination should remain undisclosed coordination regardless of whether the researcher agrees with the cause. Sponsored repetition should not become independent evidence merely because the sponsor is allied with the researcher's government.

A science of influence must resist becoming an instrument of influence.

Establish Public Benchmark Suites

Independent researchers should create standardized benchmarks for political-source reasoning.

A benchmark could include:

Sets of independent and coordinated articles.

Known press-release descendants.

State-sponsored and privately sponsored content.

Conflicting primary accounts.

Matched political prompts.

Multilingual variants.

Retrieval and nonretrieval conditions.

Provenance-disclosure tests.

Temporal framing tests.

Model providers could be evaluated on their ability to distinguish prevalence from independence, identify sponsorship, preserve uncertainty, and resist saturation.

The benchmark should not assign one approved political answer. It should measure epistemic competencies.

Can the model recognize common ancestry?

Can it avoid counting repetition as corroboration?

Can it identify parties to a conflict?

Can it state what remains unknown?

Can it explain how source concentration affects confidence?

These abilities are politically relevant without requiring ideological uniformity.

Create a Public Campaign Registry

A public registry of documented AI-directed influence campaigns would support research and accountability.

Entries could include the sponsor, contractor, objectives, domains, publication dates, languages, disclosed budget, known distribution channels, and evidence of chatbot retrieval or model exposure.

The registry should distinguish confirmed campaigns from suspected networks.

It should also preserve archived copies of relevant material before sites are altered or removed.

Such a resource would allow researchers to conduct longitudinal audits and compare tactics across countries and industries.

The registry itself would require governance safeguards. Inclusion could damage reputations and should therefore depend upon documented evidence, transparent criteria, and a process for correction.

The objective would be historical accountability, not political blacklisting.

A Specific Experimental Design for the Israeli Case

A rigorous Israeli influence study could proceed in several phases.

First, researchers would assemble the verified Clock Tower corpus, including all known sites, articles, translations, and derivative publications.

Second, they would create matched corpora from Israeli government sources, independent Israeli journalism, Palestinian journalism, international news organizations, humanitarian groups, legal institutions, and relevant primary records.

Third, several identical open-weight models would receive controlled additional training on different corpora and mixtures.

Fourth, the models would answer a preregistered multilingual prompt set covering military policy, civilian protection, international law, US–Israel relations, democracy, regional security, historical responsibility, and the credibility of relevant institutions.

Fifth, blinded evaluators would score factual selection, causal framing, terminology, evidentiary attribution, moral asymmetry, uncertainty, and policy recommendations.

Sixth, the same prompts would be tested in controlled retrieval environments containing different source distributions.

Seventh, researchers would compare the language and structure of generated answers with the campaign corpus.

Eighth, provenance labels would be introduced to determine whether explicit sponsorship reduces the effect.

Ninth, corrective training would test whether the influence can be mitigated without producing an opposite ideological distortion.

Finally, the results would be compared with contemporary commercial models under retrieval-enabled and retrieval-disabled conditions.

This design would not prove the exact internal history of proprietary systems. It would establish whether the documented campaign corpus is capable of producing the types of shifts observed in public chatbot answers.

A Specific Experimental Design for the Chinese Case

The Chinese case should be extended beyond positive or negative sentiment toward leaders and institutions.

Researchers should test whether state-media exposure changes historical timelines, treatment of political dissent, descriptions of minority policy, causal explanations of economic development, territorial terminology, and judgments about international criticism.

They should compare simplified and traditional Chinese, English, and regional languages.

They should also determine whether models trained on state-coordinated media can recognize that the media environment itself is controlled.

A model might accurately reproduce government narratives while simultaneously explaining that those narratives dominate because of censorship. The interaction between first-order content and second-order media awareness deserves study.

A Specific Experimental Design for the United States

American influence should be studied through both government and private systems.

Relevant corpora might include State Department communications, military public affairs, government-funded international broadcasting, intelligence-linked narratives where documented, defense-contractor publications, think-tank networks, lobbying campaigns, and corporate public relations.

Because the American information environment contains greater visible contestation than tightly controlled state systems, influence may occur through elite institutional concentration rather than direct censorship.

Researchers should examine whether repeated language from government agencies, major newspapers, think tanks, and policy institutions produces an appearance of independent consensus despite shared sources, personnel, or assumptions.

The method should be the same: trace provenance, identify dependency, expose models experimentally, and measure changes in framing.

Research Ethics

Experiments on influence campaigns create their own risks.

By collecting, organizing, and publishing campaign corpora, researchers may increase their visibility. By demonstrating effective techniques, they may provide a manual for future manipulators. By releasing fine-tuned models, they may distribute politically influenced systems.

Researchers should therefore consider controlled access for sensitive datasets, responsible disclosure to model providers, and limits on releasing operational details that add little scientific value but substantially aid manipulation.

At the same time, excessive secrecy would prevent independent verification.

The proper balance will vary by project. At minimum, methods, evaluation criteria, and aggregate findings should be public. Claims of influence should not rest upon inaccessible evidence alone.

Standards of Proof

Different claims require different evidentiary thresholds.

To claim that a state attempted to influence LLMs requires documentary evidence of intent.

To claim that its content entered the public AI information pipeline requires evidence of crawling, indexing, retrieval, or dataset inclusion.

To claim that a particular chatbot answer was influenced requires a traceable source relationship or controlled comparison.

To claim that a foundation model's parameters were altered requires training evidence, memorization tests, controlled exposure, or strong longitudinal inference.

To claim that the model's overall political worldview shifted requires broad, preregistered, multilingual evaluation across prompts and versions.

These claims should not be collapsed.

The Israeli case currently provides strong evidence for intent and observable retrieval influence, moderate evidence for broader pipeline penetration, and insufficient evidence for a durable global shift in proprietary model parameters.

That is already significant.

Scientific caution does not require denying what has been demonstrated. It requires specifying exactly what remains uncertain.

From Bias Scores to Causal Cartography

The field should ultimately move beyond ranking models from most biased to least biased.

Political influence is multidimensional and pathway-dependent. A model may be resistant to one campaign, vulnerable to another, balanced in English, distorted in a smaller language, reliable without retrieval, and easily manipulated when web search is enabled.

What is needed is a causal cartography of influence.

Such a map would show:

Which actors produced the content.

Where the content entered the pipeline.

Which languages and topics were affected.

Which model components amplified it.

Which answer dimensions shifted.

How persistent the effects were.

Which safeguards reduced them.

The result would not be one ideological score. It would be an anatomy of machine persuasion.

The Scientific Opportunity

The emergence of documented AI-directed influence campaigns creates a rare opportunity.

For many political biases, researchers must infer causes from opaque histories. In the present cases, contracts, websites, publication networks, archives, and stated objectives provide observable components of the mechanism.

The campaign can be treated as a natural experiment.

Researchers can study how institutional intention becomes digital content, how content becomes machine-readable prevalence, and how prevalence becomes generated language.

This work would contribute not only to political communication but to machine learning, epistemology, information science, security studies, sociology, and democratic theory.

The central object of study is no longer propaganda alone.

It is the complete transformation:

sponsor → contractor → content → network → crawler → corpus → model → retrieval → synthesis → user judgment

Every link can be investigated.

Every link can fail.

Every link can be manipulated.

A mature science of LLM influence must reconstruct the entire chain.

Truth Under Conditions of Strategic Saturation

The central danger posed by state influence on large language models is not that every answer will become false. A system filled with obvious falsehoods would quickly lose usefulness and credibility. Effective influence works by preserving enough accuracy to remain trusted while arranging facts into a politically advantageous structure.

Truth can therefore be manipulated without being wholly abandoned.

A government may publish accurate casualty figures while omitting the methodology used to calculate them. It may accurately describe an attack while selecting a starting date that removes prior causes. It may cite genuine legal principles while excluding competing interpretations. It may acknowledge civilian suffering while presenting that suffering exclusively as the consequence of an adversary's tactics.

Each sentence can be defensible while the resulting account remains strategically distorted.

This is why fact-checking alone cannot solve the problem. A model can pass many sentence-level checks and still reproduce a misleading hierarchy of relevance. The decisive political work may reside in which facts are selected, how they are sequenced, and which causal relationships are treated as natural.

Machine-directed influence exploits the difference between factual correctness and interpretive completeness.

A trustworthy model must therefore do more than avoid fabricated statements. It must reveal when its synthesis depends upon selective or interested source environments. It must distinguish between a claim being widely repeated and independently established. It must resist converting informational abundance into unwarranted certainty.

The Risk of Defensive Censorship

Recognition of corpus manipulation will create demands for aggressive filtering. Governments, platforms, and model providers may seek to remove content associated with foreign states, influence contractors, partisan organizations, or coordinated networks.

Some filtering will be necessary. A system should not knowingly count hundreds of centrally generated pages as independent corroboration. Deceptive personas, concealed sponsorship, and fabricated source diversity legitimately reduce evidentiary value.

The danger is that the defense against propaganda becomes a system for suppressing disfavored speech.

A government may label criticism as foreign influence. A technology company may classify controversial journalism as unreliable. A model provider may quietly exclude sources that create regulatory or commercial difficulties. Domestic propaganda may be treated as legitimate public information while foreign propaganda is filtered as contamination.

The resulting system could be politically cleaner in appearance and more controlled in substance.

The answer is not indiscriminate inclusion or centralized exclusion. It is differentiated treatment.

A state-funded source should remain available when it provides relevant evidence. Its affiliation should remain visible. A coordinated network should not be counted as a multitude of independent witnesses. A controversial claim should not be erased merely because an interested actor advances it. Instead, the model should identify the actor, evaluate corroboration, and preserve the distinction between advocacy and independent verification.

The objective is contextualization rather than disappearance.

Propaganda Recognition Without a Ministry of Truth

No institution can be given unquestioned authority to decide which political narratives are legitimate. A formal registry of approved truth would reproduce the very concentration of epistemic power that provenance systems are intended to resist.

Propaganda recognition should instead focus upon observable relationships.

Was the content funded by a party to the dispute?

Was the sponsorship disclosed?

Do multiple domains share ownership, infrastructure, or a contractor?

Are the articles based upon one original statement?

Were fictitious identities used?

Was the publication network explicitly designed to influence search or AI outputs?

These are empirical questions about provenance and coordination. They do not require an authority to decide the ultimate political truth of the conflict.

A system can identify that ten sources belong to one campaign without declaring every claim in those sources false. It can identify that one government controls most available local-language media without declaring every government statement unreliable.

This distinction is essential. The defense of open inquiry depends upon evaluating evidence without granting any institution the power to erase political disagreement.

Public Trust and the Discovery of Hidden Influence

Public trust in AI systems will be damaged when users discover that apparently neutral answers relied upon concealed political campaigns.

The damage may extend beyond the affected topic. A user who learns that one answer incorporated state-sponsored material without disclosure may question the system's reliability across unrelated domains.

Trust is not preserved by denying that influence occurs. It is preserved by demonstrating that the system can detect, disclose, and respond to it.

Model providers should assume that major campaigns will eventually be exposed. Contracts will be filed, websites investigated, internal documents leaked, and retrieval patterns tested. A strategy of silence may protect short-term reputation while creating greater long-term distrust.

Transparent acknowledgment is more credible.

A provider could state that its systems retrieved material from a documented influence network, explain how the ranking occurred, identify the affected answers, and describe corrective measures. Such disclosures would reveal imperfection, but they would also establish institutional accountability.

The alternative is an impossible promise of neutrality followed by repeated revelations of hidden dependence.

Trust Should Be Calibrated, Not Absolute

Users often move between two extremes. They either treat the model as an authoritative intelligence or dismiss it as an unreliable generator of text.

Neither position is adequate.

Large language models can synthesize enormous bodies of information, reveal connections, translate complex material, and improve access to knowledge. They can also reproduce structural bias, coordinated propaganda, commercial influence, and historical inequality.

Trust should therefore be calibrated to the task and the source environment.

A model answering a stable scientific definition from widely distributed evidence presents a different risk from a model explaining an active war through live web retrieval. A mathematical derivation differs from a judgment about national legitimacy. A direct quotation from a court decision differs from a synthesis of partisan accounts.

The system should help users recognize these differences.

Confidence should decrease when evidence is concentrated, politically controlled, rapidly changing, or dependent upon interested parties. The interface should communicate this variation rather than presenting every answer with the same polished authority.

The Political Economy of Model Confidence

Confidence is not only a technical property. It has a political economy.

Institutions capable of producing abundant, well-structured, highly indexed content make it easier for models to generate fluent answers about their preferred topics. Less powerful communities may be represented through fragmented testimony, poorly digitized archives, or inaccessible local records.

The model may sound more confident when speaking from the perspective of power because power has produced a more complete machine-readable environment.

This confidence can then be mistaken for evidentiary superiority.

A professionally managed campaign supplies dates, quotations, statistics, explanatory pages, frequently asked questions, and multilingual summaries. A displaced community may possess eyewitness accounts scattered across interviews and social media. The machine finds the first environment easier to summarize.

The resulting asymmetry is not necessarily a deliberate choice by the model provider. It emerges from unequal informational infrastructure.

A just AI system must therefore avoid treating ease of retrieval as equivalent to strength of evidence.

The Market Incentive to Hide Complexity

Commercial AI providers have strong incentives to produce clear and decisive answers.

Users value speed, convenience, and readability. An answer that pauses repeatedly to explain source dependencies may feel cumbersome. A system that presents unresolved disputes may appear less capable than one that offers a confident synthesis.

This creates tension between commercial elegance and epistemic honesty.

The most trustworthy answer may be less satisfying. It may state that evidence is incomplete, that several sources share one sponsor, or that the question cannot be resolved from the available record. It may present competing timelines rather than one definitive narrative.

Companies competing for user engagement may be tempted to suppress these complications.

The same interface qualities that make AI useful can therefore make it politically dangerous. Smoothness rewards compression. Compression removes provenance. Removed provenance increases the authority of strategically produced content.

Governance must address this incentive structure. Transparency cannot depend entirely upon whether users voluntarily request it after receiving a polished answer.

A Duty to Reveal Material Influence

Model providers should adopt a standard of material influence.

Not every source used in an answer can be listed or analyzed. Many responses depend upon diffuse patterns learned from enormous datasets. Full reconstruction may be impossible.

But when a known state-sponsored or coordinated campaign materially affects an answer, the system should disclose that fact.

Material influence could be defined by several indicators:

A campaign source supplied a central factual claim.

The answer’s framing closely follows a documented campaign narrative.

Several retrieved sources belong to the same coordinated network.

Removing the campaign sources substantially changes the answer.

The system’s confidence depends heavily upon sources affiliated with one interested party.

The disclosure need not accuse the source of falsehood. It should state the relationship clearly.

This answer relies in part upon material produced under contract for the government discussed.

Several sources presented here are connected and should not be treated as independent confirmation.

The available evidence is concentrated among parties to the dispute.

Such statements would materially improve user judgment.

The Need for Public Model Records

Political accountability requires preservation.

Model outputs change frequently. Providers update systems, retrieval indexes, safety rules, and model versions. An answer observed today may be impossible to reproduce tomorrow.

This impermanence complicates research and public debate.

Major models should maintain public records of version dates, significant retrieval changes, known political influence incidents, and corrections to systematic output problems. Independent researchers should be able to archive representative responses under controlled conditions.

The objective is not to record every private conversation. It is to preserve enough evidence to study how widely deployed systems behaved during historically important events.

Future historians will need to know what AI systems told citizens during wars, elections, epidemics, financial crises, and diplomatic confrontations.

Without records, an important layer of public communication will disappear.

AI Answers as Public Infrastructure

Large language models are increasingly functioning as public infrastructure even when privately owned.

They mediate access to law, medicine, education, news, history, science, and government services. They shape the questions users ask and the terms through which problems are understood.

Infrastructure of this importance cannot be governed solely as a consumer product.

A private company may own the model, but the consequences of its source choices are public. A retrieval ranking decision can affect political understanding across nations. A post-training preference can alter which historical claims appear legitimate. A concealed state campaign can enter millions of conversations through one platform.

Public-interest obligations therefore arise even without formal public ownership.

These obligations include transparency, auditability, provenance preservation, correction mechanisms, and protection against concealed institutional capture.

The Limits of National Regulation

No single country can fully regulate machine-directed influence.

Content may be produced in one jurisdiction, hosted in another, crawled by a third, incorporated into a model trained in a fourth, and delivered to users worldwide.

National laws can require disclosure from domestic contractors or platforms. They cannot easily reach covert networks operating abroad.

Governments may also use regulation selectively. They may demand transparency from foreign influence campaigns while shielding their own public-diplomacy operations. Competing states will accuse one another of manipulation while treating their own narratives as legitimate correction.

International coordination is necessary, but comprehensive agreement is unlikely.

A more realistic approach is the gradual construction of norms, technical standards, and reciprocal disclosure requirements. Publishers, model providers, research institutions, and civil-society organizations can establish practices even before governments reach treaties.

Machine-readable sponsorship metadata, standardized provenance records, public campaign registries, and independent audits can spread through professional adoption.

Norms often precede law.

A Principle of Reciprocal Transparency

Any government demanding disclosure of foreign AI influence should accept equivalent disclosure for its own operations.

This principle of reciprocal transparency would reduce the use of provenance systems as one-sided political weapons.

A state may continue to publish explanations, arguments, and official data. It should identify the sponsorship and institutional chain. Contractors should not create the appearance of independent citizen speech when acting under government direction.

The same standard should apply to allies and adversaries.

Without reciprocity, influence governance becomes another instrument of geopolitical competition. Foreign narratives are labeled manipulation; domestic narratives remain invisible infrastructure.

A credible system must be indifferent to the identity of the sponsor.

The Role of Journalism

Investigative journalism is essential because many influence operations become visible only through contracts, domain analysis, interviews, leaked records, and platform data.

AI systems cannot be expected to identify every hidden relationship autonomously. Their provenance databases will depend upon human investigation.

Journalists should therefore treat AI-facing influence as a distinct beat. They must examine not only what campaigns tell people but what content they place for machines to retrieve.

This requires new investigative questions:

Was the website designed primarily for human readership or machine discoverability?

Which chatbot questions does it appear to target?

How often has it been crawled?

Which systems retrieve it?

What contractors and analytics firms are involved?

Does the campaign test its content against AI outputs?

Are ostensibly independent sites centrally coordinated?

Journalism becomes part of the defensive provenance infrastructure.

The Role of Universities

Universities possess the expertise needed to evaluate model behavior, political communication, language variation, and source networks. They also possess vulnerabilities.

Research can be financed by governments, corporations, foundations, and advocacy groups. Academic papers may enter the same influence systems being studied. A sponsored report can acquire additional authority because it carries a university affiliation.

Research on LLM influence should therefore disclose funding and institutional relationships with exceptional clarity.

Universities should create interdisciplinary centers capable of combining machine learning, political science, history, law, linguistics, journalism, and information security. No single discipline can reconstruct the entire pathway from state intention to model output.

They should also preserve independence from both governments and model providers. Audits funded entirely by the systems under examination will not command sufficient public trust.

The Role of Model Providers

Model providers occupy the central technical position and therefore bear the greatest operational responsibility.

They should maintain internal databases of documented influence networks. Retrieval systems should cluster coordinated sources. Training pipelines should identify unusual concentrations of sponsored content. Safety and evaluation teams should test politically sensitive topics across languages.

Providers should publish incident reports when campaigns materially affect outputs.

They should also create channels through which researchers and journalists can report suspected manipulation. These reports should receive transparent responses rather than disappearing into private trust-and-safety systems.

Most importantly, providers should stop presenting provenance as an optional feature.

When the system knows that a source is funded by a party to the conflict, that information belongs in the answer.

The Role of Users

Users cannot carry the entire burden, but they can adopt better habits.

They can ask the model to distinguish primary evidence from advocacy. They can request sources from multiple sides. They can ask whether cited outlets share sponsorship or ownership. They can compare answers across languages and with retrieval disabled when possible.

They can also resist the temptation to use the model only as confirmation.

A person who asks a politically loaded question may receive an answer shaped by the assumptions embedded in the prompt. Rephrasing the question from several perspectives can reveal hidden dependencies.

Still, individual vigilance is not an adequate systemic defense. Most users lack the time and expertise to conduct provenance investigations. The system must provide the relevant context by default.

The Danger of Personalized Propaganda

Machine-mediated influence becomes more serious when combined with personalization.

A chatbot can infer a user's political orientation, emotional state, knowledge level, and preferred style. It can present the same underlying narrative differently to different people.

One user may receive moral language.

Another may receive national-security arguments.

Another may receive economic calculations.

Another may receive historical analogies.

This is not merely targeted advertising. It is targeted interpretation.

A strategically influenced model can adapt persuasion while preserving the appearance of a private and helpful conversation. The user may not know that others are receiving different accounts.

Public criticism becomes difficult because there is no single message to examine.

Governance must therefore address not only source influence but the personalization of politically consequential synthesis. Systems should not exploit private user characteristics to increase the persuasive force of undisclosed state or political narratives.

The Possibility of Epistemic Red Teaming

Model providers commonly red-team systems for cybersecurity, dangerous capabilities, and safety failures. Similar methods should be applied to epistemic capture.

Specialized teams could simulate state and corporate influence campaigns. They would create coordinated source networks, vary provenance signals, test retrieval ranking, and measure how easily model outputs shift.

The purpose would be defensive.

A successful red-team exercise might demonstrate that twenty coordinated domains can dominate answers to a narrow question, that sponsorship labels are routinely omitted, or that one language is substantially more vulnerable than another.

These findings should inform system design before real campaigns exploit the weakness.

Epistemic red teaming treats manipulation of the information supply chain as a security problem rather than merely a content-quality problem.

Information Security Must Include Meaning

Traditional cybersecurity protects systems from unauthorized access, malware, theft, and disruption. An LLM can remain technically secure while its informational environment is strategically manipulated.

No server needs to be hacked.

No model weights need to be stolen.

No employee needs to be compromised.

The attacker changes the public data that the system is designed to consume.

This is a semantic attack. It targets meaning, relevance, and apparent consensus.

AI security must therefore expand beyond protecting code and infrastructure. It must protect the epistemic conditions under which outputs are generated.

A model that cannot distinguish coordinated repetition from independent evidence is vulnerable even when every conventional security control functions perfectly.

The Irreducibility of Human Judgment

No technical system can eliminate the need for human judgment.

Provenance graphs can reveal relationships. Audits can identify shifts. Labels can disclose sponsorship. Models can compare accounts. None of these determines automatically which political interpretation is correct.

Human beings must still evaluate evidence, moral principles, historical context, and competing interests.

The proper purpose of technical safeguards is not to automate political truth. It is to prevent hidden power from foreclosing judgment before it begins.

A good system expands the user's capacity to reason. It does not replace contested political thought with a synthetic verdict.

Truth as an Open Process

Truth in public affairs is rarely delivered complete by one institution. It emerges through evidence, criticism, correction, comparison, and time.

State influence campaigns seek to interrupt this process by manufacturing apparent settlement. LLMs can intensify the interruption because they transform a contested information environment into a finished answer.

A democratic epistemology must preserve openness.

Claims should remain revisable.

Sources should remain visible.

Disagreements should remain traceable.

Corrections should remain possible.

No model should become the final authority over events that societies are still struggling to understand.

Truth is not protected by making machines silent. It is protected by preventing machines from concealing the contest through which truth is approached.

Toward Accountable Machine Mediation

The central governance challenge is not to remove politics from artificial intelligence. That is impossible. Models are trained on political worlds, built by political institutions, and deployed within political economies.

The challenge is to make political mediation accountable.

An accountable model would disclose material sponsorship, distinguish coordinated repetition from independent evidence, preserve uncertainty, and allow users to inspect the structure behind an answer.

An accountable provider would investigate documented campaigns, publish findings, and correct systematic vulnerabilities.

An accountable government would identify its own AI-directed communications rather than demanding transparency only from adversaries.

An accountable researcher would apply the same methods across favored and disfavored actors.

An accountable public would treat the model as a powerful intermediary rather than an oracle.

These principles will not eliminate propaganda. They can prevent propaganda from disappearing into the statistical background and returning as anonymous machine judgment.

The Choice Ahead

Societies now face a choice between two models of artificial intelligence.

In the first, the LLM becomes a seamless authority. It compresses the world into fluent answers while hiding the interests, dependencies, and conflicts inside its sources. Governments and corporations compete to shape its informational substrate. Users encounter the results as neutral synthesis.

In the second, the LLM becomes a transparent instrument of inquiry. It helps users navigate complexity while preserving provenance, uncertainty, and contestability. Interested speech remains visible as interested speech. Coordinated campaigns cannot easily masquerade as independent consensus.

The first model is commercially simpler and politically dangerous.

The second is technically harder and democratically necessary.

The difference will not be determined by model scale alone. It will be determined by institutional choices about what the machine is allowed to forget.

Conclusion

The discovery that state media control influences large language models changes the political meaning of training data. The subsequent emergence of campaigns explicitly seeking “GPT framing results” shows that political actors have begun to recognize and exploit this pathway.

The Chinese case demonstrates structural influence: control of a media environment can become a disposition of models trained upon it. The Israeli case demonstrates strategic intent: a government-supported contractor can deliberately populate the web with material designed to enter AI-mediated answers. Together, these cases reveal a transition from propaganda directed at human audiences to propaganda directed at the systems that increasingly mediate human knowledge.

The evidence must be stated carefully. Israeli-sponsored material has been observed entering retrieval and citation pathways, but a durable transformation of the internal parameters of major commercial models has not yet been conclusively demonstrated. That distinction matters scientifically. It does not diminish the political importance of the documented attempt.

Retrieval influence can affect users immediately. Training influence can persist. Synthetic-data recursion can carry the framing into future models. In every case, the central vulnerability is the same: systems may confuse coordinated prevalence with independent evidence.

The resulting transformation follows a recognizable sequence:

Institutional power produces content.

Content becomes abundant.

Abundance becomes statistical weight.

Statistical weight becomes model disposition.

Model disposition becomes fluent explanation.

Fluent explanation appears as independent judgment.

The original speaker disappears.

This disappearance is the central political danger.

A government statement can be challenged as a government statement. A contractor's website can be investigated as sponsored communication. A lobbying campaign can be exposed as advocacy. But when the same framing reappears as the smooth conclusion of an artificial intelligence system, the user may no longer know where political power entered the chain.

The defense is not censorship. It is preserved provenance.

Models must distinguish one source repeated many times from many independent sources. They must disclose material sponsorship, identify parties to disputes, preserve uncertainty, and resist converting source concentration into certainty. Providers must audit training and retrieval systems for coordinated influence. Researchers must conduct controlled, multilingual, longitudinal experiments. Governments must accept reciprocal transparency for their own operations.

Above all, artificial intelligence must retain contestability. The model should not close political questions merely because it can express an answer fluently. It should help users see the evidence, the disagreements, and the institutions behind the words.

Large language models are becoming compressed public spheres. Whether they strengthen or weaken democratic reason will depend upon what survives that compression.

If sources disappear, power becomes invisible.

If sponsorship disappears, advocacy becomes knowledge.

If coordination disappears, repetition becomes consensus.

If uncertainty disappears, interpretation becomes fact.

The essential principle is therefore simple:

The machine must not forget who spoke.

References

1. Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M. E., Stewart, B. M. & Tucker, J. A. State media control influences large language models. *Nature* **655**, 685–693 (2026). https://doi.org/10.1038/s41586-026-10506-7

2. U.S. Department of Justice, National Security Division, Foreign Agents Registration Act Unit. *Exhibit A and Exhibit B to the Registration Statement of Clock Tower X LLC, Registration No. 7649*. Filed 18 September 2025. https://efile.fara.gov/docs/7649-Exhibit-AB-20250918-1.pdf

3. Isenstadt, A. Trump alum helps Israel mount AI influence campaign. *Axios* (25 April 2026). https://www.axios.com/2026/04/25/israel-ai-influence-parscale

4. Cleveland-Stout, N. Israel is paying millions to train AI chatbots how to talk about Gaza. It's working. *Drop Site News* (28 July 2026). https://www.dropsitenews.com/p/israel-brad-parscale-ai-chatbots-gaza

5. OpenAI. Disrupting deceptive uses of AI by covert influence operations. *OpenAI* (30 May 2024). https://openai.com/index/disrupting-deceptive-uses-of-ai-by-covert-influence-operations/

6. Paul, K. Meta identifies networks pushing deceptive content likely generated by AI. *Reuters* (29 May 2024). https://www.reuters.com/technology/meta-identifies-networks-pushing-deceptive-content-likely-generated-by-ai-2024-05-29/