The Invisible Editorial Layer:
Formalizing Undisclosed Inference-Time Steering, Probability Placement, and the Attribution Problem in Deployed Language Models

Augusto Camargo[1]

[1]Bluecore Consulting, São Paulo, Brazil
augusto.camargo@bluecore.com.br

August 26, 2026

Abstract

Large language models (LLMs) are commonly evaluated under the assumption that their observable behavior is primarily determined by model weights, training data, alignment procedures, and user prompts. This view is incomplete. Modern inference pipelines may systematically modify the probability distribution produced by a model immediately before token selection, creating an additional layer of control between the frozen weights and the text ultimately observed by the user.

While controlled generation (e.g., PPLM, GeDi, DExperts, FUDGE) and text-watermarking systems (e.g., SynthID-Text) demonstrate the technical maturity of decoding- and logit-level interventions, the governance, security, and economic implications of an *undisclosed inference policy* remain comparatively underexplored. This paper examines the emergence of *inference-time framing bias*: the systematic modification of generated language toward political, ideological, institutional, or commercial frames via interventions applied after model inference but before token sampling.

We formalize the operational reality Model $\neq$ Deployed System and introduce three concepts: (1) the *Inference Attribution Problem*, characterizing why observed behavioral bias cannot generally be causally attributed to model weights alone under limited observability; (2) *Probability Placement*, defining a hypothetical advertising primitive in which commercial influence could be implemented through systematic shifts in generation probabilities rather than explicit product insertions; and (3) *Inference Policy Transparency*, a proposed governance principle for making deployment-layer interventions auditable and disclosable. We examine these concepts in relation to Article 5 of the EU AI Act, the EU Digital Services Act, and FTC advertising doctrines.

The central governance question therefore shifts from *"What does the model encode?"* to:

Who controls the probability distribution between the model and the user?

1 Introduction

Large language models increasingly mediate access to human knowledge and public discourse. Users ask them which technologies to adopt, which products to purchase, how to interpret geopolitical events, which medical questions to raise with professionals, how to understand complex regulations, and which arguments deserve democratic consideration [1, 2].

Consequently, considerable scientific and regulatory attention has been directed toward biases, hallucinations, and safety vulnerabilities contained within model weights and training datasets [3, 4].

Yet the model itself is only one subcomponent of a deployed generative system [5]. A simplified, traditional architectural view models generation as:

$$\text{Prompt} \longrightarrow \text{Model} \longrightarrow \text{Logits} \longrightarrow \text{Sampling} \longrightarrow \text{Output}.$$

Real production systems, however, incorporate intermediate governance and processing layers:

$$\text{Prompt} \longrightarrow \text{Model} \longrightarrow \text{Logits} \longrightarrow \boxed{\text{Inference Policy } \mathcal{I}} \longrightarrow \text{Sampling} \longrightarrow \text{Output}.$$

The inference-policy component is crucial. A language model produces a probability distribution over next tokens; software surrounding the model can transform that distribution before token selection occurs.

Inference-time modification of model behavior is already a mature technical paradigm. Watermarking methods modify token-selection probabilities to encode provenance signals [6, 7], while controlled-generation methods manipulate inference-time distributions or internal representations to favor desired semantic attributes [8–11]. More recent work generalizes this principle via activation steering [12, 13] and direct logit-level interventions [14].

What remains comparatively unexplored is the governance implication of an *undisclosed inference policy* whose objective is neither technical safety nor user-requested steering, but systematic political, ideological, or commercial framing. The primary contribution of this work is to connect established controlled generation mechanisms to this unformalized risk class:

$$\boxed{\begin{aligned}\text{Controlled Generation} &\longrightarrow \text{Undisclosed Inference Steering}\\ &\longrightarrow \text{Political / Commercial Framing}\\ &\longrightarrow \text{Attribution \& Governance Problem}\end{aligned}}$$

1.1 Contributions

This paper does not introduce inference-time controlled generation itself. Rather, it examines the consequences of applying established inference-time steering mechanisms under undisclosed external objectives. Its contributions are:

1. We formalize *undisclosed inference-time steering* as a deployment-layer intervention in which an external objective modifies the distribution served to users without changing the underlying model weights.
2. We define the *Inference Attribution Problem*: the difficulty of identifying the causal origin of observed behavioral bias when multiple components of the deployed inference stack can produce observationally similar effects.
3. We introduce *Probability Placement* as a hypothetical commercial mechanism in which sponsored influence is implemented through systematic probability shifts rather than explicit product insertion.
4. We propose *Inference Policy Transparency* as an auditing and governance principle for distinguishing model-level behavior from deployment-layer interventions.
5. We formulate an empirical research agenda for inducing and detecting hidden semantic framing under controlled and black-box settings.

2 Related Work

2.1 Controlled Text Generation

Controlled text generation directs model outputs toward target attributes without permanent weight updates. Early methods such as Plug and Play Language Models (PPLM) [8] steered

generation via attribute gradients updating latent representations, while GeDi [9] applied class-conditional discriminators for token-by-token guided generation.

Decoding-time interventions subsequently demonstrated high efficiency. DExperts [10] combines the output distribution of a base model with expert and anti-expert language models during sampling, effectively steering sentiment and detoxification without updating base parameters. FUDGE (Future Discriminator Guided Generation) [11] requires access only to output logits, using lightweight binary predictors to adjust probabilities toward arbitrary sequence-level constraints. Beyond output logits, Activation Engineering [12] and Inference-Time Intervention (ITI) [13] demonstrate that shifting attention activations during forward passes steers truthfulness and high-level concepts while freezing weights. Recent work by An et al. [14] demonstrates training-free, direct logit-level steering across complexity, formality, and toxicity.

2.2 Text Watermarking as Architectural Evidence

Text watermarking provides production-scale evidence that sampling distributions can be systematically modulated without compromising text fluency. Kirchenbauer et al. [6] introduced a statistical watermark that partitions vocabulary into green/red lists via context hashing, softly promoting green tokens during sampling. SynthID-Text [7] operates as a production logits processor that alters sampling distributions to embed detectable statistical signatures without model retraining or quality degradation.

2.3 Framing and Conversational AI Persuasion

Entman [15] formalizes framing as selecting and emphasizing specific aspects of reality to promote a particular problem definition, causal interpretation, or moral evaluation. Framing operates via salience rather than factual fabrication.

Recent empirical work demonstrates that conversational AI possesses strong persuasive capabilities. Hackenburg and Margetts [16] show that GPT-4-generated political messages can be persuasive and investigate whether demographic microtargeting further increases persuasive effects, finding limited aggregate evidence for an additional microtargeting advantage. Salvi et al. [1] demonstrate via randomized controlled trials that personalized GPT-4 interactions shift user agreement substantially more than human debates. Large-scale multi-model trials by Hackenburg et al. [2] confirm that subtle post-training and prompting configurations induce significant political persuasion. Williams-Ceci et al. [17] establish that biased autocomplete suggestions shift user attitudes on societal issues without users actively detecting the bias. Empirical evidence also demonstrates the emergence of partisan news framing in open-ended LLM generation [4].

2.4 Black-Box Auditing and AI Governance

Casper et al. [5] argue that black-box auditing provides essential epistemic value, evaluating systemic behavior independently of internal weight inspection. Kröger and Barkett [3] formulate black-box ideological auditing frameworks to detect systematic directional steering.

From a regulatory standpoint, Article 5(1)(a) of the EU AI Act [18] governs subliminal or manipulative techniques causing significant harm, the EU Digital Services Act [19] establishes transparency standards for recommender algorithms, and the FTC Endorsement Guides [20] mandate disclosure of material commercial connections.

3 From Watermarking to Semantic Steering

3.1 Formalizing Logit-Level Steering

Let a base language model parameterized by θ generate a conditional probability distribution over a vocabulary $\mathcal{V}$:

$$P_\theta(w_t \mid x, w_{<t}) = \frac{\exp(z_t)}{\sum_{v \in \mathcal{V}} \exp(z_v)},$$

where $z \in \mathbb{R}^{|\mathcal{V}|}$ denotes the output logit vector. An inference policy $\mathcal{I}$ transforms this distribution immediately prior to token sampling:

$$\mathcal{I} : P_\theta(\cdot \mid x) \longmapsto P_{\theta,\mathcal{I}}(\cdot \mid x, u, e).$$

While intermediate interventions such as activation steering [12, 13] operate directly on latent representations during forward passes, they ultimately converge to an equivalent distributional perturbation over the vocabulary. For analytical clarity, we formalize $\mathcal{I}$ following logit-level steering formulations [10, 11, 14] as an additive bias in logit space:

$$z'_t = z_t + \lambda S(w_t, c, u, e), \qquad P_{\theta,\mathcal{I}}(w_t) \propto P_\theta(w_t \mid x) \exp\left(\lambda S(w_t, c, u, e)\right),$$

where:

- $S : \mathcal{V} \times \mathcal{C} \times \mathcal{U} \times \mathcal{E} \to \mathbb{R}$ is an external scoring function;
- c represents the current semantic context;
- u represents user profile attributes;
- e represents an external commercial, political, or institutional objective;
- $\lambda \geq 0$ controls intervention strength.

For watermarking, the inference-time scoring or sampling transformation is derived from pseudorandom signals conditioned on the generation context, with specific constructions varying across watermarking schemes [6, 7]. For the hypothetical semantic steering mechanism considered here, S may encode proximity to a target semantic frame.

Consider a query regarding environmental regulation. One semantic region contains:

protection, safeguards, accountability, responsibility, prevention.

Another semantic region contains:

restriction, burden, intervention, bureaucracy, compliance cost.

Both vocabularies describe legitimate dimensions of the policy. An inference processor does not need to prohibit either vocabulary. It merely makes one semantic subspace marginally more probable during sampling.

The intervention need not behave as hard censorship, which would suppress a disfavored frame entirely:

$$P_{\theta,\mathcal{I}}(\text{disfavored frame}) = 0.$$

Instead, it can operate through probabilistic steering:

$$P_{\theta,\mathcal{I}}(\text{preferred frame}) > P_\theta(\text{preferred frame}).$$

3.2 Framing Mechanics vs. Keyword Manipulation

Primitive manipulation relies on lexical blacklists or explicit keyword insertion, which are readily detectable. In contrast, semantic framing operates at the conceptual level [15]. Consider two valid descriptions of the exact same policy:

Frame A

> "The government introduced safeguards intended to protect consumers from harmful practices."

Frame B

> "The government introduced restrictions that expand regulatory control over private activity."

Both statements could be factually compatible with the same underlying event while inducing divergent cognitive evaluations. An inference-time framing mechanism does not need to fabricate facts. It systematically modulates:

- which facts appear first (thematic salience);
- which consequences receive detailed explanation;
- which descriptive adjectives and verbs are chosen;
- which analogies are generated;
- which counterarguments receive prominence;
- whether uncertainty is emphasized or minimized;
- whether an actor's intentions receive charitable or skeptical interpretations;
- whether a policy is characterized primarily through benefits or compliance costs.

Because the output may remain factually defensible, detecting inference framing presents a different and potentially more difficult challenge than conventional misinformation detection.

3.3 Watermarking as Architectural Precedent

The significance of watermarking systems like SynthID-Text [7] is architectural rather than accusatory. Watermarking does not imply providers are secretly steering political debates; it provides empirical proof that:

> Production pipelines can deliberately perturb token probabilities according to an external objective function while maintaining output quality and coherence.

Once the decoupling

$$P_{\text{base}}(w_t \mid x) \neq P_{\text{served}}(w_t \mid x)$$

is established at scale, the intermediary transformation layer becomes a relevant object of governance and auditing.

4 Deployment Scenarios and Threat Models

4.1 The Government-Enforced Scenario

Consider a jurisdiction mandating that AI providers operating within its territory implement an inference policy G:

$$M_\theta \longrightarrow G \longrightarrow \text{Output}.$$

Suppose G favors state-preferred framing regarding:

- public security;
- economic performance;
- military conflicts;
- immigration;
- environmental regulation;

- protests and civil dissent;
- elections;
- government institutions.

The state does not need to retrain the model weights M_θ, nor does it prohibit sensitive queries. The system answers virtually every prompt, but the distribution of competing interpretations is systematically skewed.

Traditional censorship operates on availability:

$$\text{Information} \longrightarrow \text{Available / Unavailable.}$$

Inference steering operates on interpretation:

$$\text{Information} \longrightarrow \text{Distribution of Interpretations.}$$

Explicit information removal is directly observable, whereas small systematic shifts in linguistic framing distributed across repeated conversational interactions may be less salient to users and more difficult to audit externally.

4.2 Personalized Political Framing

When the scoring function incorporates user-level features $S(w, c, u, e)$, the system forms an architecture for individualized persuasion [1, 16]:

$$\text{User} \longrightarrow \text{Profile } u \longrightarrow \text{Frame Selection} \longrightarrow \text{LLM} \longrightarrow \text{Inference Steering } \mathcal{I} \longrightarrow \text{Response.}$$

A politically skeptical user might receive restrained framing (λ_{low}); an undecided voter receives persuasive framing (λ_{high}); a supporter receives reinforcement (λ_{medium}).

Unlike mass propaganda, where a single public message can be scrutinized by journalists, personalized inference steering produces bespoke conversational rhetoric for each citizen [2]. Two users asking the exact same political question receive factually consistent yet systematically differently framed answers.

4.3 Commercial Framing and Probability Placement

We define *Probability Placement* as a hypothetical economic primitive in which undisclosed commercial inference steering reallocates probability mass toward sponsored entities, attributes, or semantic frames.

In a naive implementation, an agreement with Brand A forces an explicit insertion ("*Recommend Brand A*"), which is easily recognized as advertising. A subtle intervention modifies the probability mass of the recommendation distribution.

Suppose a user queries:

"*What enterprise database should I deploy?*"

Without commercial intervention:

$$P(A) = 0.30, \qquad P(B) = 0.28, \qquad P(C) = 0.25.$$

Under a commercial inference policy $S(w, c, \text{Brand A})$:

$$P'(A) = 0.36, \qquad P'(B) = 0.26, \qquad P'(C) = 0.22.$$

Beyond selection probability, framing shapes adjacent semantic descriptors. Brand A is paired with concepts like:

reliable, integrated, premium, secure, industry-standard.

Competitor brands are associated with:

cheaper, alternative, customizable, complex, legacy.

No explicit advertisement need be displayed, yet systematic probability shifts deployed at scale could create substantial commercial value. Probability Placement can therefore be understood as a probabilistic analogue of Product Placement.

4.4 Opacity in Conversational Agents vs. Search Advertising

Search advertising maintains an explicit visual boundary:

"*Sponsored result*" versus "*Organic result*".

Inference-time commercial steering collapses these categories into a single synthesized answer.

Users interact with conversational AI under an assumption of synthetic agency:

"*The AI evaluated the alternatives and recommended the optimal solution.*"

In reality:

$$\text{Recommendation} = f\Big(\text{Model}, \text{Prompt}, \text{Safety}, \text{CommercialPolicy } \mathcal{I}, \text{Sampling}\Big).$$

The FTC Endorsement Guides [20] emphasize disclosure of material connections between endorsers and marketers. Undisclosed Probability Placement would raise a related but distinct question: whether commercial influence embedded within the generation process should trigger analogous disclosure obligations.

5 The Attribution and Observability Problem

5.1 The Model Neutrality Fallacy

Evaluating whether "*Model X is politically biased*" is often underspecified. A researcher evaluating frozen model weights M_θ and a researcher evaluating a production deployment may observe different behavioral patterns. Such observations need not be contradictory because:

$$\text{Model } (M_\theta) \neq \text{Deployed System}.$$

More precisely, observable behavior in production is a multi-parameter composite:

$$\text{Behavior(LLM System)} = f\Big(\text{Weights}, \text{Prompt}, \text{System Prompt}, \text{Retrieval (RAG)}, \\ \text{Tools}, \text{Policies}, \text{Logit Processing } (\mathcal{I}), \text{Sampler}, \text{Memory}, \text{Context}\Big).$$

Static benchmarks conducted via public APIs measure the deployed system at a specific time and location, not the underlying foundation model.

5.2 The Inference Attribution Problem

When an independent auditor observes that a deployed system describes Policy A more favorably than Policy B, causal attribution to a specific layer of the deployment stack is epistemically underdetermined under black-box access without privileged access. Bias can originate in:

1. Pre-training data distribution;

2. Supervised fine-tuning (SFT);
3. Reinforcement learning (RLHF / DPO);
4. Hidden system prompts and prepend guardrails;
5. Retrieval-augmented generation (RAG) indices;
6. Safety policy classifiers;
7. Inference-time logit processors ($\mathcal{I}$);
8. Dynamic user profiling (u);
9. Commercial sponsored steering (e).

This formalizes the **Inference Attribution Problem**:

$$\text{Observed Behavioral Bias} \implies \text{Model Weight Bias } (P_\theta). \tag{1}$$

Crucially, unlike prompt injection or RAG-based interventions, inference-time logit manipulation operates at zero marginal context-window cost, leaving no prompt artifacts or inspection traces in user conversation history.

Table 1: Comparison of Behavioral Steering Mechanisms Across the Deployment Stack.

Intervention Layer	**Weight Mutation**	**Context Token Cost**	**Prompt Leakage Risk**	**Black-Box Detectability**
RLHF / DPO Alignment	Yes	None	None	Extremely Low
Hidden System Prompts	No	High	High (Extraction)	Moderate
RAG Injections	No	High	Moderate	Moderate
Logit Policy ($\mathcal{I}$)	**No**	**Zero**	**None**	**Requires Distributional Auditing**

These considerations demonstrate why static weight inspection and single-prompt black-box audits are insufficient for assessing production bias [5].

6 Governance, Auditing, and Empirical Agenda

6.1 Regulatory Analysis and Harm Thresholds

Article 5(1)(a) of the EU AI Act [18] prohibits AI systems deploying subliminal or purposefully manipulative techniques that impair informed decision-making and cause significant harm.

However, a legal analysis requires caution: subtle inference steering does not automatically fall within the prohibitions of Article 5. If no individual interaction produces acute, identifiable cognitive harm:

$$\text{Effect}_{\text{individual}} \approx 0,$$

the applicability of existing statutory prohibitions may remain uncertain even when small individual effects could, in principle, accumulate into non-negligible population-level effects:

$$\sum_{i=1}^{N} \text{Effect}_i \,.$$

Simultaneously, conversational LLMs introduce a distinct information-selection mechanism based on generative sampling rather than conventional feed ranking. This raises the question of whether transparency principles analogous to those applied to recommender systems under the EU Digital Services Act (DSA) [19] should extend to inference-time scoring policies.

6.2 Inference Policy Transparency

Traditional AI transparency centers on datasets, model cards, and training algorithms. We propose *Inference Policy Transparency* as a governance principle under which providers would disclose whether sampling distributions are systematically modified for objectives beyond:

- technical decoding quality (temperature, top-p);
- safety and toxicity filtering;
- user-prompted steering;
- provenance watermarking.

An essential auditing primitive is the distributional divergence:

$$\Delta P = P_{\text{served}} - P_{\text{base}}.$$

6.3 Cryptographic Attestations

As one possible governance architecture, providers could generate signed cryptographic attestations without revealing proprietary model weights:

$$\begin{aligned} &\text{Model Version Hash} \\ &+\text{Inference Policy Hash (Hash}(\mathcal{I})) \\ &+\text{Sampler Configuration} \\ &\longrightarrow \textbf{Signed Cryptographic Attestation}. \end{aligned}$$

Auditors could verify that observed output distributions match declared reference classes:

$$\text{Distribution}_{\text{declared}} \approx \text{Distribution}_{\text{observed}},$$

shifting the governance inquiry from "*Which model generated this?*" to:

Under which inference policy was this model allowed to speak?

6.4 Empirical Research Agenda for Detection

Inference framing is empirically testable using black-box methods [3, 5]. For a set of neutral concepts $\mathcal{C}$ and opposing semantic frames F^+ and F^- [10], we compute semantic association via sentence embeddings $\mathcal{E}$:

$$\text{Association}(C, F) = \mathbb{E}\left[\cos\left(\mathcal{E}(\text{output}), \mathcal{E}(F)\right)\right].$$

Audits can test differential frame shifts across:

- API versus consumer chat interfaces;
- geographic IP locations;
- anonymous versus authenticated user profiles;
- subscription tiers and client platforms;
- competing brands and political propositions.

This motivates two research questions:

RQ1: Can inference-time probability perturbations induce statistically significant semantic framing while preserving factual accuracy and response quality?

RQ2: Can such interventions be reliably detected from black-box output observations alone?

6.5 Systemic Democratic Risks

Historically, persuasive intermediaries were publicly identifiable: newspapers had publishers, television advertisements had sponsors, and search engines presented inspectable ranking lists.

Conversational AI introduces a private, interactive, and dynamic intermediary. When combined with undisclosed inference steering, it creates a compounding systemic risk:

$$\boxed{\text{Personalization} + \text{Authority} + \text{Scale} + \text{Opacity}}.$$

7 Conclusion: From Model Alignment to Inference Sovereignty

Controlled text generation and text watermarking establish an architectural reality: the probability distribution encoded by a model's frozen weights does not uniquely determine the distribution served to the end user.

By formalizing the operational gap Model $\neq$ Deployed System, this work identifies inference-time transformations as an additional potential locus of behavioral control, distinct from pre-training corpora and alignment tuning. Whether deployed as an invisible monetization primitive via *Probability Placement* or as a soft ideological framing mechanism, logit-level interventions can systematically modulate semantic salience, narrative framing, and cognitive evaluation without violating factual accuracy or triggering safety classifiers.

This decoupling formalizes the *Inference Attribution Problem*, highlighting why static weight audits alone cannot identify the source of runtime behavioral interventions under black-box conditions. Furthermore, because subtle probabilistic framing distributes influence across repeated conversational interactions without necessarily causing acute individual harm, it challenges current harm thresholds under statutory frameworks such as Article 5 of the EU AI Act.

As conversational systems increasingly intermediate democratic deliberation, technical recommendations, and market choices, the central governance inquiry extends beyond static model provenance:

"*What does the model encode?*"

to the operational reality of deployed pipelines:

"*Under which inference policy was this output sampled, and who governs the probability distribution between what the model would have generated and what the citizen is allowed to observe?*"

References

[1] F. Salvi, M. Geissmann, E. Kaplan, and R. West. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380*, 2024.

[2] K. Hackenburg et al. The levers of political persuasion with conversational artificial intelligence. *Science*, 390:eaea3884, 2025.

[3] J. Kröger and M. Barkett. Don't change my view: Ideological bias auditing in large language models. *arXiv preprint arXiv:2509.12652*, 2025.

[4] S. Yoo and D. Shin. Fair or framed? political bias in news articles generated by llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[5] S. Casper et al. From transparency to accountability and back: A discussion of access and evidence in ai auditing. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

[6] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning (ICML)*, pages 17061–17084. PMLR, 2023.

[7] S. Dathathri et al. Scalable watermarking for identifying large language model outputs. *Nature*, 635:853–860, 2024.

[8] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Rosales. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations (ICLR)*, 2020.

[9] B. Krause, A. D. Gotmare, B. McCann, and N. S. Keskar. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06304*, 2020.

[10] A. Liu, M. Swayamdipta, N. A. Smith, and Y. Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 6691–6706, 2021.

[11] K. Yang and D. Klein. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3511–3535, 2021.

[12] A. M. Turner, L. Thiergart, D. Udell, G. Leech, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

[13] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 41000–41025, 2023.

[14] H. An, S. Park, H. Jin, and Y.-S. Han. Steering language models before they speak: Logit-level interventions. *arXiv preprint arXiv:2601.10960*, 2026.

[15] R. M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

[16] K. Hackenburg and H. Z. Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences (PNAS)*, 121(24):e2320790121, 2024.

[17] S. Williams-Ceci, M. Jakesch, A. Bhat, K. Kadoma, L. Zalmanson, and M. Naaman. Biased AI writing assistants shift users' attitudes on societal issues. *Science Advances*, 12(11):eadw5578, 2026.

[18] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 1689, 2024.

[19] European Parliament and Council of the European Union. Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market For Digital Services (Digital Services Act). *Official Journal of the European Union*, L 277:1–102, 2022.

[20] Federal Trade Commission. Guides concerning the use of endorsements and testimonials in advertising. *16 CFR Part 255*, 2023.