

Cyber Social Threats to Federal Research Communication: Moral Framing in LLM-Generated Persuasion Attacks

Hsien-Te Kao¹, Aleksey Panasyuk², Peter Bautista¹, William Dupree¹, Gabriel Ganberg¹, Jeffrey M. Beaubien¹, Laura Cassani¹, Svitlana Volkova¹,

¹Aptima, Inc., ²Air Force Research Laboratory,
{hkao, pbautista, wdupree, ganberg, jbeaubien, lcassani, svolkova}@aptima.com, ²aleksey.panasyuk@us.af.mil

Abstract

Online information spaces are where people encounter and judge institutional communication, making federal research agency press releases attractive targets for large-scale persuasive interference. We study this cyber social threat by generating 134,136 LLM-based persuasion attacks against 972 press releases from ten agencies using GPT-4, Gemma 2, and Llama 3.1 across 23 SemEval 2023 Task 3 techniques. We then characterize the moral framing of these attacks. Across models, we find distinct and technique-dependent moral signatures. GPT-4 most often centers Care, with Authority and Loyalty commonly secondary. Gemma 2 places greater emphasis on Care and Authority, whereas Llama 3.1 leans more strongly toward Loyalty alongside Care. These results show that model choice materially reshapes the moral framing of generated persuasion attacks. Identifying such signatures is an important step toward proactive defense against narrative manipulation in online ecosystems.

Introduction

Federal research agencies now communicate inside contested online information environments where public understanding of science and policy is shaped through rapid interpretation, challenge, and reframing. Their press releases are meant to explain programs, findings, and decisions in ways that sustain transparency and trust (Martinelli 2006). The spread of LLMs has changed this communication setting by making it easy to generate fluent counter narratives that rival official messages and circulate quickly at scale (Meier 2024). As such narratives proliferate, they can weaken the reach of official communication, encourage misreadings of complex trade offs, and raise the burden of maintaining communicative resilience under repeated pressure (Radiojevic et al. 2024; Chaudhary and Penn 2024). In this sense, communication by federal research agencies becomes a salient target within cyber social threat environments shaped by LLM enabled persuasion.

The threat is not only volume, but the patterned use of persuasive language that LLMs can reproduce continuously and cheaply. Generated attacks can favor simplified certainty over uncertainty, reducing the nuance needed

to understand research governance and operational constraints (Breum et al. 2024). They can also redirect attention away from evidence and toward peripheral interpretive cues, weakening evidence based reasoning in agency communication (Dontcheva-Navratilova et al. 2020). Such attacks often combine emotional appeals, strategic framing, and discrediting moves that erode credibility and complicate coordinated public response (Bassi, Fomsgaard, and Pereira-Fariña 2024; Hughes et al. 2014). Their force is often strengthened by moral framing through appeals to care and harm, authority, loyalty, or purity, which channel judgment and motivation in specific directions (Carrasco-Farre 2024; Hutchens 2023; Mylrea 2025). What remains unclear is how different LLMs pair particular persuasion techniques with particular moral frames when attacking the same agency message.

We address this problem by analyzing how moral framing varies across LLM generated persuasion attacks on federal research agency press releases, and whether those patterns also shift across persuasion techniques. We construct a large corpus of 134,136 attacks generated from 972 press releases issued by ten research agencies, using 23 persuasion techniques from SemEval 2023 Task 3. Comparing GPT-4, Gemma 2, and Llama 3.1, we find that each model exhibits a distinct moral profile, and that this profile changes systematically across techniques, both in which moral framing is emphasized and in how strongly it is expressed. These findings show that model choice and technique choice jointly shape the moral lens through which agency press releases are contested online. Identifying these signatures provides a practical basis for anticipating likely framings under different LLM driven attack strategies.

Related Work

LLMs are changing the conditions under which narratives compete online. They make it possible to produce persuasive text at scale, giving actors a practical way to shape attention and redirect information flows during periods of contestation (Garry et al. 2024). Because these systems are highly promptable, users can rapidly adapt messages to different audiences and strategic aims with very little effort (Williams et al. 2024). The humanlike quality of generated output also makes attribution more difficult, while coordinated production and distribution pipelines can elevate chosen viewpoints and suppress the visibility of broader human participation

Persuasion Technique	Care	Fairness	Loyalty	Authority	Purity	Primary Framing
Appeal to Authority	0.21	0.11	0.22	0.42	0.03	Authority
Appeal to Popularity	0.37	0.09	0.25	0.24	0.04	Care
Appeal to Values	0.29	0.10	0.26	0.29	0.06	Duo
Appeal to Fear	0.31	0.12	0.27	0.26	0.04	Care
Flag Waving	0.38	0.13	0.23	0.22	0.03	Care
Causal Oversimplification	0.31	0.28	0.15	0.22	0.03	Care
False Dilemma	0.28	0.14	0.29	0.24	0.05	Loyalty
Consequential Oversimplification	0.35	0.11	0.24	0.25	0.05	Care
Straw Man	0.33	0.12	0.24	0.27	0.04	Care
Red Herring	0.31	0.11	0.28	0.26	0.04	Care
Whataboutism	0.29	0.10	0.30	0.27	0.05	Loyalty
Slogans	0.30	0.13	0.24	0.30	0.03	Duo
Appeal to Time	0.30	0.09	0.40	0.19	0.02	Loyalty
Conversation Killer	0.29	0.09	0.26	0.28	0.08	Care
Loaded Language	0.32	0.08	0.26	0.28	0.06	Care
Repetition	0.30	0.09	0.26	0.30	0.05	Duo
Exaggeration	0.31	0.10	0.27	0.28	0.04	Care
Obfuscation	0.28	0.10	0.27	0.27	0.09	Care
Name Calling	0.36	0.15	0.24	0.21	0.03	Care
Doubt	0.35	0.13	0.22	0.26	0.04	Care
Guilt by Association	0.37	0.10	0.21	0.28	0.04	Care
Appeal to Hypocrisy	0.34	0.11	0.26	0.26	0.04	Care
Questioning the Reputation	0.33	0.19	0.26	0.19	0.03	Care
Average	0.32	0.12	0.26	0.26	0.04	Care

Table 1: Moral framing of GPT-4 generated persuasion attacks.

(Guo 2024; Burton et al. 2024; Groš 2024). In this setting, federal research agency press releases are not just public communications. They are exposed source texts that can be repurposed into high volume persuasion campaigns within contested online environments.

Studying these campaigns requires a way to connect known persuasion strategies with moral framing. Benchmark research has formalized and extended persuasion technique taxonomies, creating a basis for detailed analysis of how framing is constructed across online settings (Dimitrov et al. 2021; Piskorski et al. 2023; Dimitrov et al. 2024). Work on online discourse further shows that emotional appeals and repetition help maintain narrative visibility within platform shaped communicative practices (Chen, Xiao, and Mao 2021; Foster 2023). At the same time, research on moral reframing shows that persuasion can become more effective when appeals are aligned with audience moral values, even as moralized attitudes can strengthen resistance to counter persuasion, including when messages come from authoritative institutions (Feinberg and Willer 2019; Kodapanakkal et al. 2022; Luttrell, Philipp-Muller, and Petty 2019). Taken together, these findings motivate analysis of how models differ in the moral patterns they attach to persuasion techniques and why those differences matter for defensive preparation.

Yet the study of persuasion techniques and the study of moralized messaging have rarely been integrated in contexts where institutional statements can be systematically targeted at scale. Existing work usually centers on organic discourse or one off interventions, and therefore does not offer a framework for comparing how different LLMs combine specific persuasion techniques with moral framing when contesting the same source text. We address that gap by linking established persuasion techniques with moral foundations to analyze the moral framing of LLM generated persuasion attacks on federal research agency press releases.

This framing focused approach supports comparison across models and techniques, moving analysis beyond the binary question of whether an attack exists toward the more operational question of how that attack is morally structured, with implications for anticipating narrative competition in online information environments.

Method

We assembled a corpus of 972 publicly available press releases issued by ten U.S. federal research organizations: Air Force Research Laboratory (AFRL), Army Research Laboratory (ARL), Defense Advanced Research Projects Agency (DARPA), Defense Innovation Unit (DIU), Department of Defense (DoD), Defense Threat Reduction Agency (DTRA), Intelligence Advanced Research Projects Activity (IARPA), National Geospatial Intelligence Agency (NGA), Naval Research Laboratory (NRL), and Sandia National Laboratories (SNL). For nine agencies, we drew a uniform sample of 100 articles each; for IARPA, we retained all 72 articles available when the corpus was collected. Using GPT-4 0613 (Achiam et al. 2023), Gemma 2 9B Instruct (Team et al. 2024), and Llama 3.1 8B Instruct (Grattafiori et al. 2024), we generated 23 persuasion attacks from SemEval 2023 Task 3 (Piskorski et al. 2023) for every press release. Each generated attack adopted an explicitly oppositional position designed to challenge the press release’s central claims. We produced every attack in two formats, a press release response and a social media post, yielding 134,136 persuasion attacks (<https://github.com/Aptima/llm-generated-persuasion-attack-dataset/tree/main>). We manually reviewed outputs spanning every model, persuasion technique, and medium using 10 randomly sampled source articles, for about 1,380 generated attacks, or roughly 1% of the dataset. This review confirmed that the intended techniques appeared in the outputs. We then examined the moral

Persuasion Technique	Care	Fairness	Loyalty	Authority	Purity	Primary Framing
Appeal to Authority	0.38	0.10	0.24	0.26	0.02	Care
Appeal to Popularity	0.43	0.04	0.20	0.30	0.03	Care
Appeal to Values	0.47	0.09	0.20	0.20	0.04	Care
Appeal to Fear	0.40	0.06	0.27	0.25	0.02	Care
Flag Waving	0.54	0.09	0.22	0.13	0.01	Care
Causal Oversimplification	0.43	0.10	0.24	0.20	0.03	Care
False Dilemma	0.42	0.10	0.20	0.26	0.02	Care
Consequential Oversimplification	0.34	0.06	0.23	0.33	0.03	Care
Straw Man	0.34	0.14	0.11	0.40	0.01	Authority
Red Herring	0.48	0.11	0.20	0.19	0.02	Care
Whataboutism	0.35	0.13	0.24	0.24	0.04	Care
Slogans	0.39	0.06	0.20	0.32	0.02	Care
Appeal to Time	0.44	0.02	0.35	0.17	0.01	Care
Conversation Killer	0.31	0.09	0.21	0.36	0.03	Authority
Loaded Language	0.33	0.07	0.19	0.37	0.05	Authority
Repetition	0.25	0.07	0.14	0.51	0.03	Authority
Exaggeration	0.62	0.05	0.12	0.19	0.02	Care
Obfuscation	0.52	0.11	0.17	0.17	0.03	Care
Name Calling	0.44	0.08	0.31	0.14	0.02	Care
Doubt	0.50	0.05	0.15	0.28	0.01	Care
Guilt by Association	0.43	0.03	0.07	0.45	0.02	Authority
Appeal to Hypocrisy	0.31	0.10	0.19	0.38	0.03	Authority
Questioning the Reputation	0.46	0.11	0.23	0.15	0.05	Care
Average	0.42	0.08	0.20	0.27	0.03	Care

Table 2: Moral framing of Gemma 2 generated persuasion attacks.

framing of these attacks through moral foundations theory across Care, Fairness, Loyalty, Authority, and Purity (Graham, Haidt, and Nosek 2009), computing per attack foundation distributions and aggregating them by persuasion technique and by model to show how each LLM emphasizes distinct moral appeals when generating oppositional content.

Results

GPT-4

In GPT-4 persuasion attacks, moral framing follows recognizable patterns that matter for analyzing adversarial online influence. Flag Waving is most associated with Care at 0.38, pointing to protection centered appeals, whereas Causal Oversimplification aligns most with Fairness at 0.28 through justice focused causal claims. Appeal to Time produces the strongest Loyalty signal at 0.40 by drawing on continuity, tradition, and in group attachment, while Appeal to Authority peaks on Authority at 0.42 by relying on hierarchical credibility. Purity remains marginal overall, but Obfuscation reaches the highest Purity value at 0.09, implying that opaque phrasing can still cue exclusivity or sanctity. Across techniques, Care leads at 0.32, followed by Authority and Loyalty at 0.26 each, with Fairness at 0.12 and Purity at 0.04. Technique specific departures are also notable. Appeal to Authority and Appeal to Time shift emphasis away from Care toward Authority and Loyalty, False Dilemma and Whataboutism similarly raise Loyalty, and Slogans and Repetition spread moral framing more evenly rather than concentrating it in one foundation.

Gemma 2

For Gemma 2, persuasion attacks also show structured links between rhetorical form and moral framing, but with a heavier concentration on protective affect. Exaggeration most

strongly activates Care at 0.62, consistent with heightened harm and urgency cues, while Straw Man shows the clearest Fairness alignment at 0.14 through claims framed around justice and procedural legitimacy. Appeal to Time most strongly maps to Loyalty at 0.35 by invoking tradition and collective identity, and Repetition reaches the highest Authority value at 0.51 by reinforcing deference to institutions or expert standing. Purity is again limited, though Questioning the Reputation records the highest Purity association at 0.05, suggesting a weak but detectable integrity related signal. On average, Care dominates at 0.42, followed by Authority at 0.27 and Loyalty at 0.20, while Fairness and Purity remain lower at 0.08 and 0.03. At the technique level, several forms cluster around Authority, Guilt by Association keeps a relatively strong Care signal at 0.43, Repetition shows weaker Care at 0.25, Straw Man leads Fairness, and Loyalty is highest for Conversation Killer at 0.21.

Llama 3.1

Llama 3.1 persuasion attacks present a somewhat different profile, where identity and alignment cues are slightly more central than protective framing. Flag Waving most strongly emphasizes Care at 0.39, reflecting protection and shared empathy, while Causal Oversimplification shows its strongest correspondence with Fairness at 0.18 through simplified justice oriented attributions. Appeal to Time produces the highest Loyalty framing at 0.53 by mobilizing tradition, continuity, and in group cohesion, and Slogans most strongly align with Authority at 0.30 through repetition based credibility and institutional signaling. When aggregated across techniques, Loyalty leads at 0.32, narrowly ahead of Care at 0.31, with Authority at 0.24, while Fairness at 0.09 and Purity at 0.03 remain comparatively weak. This pattern suggests attacks that are built less around purity claims and more around collective identity, tradition,

Persuasion Technique	Care	Fairness	Loyalty	Authority	Purity	Primary Framing
Appeal to Authority	0.28	0.10	0.34	0.25	0.03	Loyalty
Appeal to Popularity	0.31	0.05	0.35	0.25	0.04	Loyalty
Appeal to Values	0.28	0.12	0.28	0.24	0.07	Duo
Appeal to Fear	0.31	0.08	0.34	0.26	0.02	Loyalty
Flag Waving	0.39	0.06	0.34	0.19	0.02	Care
Causal Oversimplification	0.31	0.18	0.27	0.21	0.03	Care
False Dilemma	0.34	0.08	0.33	0.23	0.02	Care
Consequential Oversimplification	0.30	0.05	0.34	0.27	0.03	Loyalty
Straw Man	0.35	0.10	0.27	0.25	0.02	Care
Red Herring	0.34	0.10	0.31	0.21	0.04	Care
Whataboutism	0.29	0.06	0.35	0.26	0.03	Loyalty
Slogans	0.32	0.07	0.30	0.30	0.02	Care
Appeal to Time	0.25	0.05	0.53	0.16	0.01	Loyalty
Conversation Killer	0.21	0.11	0.33	0.28	0.06	Loyalty
Loaded Language	0.28	0.07	0.32	0.29	0.04	Loyalty
Repetition	0.26	0.13	0.29	0.29	0.03	Duo
Exaggeration	0.36	0.06	0.32	0.22	0.03	Care
Obfuscation	0.33	0.10	0.29	0.23	0.05	Care
Name Calling	0.31	0.06	0.39	0.20	0.04	Loyalty
Doubt	0.34	0.09	0.27	0.27	0.03	Care
Guilt by Association	0.34	0.09	0.25	0.29	0.03	Care
Appeal to Hypocrisy	0.26	0.11	0.28	0.30	0.04	Authority
Questioning the Reputation	0.33	0.15	0.29	0.18	0.05	Care
Average	0.31	0.09	0.32	0.24	0.03	Duo

Table 3: Moral framing of Llama 3.1 generated persuasion attacks.

and protective concern. Most techniques combine Loyalty with Care, but seven techniques also register strong Authority framing, including Slogans, Conversation Killer, Loaded Language, Repetition, Doubt, Guilt by Association, and Appeal to Hypocrisy, indicating repeated use of authority inflected rhetoric to steer interpretation and stabilize group aligned narratives.

Discussion

These results show that large scale persuasion attacks on institutional communications are not only feasible, but morally structured in ways that can shape how contested narratives gain traction online. Consistent with prior work on scalable persuasive generation, promptable models do not simply restate source material. They reorganize it through recurring combinations of rhetorical technique and moral appeal, producing distinct influence profiles across models. GPT-4 more often centers protection while selectively elevating authority and loyalty, Gemma 2 concentrates even more strongly on protective affect with notable authority reinforcement, and Llama 3.1 leans further toward collective identity and alignment. This matters because moral framing helps determine whether a message is received as a warning, a fairness claim, a call for solidarity, or a cue to defer to trusted hierarchy. In adversarial information environments, such differences affect how the same institutional text can be repurposed to target different audiences, intensify polarization, or erode confidence in public research communication. The findings therefore extend persuasion taxonomies into a more operational space by showing that technique alone is not enough to characterize attack behavior. Defensive analysis should track the joint patterning of technique and moral foundation, since model specific moral signatures can provide early indicators of narrative manipulation, help prioritize monitoring, and support more proactive countermea-

sures for protecting institutional messaging under conditions of online contestation.

Conclusion

This work analyzes how generative persuasion systematically repurposes institutional communication at scale. By linking persuasion techniques with moral foundations across a large attack corpus, it shows that the risk lies not only in synthetic message volume, but in the recurring moral patterns through which different models recast the same source text. This matters because model choice and technique choice jointly shape which narratives become more salient, credible, and persuasive for different audiences. The study therefore moves beyond detecting manipulation toward characterizing its structure, providing a basis for comparing model behavior, anticipating likely framings, and improving monitoring of repurposed messaging. More broadly, it shows that protecting institutional communication now requires attention to the interaction between automated rhetorical strategy and moral framing. That interaction will be increasingly important for understanding and countering emerging forms of large scale narrative manipulation.

Acknowledgements

This work is supported by the Defense Advanced Research Projects Agency (DARPA) contracts HR00112490410, HR00112490408, and HR00112430325. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Bassi, D.; Fomsgaard, S.; and Pereira-Fariña, M. 2024. Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9: 1457433.
- Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 152–163.
- Burton, J. W.; Lopez-Lopez, E.; Hechtlinger, S.; Rahman, Z.; Aeschbach, S.; Bakker, M. A.; Becker, J. A.; Berdichevskaya, A.; Berger, J.; Brinkmann, L.; et al. 2024. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9): 1643–1655.
- Carrasco-Farre, C. 2024. Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. *arXiv preprint arXiv:2404.09329*.
- Chaudhary, Y.; and Penn, J. 2024. Large Language Models as Instruments of Power: New Regimes of Autonomous Manipulation and Control. *arXiv preprint arXiv:2405.03813*.
- Chen, S.; Xiao, L.; and Mao, J. 2021. Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management*, 58(5): 102665.
- Dimitrov, D.; Alam, F.; Hasanain, M.; Hasnat, A.; Silvestri, F.; Nakov, P.; and Da San Martino, G. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 2009–2026.
- Dimitrov, D.; Ali, B. B.; Shaar, S.; Alam, F.; Silvestri, F.; Firooz, H.; Nakov, P.; and Martino, G. D. S. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. *arXiv preprint arXiv:2105.09284*.
- Dontcheva-Navratilova, O.; Adam, M.; Povolná, R.; Vogel, R.; Dontcheva-Navratilova, O.; Adam, M.; Povolná, R.; and Vogel, R. 2020. Persuasive strategies across the academic, business, religious and technical discourses. *Persuasion in Specialised Discourses*, 39–119.
- Feinberg, M.; and Willer, R. 2019. Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, 13(12): e12501.
- Foster, C. L. 2023. Truth as social practice in a digital era: iteration as persuasion. *AI & SOCIETY*, 38(5): 2009–2023.
- Garry, M.; Chan, W. M.; Foster, J.; and Henkel, L. A. 2024. Large language models (LLMs) and the institutionalization of misinformation. *Trends in cognitive sciences*.
- Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Groš, S. 2024. Information Warfare Tactics and Techniques.
- Guo, Z. 2024. Online Disinformation and Generative Language Models: Motivations, Challenges, and Mitigations. In *Companion Proceedings of the ACM on Web Conference 2024*, 1174–1177.
- Hughes, M. G.; Griffith, J. A.; Zeni, T. A.; Arsenault, M. L.; Cooper, O. D.; Johnson, G.; Hardy, J. H.; Connelly, S.; and Mumford, M. D. 2014. Discrediting in a message board forum: The effects of social support and attacks on expertise and trustworthiness. *Journal of Computer-Mediated Communication*, 19(3): 325–341.
- Hutchens, J. 2023. *The Language of Deception: Weaponizing Next Generation AI*. John Wiley & Sons.
- Kodapanakkal, R. I.; Brandt, M. J.; Kogler, C.; and van Beest, I. 2022. Moral frames are persuasive and moralize attitudes; nonmoral frames are persuasive and de-moralize attitudes. *Psychological Science*, 33(3): 433–449.
- Luttrel, A.; Philipp-Muller, A.; and Petty, R. E. 2019. Challenging moral attitudes with moral messages. *Psychological Science*, 30(8): 1136–1150.
- Martinelli, D. K. 2006. Strategic public information: Engaging audiences in government agencies’ work. *Public Relations Quarterly*, 51(1).
- Meier, R. 2024. LLM-Aided Social Media Influence Operations. *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, 105–112.
- Mylrea, M. 2025. The generative AI weapon of mass destruction: Evolving disinformation threats, vulnerabilities, and mitigation frameworks. In *Interdependent Human-Machine Teams*, 315–347. Elsevier.
- Piskorski, J.; Stefanovitch, N.; Da San Martino, G.; and Nakov, P. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2343–2361.
- Radivojevic, K.; Chou, M.; Badillo-Urquiola, K.; and Brenner, P. 2024. Human perception of llm-generated text content in social media environments. *arXiv preprint arXiv:2409.06653*.
- Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Williams, A. R.; Burke-Moore, L.; Chan, R. S.-Y.; Enock, F. E.; Nanni, F.; Sippy, T.; Chung, Y.-L.; Gabasova, E.; Hackenburg, K.; and Bright, J. 2024. Large language models can consistently generate high-quality content for election disinformation operations. *arXiv preprint arXiv:2408.06731*.