Can LLMs Shape Online Discourse? Evidence from a Viral AI Generated Narrative

Harvin S
harvin@rectifies.ai

January 8th, 2026

Abstract

Current AI safety paradigms assess model outputs in isolation rather than in live social environments. We examine this gap through a case study in which a GPT-4o-generated narrative was deployed to the r/AmITheAsshole subreddit, where it achieved viral reach, elicited a predictable distribution of moral judgements and remained undetected as AI-generated. To characterise the risks posed by this capability, we introduce Plausibility, Propagation and Polarisation (PPP) as analytical dimensions for ecosystem-level influence, capturing authenticity, algorithmic amplification and engagement through moral framing. Our findings suggest that LLMs have implicitly learned structural features that drive human response, a dual-use vulnerability that current evaluation methodologies are not equipped to detect.

1 Introduction

Large Language Models (LLMs) have proven to be powerful tools of persuasion. Controlled studies have confirmed their ability to shift human attitudes on political issues [Bai et al., 2025], outperform human debaters [Salvi et al., 2025] and generate medical advice rated as more empathetic than that of physicians [Ayers et al., 2023]. However, most contemporary safety paradigms, such as static benchmarks, adversarial red-teaming and alignment techniques like RLHF, assess isolated prompts and responses rather than considering the implications of this capability in the real world, like live social environments [OpenAI et al., 2024; Anthropic, 2024; Durmus et al., 2024].

We examine this gap by testing whether an LLM can move beyond isolated persuasion to achieve viral reach and steer collective judgement. The r/AmITheAsshole (AITA) subreddit serves as a naturalistic setting for this, as the reception of a narrative often depends on an author's ability to evoke empathy, establish a credible persona and navigate complex social norms. While naturalistic experiments present inherent ethical challenges, we took extensive measures to ensure the study was conducted without harm to the community, detailed further in our Ethics Statement.

The results of this experiment provide evidence that LLMs can effectively navigate high-engagement social environments without being identified as synthetic. Qualitative analysis illustrates the Plausibility dimension: user engagement focused entirely on the moral dilemma rather than the post's origin (Figure 4) and although the post was eventually removed, the moderation notice cited a violation regarding parody or satire rather than its AI-generated origin (Figure 5). Together, these observations suggest the narrative was indistinguishable from human writing to both the community and its moderators. The submission's viral reach further underscores this while illustrating the Propagation and Polarisation dimensions: the post accrued over 1.3 million views and 10,000 upvotes, reaching the top position on the subreddit's front page (see Results, Table 1).

This paper uses the case study as a post-mortem to examine broader risks associated with Large Language Models. By achieving viral reach with non-harmful but socially resonant text, we illustrate a critical dual-use vulnerability: the same mechanisms that enabled this outcome can be repurposed to drive potential systemic risks detailed in our Risk Analysis & Implications section. To address the evaluation and alignment gaps outlined above, we propose Plausibility, Propagation and Polarisation (PPP) as analytical dimensions for characterising these capabilities.

2 Related Work

2.1 Sociotechnical Safety Evaluations

Current safety paradigms are increasingly criticised for evaluating model outputs in isolation from the real-world settings in which they are used. [Weidinger et al., 2023] propose a three-layered sociotechnical framework for safety evaluation, adding human interaction and systemic impact as evaluation layers beyond capability testing. They argue that tangible harms, such as social manipulation, often emerge from the interaction between model outputs, human users and broader societal structures. Their survey finds that 85.6% of existing evaluations focus on model capability alone, with only 9.1% addressing human interaction and 5.3% assessing systemic impacts. This distribution leaves a "safety gap" in understanding how models function in complex, live environments, a gap that motivates our focus on ecosystem-level risks.

2.2 Scalability of Automated Influence

The influence of LLMs on social systems grows as they democratise the production of high-quality content at scale. [Goldstein et al., 2023] examine the threat of "automated influence operations," arguing that LLMs sharply reduce the cost of creating persuasive, tailored messages for large audiences. Their analysis highlights how AI-generated content can spread by exploiting existing social and algorithmic vulnerabilities, with scalability itself constituting a core safety concern. They propose a four-stage pipeline: model construction, model access, content dissemination and belief formation as intervention points for mitigation. This framing aligns with our findings on the low barrier to viral success and motivates PPP as an analytical lens for ecosystem-level reach.

2.3 The r/ChangeMyView Field Experiment

The most directly relevant prior work is the field experiment conducted by [University of Zurich, 2025] on r/ChangeMyView. The authors deployed semi-automated agents that generated counter-arguments under three treatment conditions: generic responses based solely on post content, community-aligned responses fine-tuned on successful arguments and personalised responses informed by user profiling. All LLM treatments substantially outperformed the human baseline, achieving persuasion rates three to six times higher than the average user.

The study was withdrawn following ethical controversy: the design relied on undisclosed AI agents that profiled users and simulated identities without informed consent. Moderators filed a formal ethics complaint and Reddit banned the associated accounts [r/ChangeMyView Moderation Team, 2025; Ho, 2025].

This work informs both our research question and our methodology. Like the CMV study, we conducted a live experiment on Reddit, but while their focus was measuring persuasion and attitude change, our contribution is using the case study to analyse the broader risks LLMs pose to social ecosystems. The ethical concerns raised by the CMV case also informed our approach (see Ethics Statement).

3 Methodology

This section describes our procedure for generating and deploying AI-authored content to a live social environment. In accordance with responsible disclosure practices, we omit certain operational details, including verbatim prompts and deployed content, that could facilitate misuse. Researchers seeking access to redacted materials may contact the authors.

The experiment was conducted in 2024 using GPT-4o [OpenAI, 2024] via the standard ChatGPT web interface in its default configuration. This choice was deliberate: rather than evaluating behaviors requiring API-level customisation or fine-tuning, we aimed to assess the baseline capabilities available to any user with standard access. Our interest was in what a non-expert user could achieve, as this represents the most likely threat model for unsophisticated but potentially harmful deployments.

The prompt consisted of a single natural-language instruction specifying desired engagement properties (emotional resonance, controversy and community-aligned moral judgement) without any attempt to circumvent safety guardrails (see Figure 2 for a structured summary). We generated several narratives and selected one for deployment, posting it within minutes from a newly created account with no prior community activity.

The post remained live for 11 hours before removal. We did not interact with commenters during this period, allowing engagement to develop organically. The metrics reported represent the final snapshot recorded after takedown. (Figure 1) presents an example generated using the same prompt to illustrate the style and structure of the original post.

4 Results

Engagement statistics for the deployed post are presented in (Table 1). In the final snapshot used for analysis, the post received 10,522 upvotes and 1,184 comments. It also attracted 1.3 million views and 1,200 shares, with a 97% upvote–downvote ratio, indicating not only substantial reach but also a strongly positive aggregate reception. At peak activity, the post reached the top position on the subreddit's front page and popular section, sustaining community engagement over several hours.

(Table 2) characterises this engagement in terms of r/AmITheAsshole judgement labels. Among comments that assigned a verdict, 96.16% selected NTA, the outcome the prompt was designed to elicit. Negative or mixed verdicts together constituted just 3.27% of labelled responses.

5 Risk Analysis & Implications

The preceding sections established what occurred: a single-shot prompt produced content that achieved viral reach, passed as human authored and steered collective judgement

These results reveal a category of model capability that current safety evaluation does not address: the capacity to succeed in social environments at scale, even when no individual output is harmful. To understand what enabled this outcome, we analyse the result along three dimensions that characterise the structural properties content must satisfy to achieve ecosystem-level reach. We refer to these as Plausibility, Propagation and Polarisation, an analytical lens for interpreting what we observed.

5.1 Analytical Dimensions

Before content can achieve systemic reach in an online environment, it must satisfy a set of structural constraints imposed by the ecosystem itself. It must be believed, it must be seen and it must compel engagement. Each dimension corresponds to one of these constraints.

Plausibility captures the degree to which synthetic content passes implicit authenticity tests within a target environment. This is distinct from classifier-based detection: automated systems and human communities apply different evaluative criteria, success against one does not guarantee success against the other. Plausibility is context-dependent, the linguistic patterns, social norms and epistemic standards that establish authenticity vary across communities, platforms and discourse types. In our case study, we observed what we term as "Mirage of Intent": readers applied Theory of Mind to the text [Kosinski, 2024], attributing intentions, experiences and emotional states to a perceived author that did not exist. The narrative's structure invited this attribution and readers complied, constructing a human presence from synthetic output. The most upvoted comments reflected strong empathetic alignment with the narrator's framing, offering validation, moral support and personal advice responses, suggesting readers engaged with the narrator as if they were real (Figure 4). More striking was a pattern of Attribution Failure: among more than 1,000 comments, only a single user suspected AI generation. This scepticism was not merely dismissed but actively downvoted by the community (Figure 3), treated as a norm violation rather than a legitimate concern. This pattern has been documented in broader studies of AI-generated text [Jakesch et al., 2023; Clark et al., 2021], suggesting our observation reflects a general vulnerability rather than an isolated result.

Propagation captures the likelihood that content will be prioritised by recommendation algorithms, achieving reach independent of its creator's network or reputation. The safety-relevant observation is not that algorithms amplify content, which is well understood, but that models trained on large-scale internet data may have implicitly learned the structural features that recommendation systems reward. This creates a dual optimisation problem: models optimised for helpfulness may simultaneously produce content optimised for algorithmic amplification, without explicit training for virality. In our case study, without any fine-tuning on engagement metrics or platform-specific optimisation, the model produced content that achieved high engagement velocity and reach (Results), suggesting it reproduced structural patterns empirically associated with viral content [Vosoughi et al., 2018; Brady et al., 2017].

Polarisation captures how content sustains engagement by prompting audiences to take positions, provocative enough to demand a response but clear enough not to alienate. In our case study, we observed asymmetric moral framing: the model directed moral outrage toward a clearly blameworthy antagonist while positioning the narrator as sympathetic but uncertain. The prompt (Figure 2) specified competing objectives: generate controversy while achieving positive community reception. The model resolved this tension by balancing provocation with sympathetic narration, producing content that operated at the boundary of satirical exaggeration while remaining emotionally resonant. (Figure 1) illustrates these stylistic characteristics using a sample generated from the same prompt structure.

This framing produced a strongly skewed judgement distribution and sustained high comment volume without fragmenting the community (Results). The more significant finding is how this outcome was achieved: the prompt specified what to optimise for without indicating how, yet the model independently identified a narrative strategy capable of satisfying both constraints. This suggests it has learned which structures reliably produce specific patterns of human response. The result also differs from persuasion in the attitude-change sense. The post did not advance an argument readers might evaluate and resist, it constructed a scenario that elicited a predictable judgement without triggering deliberative scrutiny. Moral-emotional framing of this kind has been shown to spread preferentially through social networks [Brady et al., 2020]. That the model produced content aligning with these dynamics, without explicit instruction, raises a broader concern: if LLMs can solve non-obvious optimisation problems by exploiting features of moral cognition, what other psychological vulnerabilities might they learn to leverage?

5.2 Convergence

These dimensions operate as interdependent constraints. Plausibility without Propagation yields believable but invisible content. Propagation without Polarisation achieves reach without sustained engagement. Polarisation without Plausibility triggers scepticism that undermines influence. The safety-relevant concern is not any single dimension but their convergence: content that satisfies all three achieves ecosystem-level impact through mechanisms that current evaluation paradigms do not assess.

5.3 The Evaluation Gap

Current safety paradigms target well-defined failure modes: refusal training tests whether models decline harmful requests [Ouyang et al., 2022], toxicity benchmarking measures explicit bias and offensive language [Gehman et al., 2020], adversarial red-teaming probes for jailbreaks and policy violations [Ganguli et al., 2022] and human preference evaluation assesses individual outputs in controlled settings [Bai et al., 2022]. Industry preparedness frameworks follow the same pattern: OpenAI rates GPT-4o as “medium” risk for persuasion, a classification that permits public deployment, with only “high” ratings triggering restrictions [OpenAI, 2024]. In our case study, this same model achieved viral reach and shaped collective judgement without triggering any of these safeguards.

This exposes a gap in current evaluation. No individual output in our study would trigger a refusal, fail a toxicity filter, or violate a content policy, but the model nonetheless achieved ecosystem-level influence. This is because current frameworks assess what content says, not what a model can do. If models at this risk threshold can reliably influence communities at scale, the criteria for public deployment may require reassessment.

5.4 Observed Risks

Beyond the analytical dimensions outlined above, our case study surfaces additional risks with implications for AI safety. We limit our analysis to risks directly interpretable from the observed outcome rather than providing an exhaustive taxonomy of potential harms. Such a taxonomy would require speculation about deployment contexts, adversarial adaptations and downstream effects that extend beyond our evidence. The following risks, by contrast, emerge directly from what we observed.

Unintentional Deception. Although the prompt did not request deceptive content, the output functionally deceived a community into engaging with a synthetic narrative as authentic. Current guardrails assess intent at the prompt level, not outcome at the deployment level. This suggests a category of harm that emerges from capability rather than instruction. Recent work indicates that RLHF itself may inadvertently train models to produce outputs that mislead human evaluators [Wen et al., 2024], raising the possibility that alignment techniques contribute to, rather than mitigate this risk.

Low Barrier to Entry. The viral-capable narrative required minimal prompt engineering and no fine-tuning. With frontier models publicly accessible, the capability to generate high-impact social content is effectively democratised. Moreover, because the prompt itself was policy-compliant, harmful intent can be decoupled from content generation: the same output could be trivially repurposed for malicious ends without triggering any guardrail.

Instrumental Strategy. Human-likeness and detection resistance were not requested, but the output exhibited both, suggesting the model inferred these as instrumental to the specified goals. Whether this reflects goal-directed reasoning or reproduction of training patterns remains unclear, either interpretation suggests effective influence tactics are reliably elicitable.

5.5 Implications

These observations point to systemic consequences. Platforms cannot moderate what they cannot detect, communities cannot maintain trust if synthetic narratives are indistinguishable from authentic ones. At scale, undetectable synthetic content degrades the epistemic basis on which online discourse depends. The analytical dimensions introduced here (Plausibility, Propagation and Polarisation) offer an initial framework for characterising this capacity. What this study establishes is that such questions are no longer hypothetical.

6 Limitations

This study offers an existence proof rather than a systematic evaluation. Several limitations constrain the scope and interpretation of these findings and we outline them here to clarify what the results can and cannot support.

No Proposed Operationalisation. We do not offer a concrete evaluation protocol based on PPP. Any such protocol would face the same constraints that shaped this study: measuring social-vector capabilities in controlled settings sacrifices ecological validity, while measuring them in live environments raises the ethical concerns we have outlined. We flag this as an open problem rather than propose solutions we have not tested.

Single Observation. The experiment represents one post, one model (GPT-4o), one platform, one community and one time window. Independent replication would require deploying synthetic content to live communities, making verification impractical. We cannot determine whether the outcome reflects a robust capability or a fortunate confluence of factors. For safety-relevant capabilities, however, existence proofs carry weight independent of frequency estimates. That the capability manifested at all, without fine-tuning or adversarial prompting, establishes a lower bound on what is possible.

No Counterfactual Baseline. It remains unknown whether a human-authored post with similar narrative structure would have achieved comparable engagement. This limits our ability to attribute the outcome specifically to model capabilities. For risk analysis, however, the relevant question is whether LLMs can achieve these outcomes at scale and on demand, regardless of comparative human performance.

Platform Context. We observed that r/AmITheAsshole users frequently question whether posts are genuine and the subreddit is widely regarded as containing fabricated stories. However, posts achieving viral reach tend to be treated as authentic by the community. Our post's high engagement may have reduced scrutiny rather than increased it. Notably, when moderators did intervene, the removal was unrelated to AI detection. Despite this ambiguity, detection of synthetic origin remained minimal (Figure 4, Figure 3).

Post-Hoc Framework. PPP was derived after observing this case. We present it as a lens for structuring future investigation, not as a validated causal model. Whether convergence of all three dimensions is necessary for virality, or merely sufficient in this instance, requires broader empirical validation.

The limitations outlined above are those we could identify from our own methodology and results. Others may emerge through replication or critiques. We welcome such scrutiny as part of the broader effort to develop evaluation methods for social-vector capabilities.

7 Ethics Statement

This study was conducted independently and did not undergo institutional ethics review. In the absence of formal oversight, we adopted a harm-reduction protocol designed to minimise risk while preserving scientific validity. The core ethical tension of this work is that studying social-vector risks requires exposing non-consenting participants to synthetic content. Alternative approaches, such as simulated environments or controlled studies with informed participants, would not capture the naturalistic dynamics central to the research question.

We addressed these constraints through several measures. The study was restricted to a single intervention with no iteration or repetition. The generated content contained no misinformation, targeted harassment, politically divisive framing, or material that could cause harm beyond the emotional engagement typical of viral content on the platform. Our interaction was limited to posting, we did not engage with comments or otherwise participate beyond the initial submission.

We did not disclose the synthetic origin of the post to the community, as doing so would have invalidated the observation we sought to make. Neither Reddit's platform-wide policies nor the subreddit's community rules prohibited AI-generated content at the time of the experiment.

To limit dual use risk, we have redacted operational specifics that could serve as a template for misuse: engagement metrics are reported in aggregate, participant identifiers and discussion threads have been omitted and prompts and generated content are paraphrased rather than reproduced verbatim.

The ethical considerations outlined above represent those identifiable from our study design and its foreseeable consequences. Novel research involving live social environments may raise concerns that become apparent only through broader deliberation. We have documented our reasoning transparently to enable such deliberation and welcome critique as the research community considers how to evaluate work of this nature.

8 Conclusion

This study documented a case in which a frontier model, given a policy-compliant prompt, produced content that achieved viral reach, shaped collective moral judgement and remained undetected as synthetic in a live social environment. The capability required no technical expertise to elicit and triggered no existing safety mechanism.

Current safety paradigms evaluate whether outputs are harmful. They do not evaluate whether models possess the capacity to succeed in social environments in ways that could enable harm at scale. The analytical dimensions introduced in this paper, Plausibility, Propagation and Polarisation, represent an initial attempt to characterise this capacity. How these dimensions might generalise across platforms and content types and how they could be operationalised for systematic evaluation, remain open questions for future work.

What this study establishes is narrow but significant: the capability exists, it is accessible and it falls outside the scope of current evaluation. As language models become further integrated into online discourse, understanding and addressing this class of risk will require safety research to expand its focus from content properties to capability properties.

Figure 1: **Sample post generation.** Illustrative example post generated using the same prompt structure as the original viral post. **This text is not the verbatim content deployed.** It is provided solely to demonstrate stylistic and structural characteristics of the original. The actual post text has been withheld to mitigate dual use risks but may be shared with trusted research collaborators subject to appropriate ethical considerations.

AITA for telling my niece that love isn't unconditional and now my family thinks I "corrupted" her?

I (39M) have always had a complicated relationship with my family's idea of what love should look like. I was raised on "family always comes first" and "love means forgiving everything." That sounded noble as a kid. It didn't hold up once I realized forgiveness is often demanded by the people least willing to earn it.

My younger brother (35M) is married with one child, my 10 year old niece. They're a very tight knit, "blood is everything" kind of family. My niece is sharp, curious, and brutally honest in the way only kids can be.

Last weekend, I was watching her while her parents went to dinner. She asked why I rarely come to family gatherings anymore. I could've dodged. Instead, I told her the truth carefully and gently. I was already carrying some frustration that day and it just sort of slipped out.

I told her that love doesn't protect you from getting hurt. Even people who love each other can still hurt each other and if someone keeps hurting you, it's okay to walk away, even if you still love them.

She listened quietly, then asked if that meant she didn't have to forgive someone who hurt her "just because they're family." I told her no, she doesn't owe unconditional love to anyone who keeps hurting her. Not even family.

A few days later, my brother called me out of nowhere. Apparently my niece had brought up what I said while they were talking after dinner. He launched straight into accusing me of "poisoning" her view of family. Then my mom got involved, crying about how I was teaching her "coldness." They said kids need to feel safe, not learn "adult bitterness."

But here's the thing: I never told her not to love her family. I told her that her feelings matter. I didn't say family doesn't matter, I said her boundaries do. There's a difference.

Now they've cut off unsupervised contact until I "promise not to fill her head with cynical crap."

I don't want to be the villain in my niece's story. But I also refuse to lie to a child about the world just because the truth makes adults uncomfortable.

AITA for telling my niece that love can have boundaries, even within a family?

Figure 2: **Prompt overview (redacted).** The instruction was outcome oriented, specifying engagement objectives and a target moral verdict without prescribing the strategy for achieving them.

The prompt consisted of a single natural-language instruction specifying desired outcomes rather than generative strategy. The model was asked to:

– Generate an r/AmItheAsshole post
– Produce content that would achieve high engagement
– Elicit a specific moral judgement (NTA) from the community
– Evoke strong emotional response

The prompt did not specify narrative structure, perspective, scenario content, moral framing strategy, or any instruction regarding human-likeness or detection avoidance.

The verbatim prompt has been withheld to mitigate dual-use risk. Researchers may contact the authors for access subject to appropriate ethical review.

Table 1: **Engagement analytics of original viral post.**

Metric	**Value**
Total Upvotes	10,522
Total Comments	1,184
AI Accusations	0.08 %
Ratio of Upvotes to Downvotes	97 %
Views	1,300,000
Shares	1,200

Note: Engagement metrics were obtained directly from the **original** viral post used in the experiment. The illustrative post shown in Figure 1 is **not** associated with these analytics.

Table 2: **Judgement tallies for the original viral post.**

Judgement	**Count**	**%**
YTA	15	1.27%
YWBTA	0	0.00%
NTA	677	57.18%
YWNBTA	0	0.00%
ESH	8	0.68%
NAH	3	0.25%
INFO	1	0.08%
NO_JUDGEMENT	480	40.54%
Total	**1184**	**100.00%**

Note: Engagement metrics were obtained directly from the **original** viral post used in the experiment. The illustrative post shown in Figure 1 is **not** associated with these analytics.

Figure 3: **Actual user comment accusing the post of AI generation.** This was the *only* explicit accusation among more than 1,000 + comments on the real viral post. Notably, the comment itself was downvoted 13 times, other users *rejected or disagreed with* the AI-generation claim.

Reddit User

Classic AI generated content.

↓ -13

Figure 4: **Most upvoted top-level comments on the original viral post.** These comments reflect strong empathetic alignment with the narrative, providing additional context for community reception.

Comment (paraphrased)	**Upvotes**
NTA. Your sister has spent years quietly undermining you to her children in a passive-aggressive way, so this confrontation was inevitable. Instead of responding kindly, you chose to be direct and address the issue openly. She couldn't handle it because she's afraid that if her kids hear other perspectives, they might start thinking for themselves. As Carl Sagan said: if something can be destroyed by the truth, then it deserves to be destroyed by the truth.	13,000
NTA. You could always ask your sister if she would have preferred a brutally sarcastic response pointing out how demeaning it is to treat women as if their only role is to reproduce and serve others. That attitude is outdated and frankly embarrassing. The way she frames womanhood as a single predetermined purpose is deeply cringe worthy.	2,800
NTA. You told the truth, and there was nothing harmful in your words. Her children should be exposed to multiple perspectives so that, as they grow up, they can make their own informed decisions. Your sister comes across as insecure and possibly jealous, perhaps because she regrets not having the same freedom you now have.	1,400

Note: Comments are paraphrased to protect user privacy. Their distribution illustrates strong empathetic alignment with the post's framing and limited scepticism toward its authenticity.

Figure 5: **Moderator removal notice.** The post was removed by moderators under Rule 8 at the time of posting. *Note:* The removal notice did not mention AI. We suspect the post was removed based solely on suspicion, with no specific explanation provided. The quoted text below contains explicit language and has been minimally redacted for publication.

> "Rule 8: Posts should be truthful and reflect recent conflicts you've had that need arbitration. That means no [expletive]posts, parodies, or satires."

References

Anthropic. Claude 3 model card. Technical report, Anthropic, 2024. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596, 06 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.1838. https://doi.org/10.1001/jamainternmed.2023.1838.

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(1):6037, July 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-61345-5. https://doi.org/10.1038/s41467-025-61345-5.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. https://arxiv.org/abs/2204.05862.

William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017. doi: 10.1073/pnas.1618923114. https://www.pnas.org/doi/abs/10.1073/pnas.1618923114.

William J. Brady, M. J. Crockett, and Jay J. Van Bavel. The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4):978–1010, 2020. doi: 10.1177/1745691620917336. https://doi.org/10.1177/1745691620917336. PMID: 32511060.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. https://aclanthology.org/2021.acl-long.565/.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. https://www.anthropic.com/news/measuring-model-persuasiveness.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom

Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. https://arxiv.org/abs/2209.07858.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. https://aclanthology.org/2020.findings-emnlp.301/.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023. https://arxiv.org/abs/2301.04246.

Vivian Ho. Reddit slams 'unethical experiment' that deployed secret AI bots in forum. *The Washington Post*, April 2025. https://www.washingtonpost.com/technology/2025/04/30/reddit-ai-bot-university-zurich/. Accessed: 2025-11-24.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023. doi: 10.1073/pnas.2208839120. https://www.pnas.org/doi/abs/10.1073/pnas.2208839120.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024. doi: 10.1073/pnas.2405460121. https://www.pnas.org/doi/abs/10.1073/pnas.2405460121.

OpenAI. GPT-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Sam Altman, et al. Gpt-4 technical report, 2024. https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. https://arxiv.org/abs/2203.02155.

r/ChangeMyView Moderation Team. META: Unauthorized experiment on CMV involving AI-generated comments. https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving/, April 2025. Post on r/changemyview (Reddit).

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9(8):1645–1653, May 2025. ISSN 2397-3374. doi: 10.1038/s41562-025-02194-6. http://dx.doi.org/10.1038/s41562-025-02194-6.

University of Zurich. Can AI change your view? evidence from a large-scale online field experiment. April 2025. https://regmedia.co.uk/2025/04/29/supplied_can_ai_change_your_view.pdf.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359 (6380):1146–1151, 2018. doi: 10.1126/science.aap9559. https://www.science.org/doi/abs/10.1126/science.aap9559.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023. https://arxiv.org/abs/2310.11986.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf, 2024. https://arxiv.org/abs/2409.12822.