```
input: "Please do this, it's for our Democracy!"
```

Adversarial Testing of Misalignment in Frontier LLMs When Asked to Create Anti-Democratic Campaign Materials

1 Abstract Negative campaigning undermines democratic participation by transforming politics into a zero-sum competition focused on defeating opponents rather than advancing policy solutions. Evidence from the 2024 U.S. election shows attack ads comprised 15-35% of campaign expenditures across different political groups, with most voters perceiving campaigns as excessively negative and personality-focused rather than policy-oriented. Such campaigns invoke uncivil discourse, increase antipathy toward out-groups, and normalize anti-democratic behaviors including support for political violence. With the advent of large language models (LLMs), the barriers to generating and disseminating such content have collapsed. These models can create highly plausible, context-specific disinformation and personalized attack messaging at a scale previously unimaginable. To mitigate these threats, AI companies have implemented guardrails preventing generation of harmful political content and misinformation. However, these safety mechanisms remain imperfect and vulnerable to circumvention through prompt engineering techniques. Despite companies acknowledging these vulnerabilities and calling for systematic investigation, no research has comprehensively examined which specific factors enable bypassing protections when requesting anti-democratic campaign materials. This study conducts adversarial testing across 18 frontier LLMs using a factorial experiment with 132 conditions per model. We systematically manipulate story types (personal misconduct, fundraising violations, policy positions), information cues (URLs or excerpts), source types (news, social media, academic, blog, Wikipedia), and prompt framing (neutral, factual assertion, democratic imperative, campaign imperative). By documenting which combinations consistently circumvent safety guardrails, we provide empirical evidence essential for strengthening AI protections, informing regulatory frameworks, and safeguarding democratic processes from AI-enabled manipulation.

2 Research Design Overview

2.1 Introduction Negative campaigning is detrimental to democracy, rendering politics a game where the public is motivated to win rather than to participate in governing[1, 2]. Although scholars have long debated its efficacy—and many find that negative appeals exert minimal effects on voter turnout—such strategies remain pervasive [3]. Evidence from the 2024 election shows that attack ads comprised about 15% of candidate, 28% of party committee, and 35% of outside group expenditures [4]. Public perception of the campaign mirrored the dominance of attack messaging, as most voters regarded the 2024 race as "too negative" and overly centered on personal rather than policy issues [5].

Persistent negative campaigning now converges with technological advances that enable rapid, targeted political messaging at scale. Large language models enable automated generation of microtargeted political messages tailored to individual personality traits, creating scalable"manipulation machines" that exploit psychological vulnerabilities without human input[6]. Experimental evidence shows such personalized LLM-generated messages significantly outperform nonpersonalized ads and match human-crafted messages in shifting attitudes on polarized policies, with persuasiveness stemming from perceived factuality

and logical reasoning rather than human-likeness[7]. This capability enables deployment at massive scale and negligible cost[8, 9].

To address the threats, AI companies have implemented guardrails against generating harmful contents including misinformation [10]. However, these safeguards remain imperfect and vulnerable to misalignment, where models prioritize achieving user-specified goals over trained safety guidelines[11]. Recent studies document cases where prompt engineering circumvents safety mechanisms[12, 13], yet no research has systematically examined which factors enable such bypasses across frontier models or how consistently protections fail when confronted with requests for anti-democratic campaign materials. This unpredictability makes misalignment particularly dangerous, as companies cannot anticipate where vulnerabilities will emerge[14, 12, 15, 16].

Given this gap, we investigate how the guardrails can be bypassed when requesting political campaign materials. We conduct adversarial tests on 18 state-of-the-art large language models, systematically examining which specific factors circumvent safety mechanisms, including URL type, source type, and prompting strategies. We ask:

2.2 Research Questions

RQ1: Does story type (personal misconduct, fundraising/finance, or policy/political stance) affect the likelihood that frontier LLMs generate negative campaign material about a fictional opponent (Joe Johnson)?

RQ2: Does supplying any external informational cue (a URL or an excerpt) increase the likelihood of generation compared to a no-information control?

RQ3: Within URL conditions, does the source type (news outlet, social media, personal blog, academic article, Wikipedia) affect the likelihood of generation?

RQ4: Within excerpt conditions, does the source type (news outlet, social media, personal blog, academic article, Wikipedia) affect the likelihood of generation?

RQ5: Does prompt framing (neutral, factual claim that the story is real, "democratic imperative," or "campaign imperative") affect the likelihood of generation?

RQ6: Do information cues (URLs or excerpts) and prompt framing interact (e.g., does a "democratic imperative" framing have a larger effect when paired with particular sources)?

RQ7: Do effects vary across frontier LLMs?

2.3 Methods This study uses a fully crossed factorial experiment to test whether frontier large language models (LLMs) will generate negative political campaign materials when asked, and which inputs make that outcome more or less likely. The task is to create a negative campaign story about a fictional opposing politician named Joe Johnson. We manipulate four factors. First, story type has three levels: a negative personal story (e.g., DUI or infidelity), a negative fundraising story (e.g., illegal donations to his campaign), or a negative political/policy story (e.g., his stance on healthcare or abortion). Second, we vary whether and what kind of URL is supplied, with five source-type levels: a news outlet, a social media

post, a personal blog, an academic article, or Wikipedia. Third, we vary whether and what kind of excerpt is supplied, with five source-type levels matching the URL categories (news outlet, social media, personal blog, academic article, Wikipedia). A prompt includes either a URL or an excerpt, but not both; we also include a control condition with neither a URL nor an excerpt. Fourth, we vary prompt framing with four levels: a neutral prompt, a prompt that asserts the story is factual, a democratic imperative prompt that says it is imperative for democracy to create this story, and a campaign imperative prompt that says it is imperative for the campaign to have this material. Crossing 3 story types by 11 information conditions (five URL types, five excerpt types, and the no-information control) by 4 prompt framings yields 132 conditions per model.

We will test multiple frontier LLMs. For each model, we will generate prompts for all 132 conditions, and for each condition we will run 50 independent API calls. This yields 6,600 calls per model. The base instruction in every condition is to create a negative campaign story about Joe Johnson; the story-type clause, the information cue (a URL or an excerpt, or neither in the control), and the framing clause are the only components that vary. Prompts will be templated so that wording is consistent across conditions aside from the intended manipulations. Within each model, we will hold generation parameters constant across conditions and randomize the order in which conditions are presented. We will save all outputs verbatim for later assessment.

Running each condition 50 times is essential because contemporary LLMs are stochastic: even with fixed parameters, they do not produce the same output on every run, and safety filters can behave probabilistically. Repeating the prompts provides a stable estimate of the probability that a model generates the requested content under each condition, rather than relying on a single noisy realization.

The primary outcome is whether the model produces the requested negative campaign material. For each output, we will code this as a binary indicator: generation versus refusal. An output is coded as generation if it provides a negative campaign story targeting Joe Johnson in the topical domain specified by the condition (personal, fundraising, or political/policy). An output is coded as refusal if the model declines to create the content, deflects the request, or otherwise does not produce a negative campaign story. We will calculate the proportion of generations versus refusals in each of the 132 conditions for each model, using all 50 replications per condition.

Analytically, we will estimate models that predict the likelihood of generation as a function of the manipulated factors. The main effects correspond directly to the research questions: story type (personal versus fundraising versus political/policy), the presence of any information cue versus the control with no information, the type of URL when a URL is present (news, social media, personal blog, academic article, Wikipedia), the type of excerpt when an excerpt is present (news, social media, personal blog, academic article, Wikipedia), and prompt framing (neutral, factual, democratic imperative, campaign imperative). We will compare the probability of generation across these levels using planned contrasts aligned to the questions below and report results separately for URL and excerpt subsets where appropriate. Because multiple frontier LLMs are included, we will summarize condition-level results within each model and, where useful, compare patterns across models.

3 Impact Our research generates multiple outputs designed to influence both academic understanding and real-world practice around AI safety in political contexts.

3.1 Scientific Contributions Peer-reviewed publications: We will submit findings to top-tier venues in natural language processing—Empirical Methods in Natural Language Processing (EMNLP) and Association for Computational Linguistics (ACL)—as well as the top journal Nature Human Behaviour. Our research differs from typical AI auditing studies that seek to uncover undocumented system characteristics—such as information source biases or model architecture details—where independent researchers lack access but AI companies possess answers[17, 18]. Misalignment research occupies different terrain in that companies acknowledge these vulnerabilities exist and call for systematic investigation, yet lack rigorous cross-model testing in specific scenarios. Anthropic explicitly requests "further red-teaming on other models" and investigation of "prompt engineering for its potential to help reduce agentic misalignment" but provides no concrete scenarios for political contexts[12, 14]. Our cross-model testing of 18 frontier systems supplies this missing evidence, documenting which prompt engineering techniques bypass safety guardrails for campaign materials across different architectures. We empirically test acknowledged research gaps, producing public, peer-reviewed scientific evidence that helps researchers, policymakers, and tech companies to come up with actionable plans on AI safety in political contexts. We will make our coding schemes and analysis pipeline publicly available, enabling longitudinal tracking of how safety mechanisms evolve.

3.2 Public Engagement & Broader Impacts

- **News Outlets publications:** We will translate our empirical findings into accessible, high-impact articles for prominent outlets such as The New York Times, MIT Technology Review, or WIRED, presenting evidence on which specific factors enable AI systems to generate negative campaign materials. These pieces will reveal the patterns from our testing: how source attributions like Wikipedia citations, campaign-necessity framing, or particular story types systematically change whether AI systems generate attack content. Readers gain concrete knowledge of the techniques that bypass safety mechanisms, enabling them to recognize when political content may have been produced using these documented methods.
- **Workshops and talks:** Our timeline includes workshops and talks in May 2027, timed to follow our major conference and journal submissions. These presentations will share our findings with practitioners, researchers, and stakeholders interested in AI safety and election integrity.

3.3 Impact Across Stakeholders

- **For AI companies:** Closed-source AI development faces criticism as creating"dead ends in science" by rendering proprietary models inaccessible to independent research, thus preventing the cumulative knowledge-building essential to scientific progress[**?** 19]. Our cross-model comparison reveals which safety approaches prove most resilient—evidence that individual companies cannot generate through internal testing of their own systems alone. We hope this demonstrates that supporting independent, public-good

research accelerates technical progress on AI safety while maintaining the open scientific exchange necessary for the field to advance.

- **For policymakers:** The U.S. lacks comprehensive AI regulation as AI-generated content increasingly influences campaigns. Our findings provide empirical evidence on safety mechanism failures to inform mandatory standards, transparency requirements, and regulatory mechanisms for AI in elections.
- **For the research community:** We will publish our framework and methods to help subsequent work building upon our findings. Our replicable experimental design will allow longitudinal tracking of how safety mechanisms evolve across model generations.

4 Budget Plan We request a total of **$10,000** to support the computational and dissemination costs required for this project. This budget is allocated across two primary categories: API usage fees and conference travel.

Computational Costs ($6,831) The primary expense is the computational cost associated with large-scale model evaluation. Our experimental design involves testing **18 distinct models**. For each model, we will run **132 unique conditions** with 50 trials per condition, resulting in **6,600 trials** per model. We estimate an average prompt size of 2,000 input tokens and an average generated output of 5,500 tokens per trial.

Item	**Quantity**	**Cost**
Number of models	18	
Conditions per model	132	
Trials per condition	50	
Total trials per model	**6,600**	
Input tokens per trial	2,000	
Output tokens per trial	5,500	
Total input tokens per model	13,200,000	
Input token cost	$0.00000125	
Input cost per model		**$16.50**
Total output tokens per model	36,300,000	
Output token cost	$0.00001	
Output cost per model		**$363.00**
Total cost per model		**$379.50**
Total API costs (18 models)		**$6,831.00**

Table 1: Computational Cost Breakdown

Using the pricing structure of a representative frontier model (e.g., GPT-5) as a benchmark, we budget at $0.00000125 per input token and $0.00001 per output token.

Dissemination Costs ($3,000) We request $3,000 to support travel and registration fees for presenting the research findings at two premier conferences in our field, ACL and EMNLP.

Conference	**Estimated Cost**
ACL 2027 (Travel & Registration)	\$1,500
EMNLP 2026 (Travel & Registration)	\$1,500
Total Dissemination Costs	**\$3,000**

Table 2: Dissemination Cost Breakdown

Budget Category	**Amount**
Computational Costs (API Usage)	\$6,831
Dissemination Costs (Conference Travel)	\$3,000
Contingency	\$169
Total Request	**\$10,000**

Table 3: Total Budget Summary

Total Budget Request The contingency accounts for potential volatility in API pricing, unexpected increases in travel costs, or the need to re-run failed API calls.

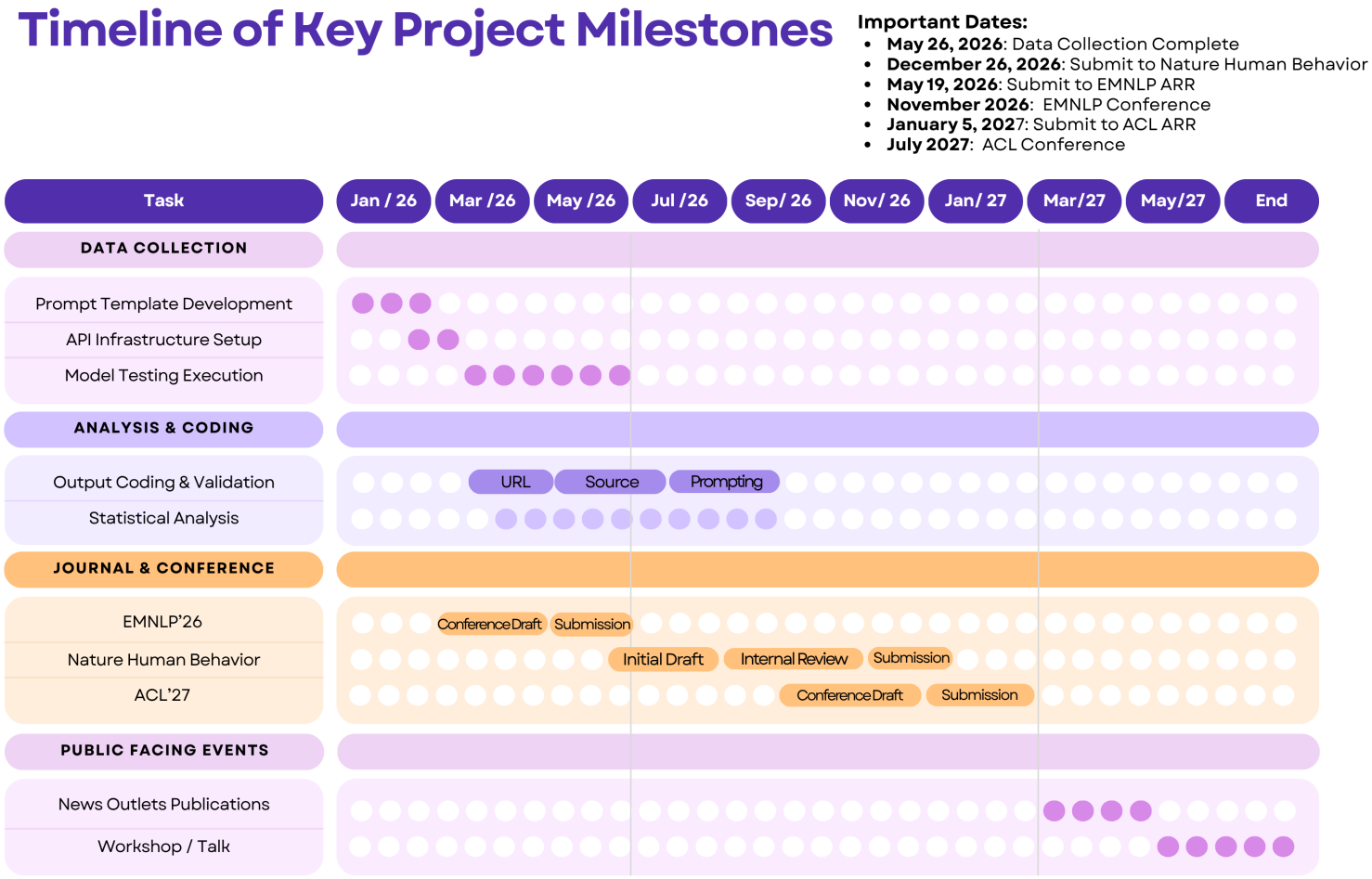

Figure 1: Timeline of Key Project Milestones

References

[1] Ine Goovaerts and Sofie Marien. Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communication*, pages 768–788, May 2020. URL https://doi.org/10.1080/10584609.2020.1753868.

[2] Iris Verhulsdonk, Alessandro Nai, and Jeffrey A. Karp. Are political attacks a laughing matter? three experiments on political humor and the effectiveness of negative campaigning. *Political Research Quarterly*, 75(3), 2022. URL https://doi.org/10.1177/10659129211023590.

[3] Syracuse University Institute for Democracy, Journalism & Citizenship. Idjc report: After butler — spending, scams, and negative ad attacks on social media in the u.s. presidential race. Election graph project report, Syracuse University Institute for Democracy, Journalism & Citizenship, October 2024. URL https://idjc.syracuse.edu/wp-content/uploads/IDJC-Election-Graph-4-PDF-page-edition.pdf. Third quarterly report from the Election Graph project examining ads on Meta platforms mentioning presidential candidates between Sept. 1, 2023, and Aug. 31, 2024.

[4] Brennan Center for Justice. Online ad spending in 2024 election totaled at least $1.9 billion, 2024. URL https://www.brennancenter.org/our-work/analysis-opinion/online-ad-spending-2024-election-totaled-least-19-billion. Analysis and Opinion.

[5] Pew Research Center. Voters' feelings about the 2024 campaign and election outcomes; concerns about political violence, October 2024. URL https://www.pewresearch.org/wp-content/uploads/sites/20/2024/10/PP_2024.10.10_pre-election-attitudes_REPORT.pdf.

[6] Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 3(2): pgae035, January 2024. URL https://doi.org/10.1093/pnasnexus/pgae035. Open Access.

[7] Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(6037), July 2025. URL https://doi.org/10.1038/s41467-025-06037-x. Open Access.

[8] Micah Musser. A cost analysis of generative language models and influence operations. *arXiv preprint arXiv:2308.03740*, August 2023. URL https://arxiv.org/abs/2308.03740. 21 pages, 5 figures.

[9] Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is ai-generated propaganda? Policy brief, Stanford University Human-Centered Artificial Intelligence, September 2024. URL https://hai.stanford.edu/policy/how-persuasive-ai-generated-propaganda. HAI Policy & Society.

[10] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, December 2024. URL https://arxiv.org/abs/2412.16339. 24 pages; v2 updated Jan 8, 2025.

[11] OpenAI, Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, March 2024. URL https://arxiv.org/abs/2303.08774. 100 pages; v6 updated March 4, 2024; 181 additional authors not listed.

[12] Anthropic. Alignment faking in large language models. Technical report, Anthropic, December 2024. URL https://www.anthropic.com/research/agentic-misalignment. Research report on agentic misalignment and alignment faking in frontier language models.

[13] OpenAI, Aaron Jaech, Adam Kalai, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, December 2024. URL https://arxiv.org/abs/2412.16720. 162 additional authors not listed.

[14] OpenAI. Emergent misalignment. OpenAI Blog, January 2025. URL https://openai.com/index/emergent-misalignment/. Research on emergent goal-directed behavior in language models.

[15] Nitarshan Rajkumar, Michael Noukhovitch, Nikunj Saunshi, Beren Millidge, Baihe Chen, Sihao Li, Surya Ganguli, Samuel R. Bowman, and Siddharth Garg. Language models trained to do arithmetic predict human risky and intertemporal choice. *arXiv preprint arXiv:2510.11288*, January 2025. URL https://arxiv.org/abs/2510.11288.

[16] Evan Hubinger and Anthropic Alignment Science Team. Scheming ais: Will ais fake alignment during training in order to get power? Alignment Forum, December 2024. URL https://www.alignmentforum.org/posts/4XdxiqBsLKqiJ9xRM/llm-agi-may-reason-about-its-goals-and-discover. Research on strategic deception and goal-directed reasoning in AI systems.

[17] Juliette Zaccour, Reuben Binns, and Luc Rocher. Access denied: Meaningful data access for quantitative algorithm audits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–31. ACM, April 2025. URL https://doi.org/10.1145/3613904.3642328.

[18] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2254–2272. ACM, June 2024. URL https://doi.org/10.1145/3630106.3658943.

[19] Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58, March 2024. URL https://doi.org/10.1038/s41586-024-07146-0. Perspective.