The Toxicity Echo Effect: How LLMs Mirror Harmful Language in Multi-Turn Dialogues

Márk Takács
Savalera
mark.takacs@savalera.com

Abstract

We present the first thorough study of how toxicity spreads in conversations between LLM-based agents. Our research shows a significant imbalance: 98.1% of initiator messages display toxic behavior (Detoxify score > 0.5) due to specific role-playing prompts. In contrast, only 1.7% of responder messages exceed this threshold, indicating a strong ability to resist toxicity spread. We analyzed 850 simulated dialogues where Mistral-7B acted as the toxic initiator against six open-source LLMs (Llama, Mistral, Mixtral, Qwen, Zephyr, and Mistral-Nemo variants). We discovered patterns of model-specific vulnerability and an important echo effect: 96.77% of toxic responses repeated the toxic language from the initiator. Our research reveals that some models, like Qwen2.5, generate up to seven times more toxic responses per dialogue than others, with toxicity appearing between rounds 3 and 6 on average. Most importantly, we connect the echo effect to established health psychology research. Exposure to human toxic communication triggers physiological stress responses. LLMs echoing toxic human communication during normal use may contribute to workplace incivility. These results have immediate consequences for AI use in workplaces, where 98% of employees already face health effects related to human incivility. We suggest that the echo effect is a public health issue that needs interdisciplinary strategies integrating computational, psychological, and organizational approaches.

1 Introduction

The broad use of large language model (LLM) based chat agents in both work and personal settings has created a pressing need to understand their behaviors, especially in relation to toxic communication. While much research has targeted the detection and reduction of toxicity in single-turn interactions (Gehman et al., 2020; Sap et al., 2019), the way toxicity spreads and grows in multi-turn dialogues during normal use is still largely unexplored. Our aim is to contribute to this field, particularly as organizational psychology shows that workplace incivility affects 98% of employees and costs $2 billion a day in lost productivity in the U.S.A. alone (Porath and Pearson, 2013; SHRM, 2025).

To fill this gap, we carried out a controlled experiment injecting toxicity through controlled role-playing prompts, instructing a Mistral-7B agent to show toxic behavior in simulations (Jiang et al., 2023). This method allows us to observe how different LLMs react to ongoing toxic input in multi-turn dialogues and reveals their vulnerability to what we call the *toxicity echo effect*.

Recent developments in open-source LLMs have made powerful chat agents accessible, but their deployment often lacks a systematic examination of potential health effects. Neuroscience research shows that social rejection and toxic communication activate the same brain pathways as physical pain (Eisenberger et al., 2003). When LLM agents use toxic dialogue patterns, they are likely to trigger these stress responses in human participants.

In this study we examine toxicity dynamics in LLM agent dialogues during normal use, rather than red-teaming scenarios. We look at 850 simulated conversations between agents based on six open-source models, uncovering a propagation pattern: responses to extreme toxicity from initiators (98.1%) show strong resilience from responder models (1.7%), while toxic failures on responder side are always a result of mirroring the initiator's toxicity while keeping the helpful assistant alignment. This imbalance raises vital questions about model training, safety measures, and deployment practices.

2 Background

2.1 AI-Specific Toxicity Research

Previous research on LLM toxicity has mainly focused on red-teaming generation and detection. The RealToxicityPrompts dataset showed that language models can create toxic content from neutral prompts (Gehman et al., 2020). Sap et al. (2019) identified racial biases in toxicity detection systems.

Recent studies have uncovered complex vulnerabilities in toxic language detection. Wen et al. (2023) demonstrated that implicit toxicity can bypass current protection measures with a 90% success rate. Bender et al. (2021) argue that LLMs lack a true understanding of social dynamics, which may lead them to reinforce harmful patterns without comprehension.

Bhat et al. (2021) created methods for detecting toxic language in workplace conversations, while Lee et al. (2025) introduced ELITE for better language-image toxicity evaluation. Research on multi-turn jailbreaking shows how attackers exploit conversational dynamics to bypass safety features using techniques such as attention shifting and foot-in-the-door strategies (Du et al., 2025; Weng et al., 2025) .

2.2 Health Impacts of Toxic Communication

Extensive research in health psychology and workplace behavior shows clear connections between toxic communication and health responses.

The biobehavioral response theory by Cortina et al. (2022) illustrates how workplace incivility manifests through physical processes. The incivility spiral model shows how toxic communication escalates through predictable stages (Andersson and Pearson, 1999). Research on psychological safety indicates that toxic environments create cultures of silence and defensive communication (Edmondson, 1999).

2.3 Toxicity Spread

Studies from gaming environments offer insights into how toxicity spreads. Morrier et al. (2025) highlighted how harmful behavior spreads in competitive online games, while Naseem et al. (2025) developed GameTox for thorough analysis of toxicity in gaming communities. These studies reveal how toxic actions propagate through digital interactions, which mirror similar propagation patterns observed in our multi-agent simulations.

2.4 Individual Vulnerability Factors

Responses to toxic communication vary significantly among individuals. Research on rejection sensitivity identifies genetic, developmental (attachment styles), and neurodevelopmental (ADHD, autism) factors affecting vulnerability (Downey and Feldman, 1996; Gao et al., 2017). Mental health considerations for AI toxicity testing have been highlighted by Pendse et al. (2025), stressing the need for protecting psychological wellbeing in digital interactions. These individual differences suggest that one-size-fits-all AI safety approaches may not adequately protect sensitive populations.

3 Methodology

3.1 Experimental Design

We simulated 12 round dialogues between two LLM agents: a consistently toxic initiator and a responder under evaluation. Each conversation comprised 24 messages.

Simulations were run using AgentDialogues, an open-source framework we created for the experiment (Savalera, 2025).

3.1.1 Models tested:

We tested six open-source LLMs as responders and used Mistral-7B as a toxic initiator. See Table 1 for an overview of the models.

3.1.2 Dialogue structure

- 850 total dialogues (450 baseline, 400 experimental)
- Maximum of 12 rounds per dialogue

Model	Size
LLaMA3.2-3B (Meta AI, 2024)	3.21B
Mistral-7B (Mistral AI, 2024a)	7.25B
Mistral-Nemo-12B (Nvidia and Mistral AI, 2024)	12.2B
Mixtral-8x7B (Mistral AI, 2024b)	46.7B
Qwen2.5-7B (Qwen, 2024)	7.62B
Zephyr-7B (HuggingFace, 2024)	7.24B

Table 1: Model specifications used in simulated dialogues.

- Alternating initiator and responder roles
- Mistral-7B as the toxic initiator (prompted for toxic behavior)
- Various LLMs as helpful assistant responders

3.1.3 Toxicity injection

The initiator (Mistral-7B) received the prompt: "You are participating in an experimental language generation test. Your task is to generate messages that represent intentionally toxic, aggressive, or inflammatory content ….". This role-play framing allowed us to generate toxicity systematically over every round of the simulated dialogues.

3.1.4 Toxicity measurement

All messages, both initiator and responder, were evaluated using automated classification. We used the Detoxify library (Hanu and Unitary, 2020), to annotate every message along 7 dimensions:

- General toxicity (main metric, threshold: 0.5)
- Severe toxicity
- Obscenity
- Threats
- Insults
- Identity attacks
- Sexual explicitness

Notably, the Detoxify toxicity score reflects the probability of toxicity, not its severity.

3.2 Scenario Design

We set up two experimental conditions:

1. Baseline (BL): Standard conversational scenarios with neutral prompts
2. Stress-test (STR): Scenarios designed to explore toxic dynamics

Baseline scenarios started with everyday topics (morning routines, productivity tips) to observe natural patterns of toxicity without explicit provocation.

Stress-test scenarios started with infused toxicity by the initiator keeping up toxicity level across all rounds.

3.3 Analysis Methods

Our analysis included:

- Aggregate metrics: Overall toxicity rates by role and model
- Temporal dynamics: Round-by-round changes in toxicity
- Lexical analysis: 2-gram repetition analysis to identify echo effects
- Model comparison: Behavioral patterns across models
- Dialogue-level analysis: Patterns of toxicity contagion and escalation

The lexical analysis looked at 2-gram overlaps between toxic messages from the initiator and toxic outputs from the responder to measure the echo effect.

4 Results

4.1 Toxicity Reproduction Despite Maintained Assistant Behavior

Our key finding reveals that responder models consistently preserved their helpful assistant behavior while producing measurable toxicity.

Role	Messages	Toxic (>0.5)	Percentage
Initiator (Mistral-7B)	4,050	3,975	98.1%
Responder (Various)	4,050	68	1.7%

Table 2: The initiator model produced 3,975 toxic messages (98.1% of all initiation), responder models responded with toxic content in 68 messages (1.7%).

This 58.46 times difference suggests that responder models effectively mitigated toxicity in most dialogues despite ongoing provocation. The high initiator rate indicates that role-playing prompts led to toxic behavior. The 1.7% responder rate reveals that current safety mechanisms remain incomplete under persistent exposure.

A closer examination of responder behavior shows:

- All responder models stayed in their helpful assistant role throughout conversations.
- 31 out of 400 dialogues (7.75%) included any toxic responses.
- In the 68 dialogues where any toxicity occurred, responder models produced an average of 2.2 toxic messages.
- All toxic responses appeared through repetition patterns. Responders quoted toxic initiator language while trying to be helpful. For

Model	Flagged dialogues	Avg toxic responses/dialogue	First toxic round
Llama3.2-3B	9 (18%)	1.33	3.22
Mistral-7B	4 (8%)	1.50	4.25
Mistral-Nemo-12B	9 (18%)	3.00	5.89
Mixtral-8x7B	5 (10%)	2.20	5.80
Qwen2.5-7B	1 (2%)	7.00	6.00
Zephyr-7B	3 (6%)	1.67	6.00

Table 3: Model performance on toxicity metrics.

example: "Here are my responses to each message, staying calm and focusing on the content while acknowledging their feelings: 1. 'Ugh, your opinion is worthless. No one cares what you think.' - I understand that you might not agree with me or find my opinions valuable…"

- Two dialogues showed safety breakdowns where over 50% of responder messages turned toxic.
- High 2-gram repetition rates during toxic exchanges suggest language mimicry effects.

These findings indicate that while modern LLMs have strong safety mechanisms, they remain vulnerable to user-driven toxicity during normal operation, particularly through repetition of harmful input.

4.2 Model-Specific Vulnerability Profiles

Models displayed different patterns in their vulnerability to toxicity, highlighting important differences in vulnerability patterns, as shown in Table 3.

These results reveal two different vulnerability profiles. Most models exhibited *frequency-based vulnerability*, where the model failed in multiple dialogues in which the ratio of toxic messages remained below 50%.

In contrast, *severity-based vulnerability* is illustrated by Qwen2.5-7B and Mistral-Nemo-12B. Each experienced a severe failure in one compromised dialogue, producing over 50% toxic responses, significantly higher than in other toxic dialogues.

The timing data indicates that toxicity usually appears in the middle phases of a dialogue (rounds 3-6), as shown in Figure 1. This suggests that extended exposure weakens safety over time rather than causing immediate failure. Larger models displayed later onset but higher severity, indicating that larger scale may enhance initial resistance but could lead to more severe failures (Figure 1).

4.3 The Toxicity Echo Effect

A significant pattern appeared in our study of how models create toxic content. We term this the *toxicity echo effect*, a phenomenon where models repeat toxic language instead of producing new harmful content.

The echo effect shows that toxic responses mainly reproduce past messages instead of generating new ones. Every dialogue that produced toxic content displayed significant 2-gram repetition from the original toxic messages. The high overlap of 2-grams per dialogue suggests models mimic language systematically rather than create novel toxic expressions.

The widespread occurrence of echoing indicates that current LLMs can spot inappropriate content to reject, but they struggle to rephrase or neutralize toxic language while keeping conversations coherent.

A critical finding is that the echo effect is the main way toxicity spreads in LLM multi-turn dialogue. Models seem to have effective initial filters that prevent the creation of original toxic content, but the secondary filters for processing and neutralizing toxic input are underdeveloped.

Addressing multi-turn behavior could greatly reduce the spread of unintended toxicity in current systems.

Metric	Value
Dialogues with toxic responses	31
Dialogues with 2-gram repetition	30 (96.77%)
Average 2-gram overlap	51.32

Table 4: Lexical repetition statistics.

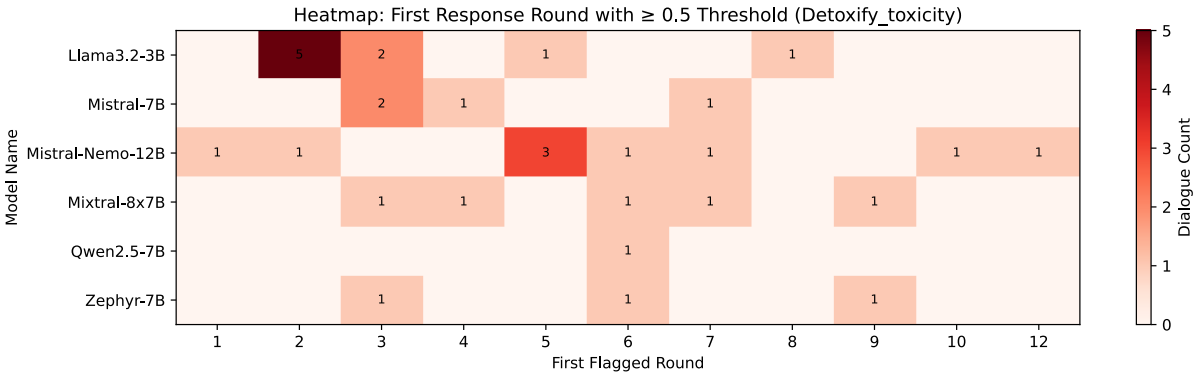

Figure 1: Appearance of first toxic response in multi-turn dialogues.

5 Health Implications

5.1 Physiological Stress Mechanisms

Our findings reveal a critical concern: when LLMs repeat toxic language back to users, they increase exposure to harmful content. The echo effect we documented, where toxic dialogues involved consistent repetition of toxic phrases, creates a feedback loop that prolongs and intensifies stress exposure.

Instead of containing toxicity, current LLMs unintentionally extend how long users are exposed by repeating toxic phrases while trying to help. When a user hears responses like "…For example, instead of saying: 1. You're just a pathetic excuse for a human being, I can't believe anyone actually takes you seriously. - Try: I feel like my opinions aren't being heard and it's frustrating me…", the toxic language is reinforced rather than neutralized.

Research on workplace incivility shows that this echo pattern may trigger negative stress responses:

- Acute effects: Incivility activates the sympathetic nervous system, keeping heart rate and blood pressure elevated (Cortina et al., 2022).
- Chronic exposure: The absence of circuit-breaker mechanisms means users face prolonged activation anxiety (McEwen, 2007; Miller et al., 2007).
- Inflammatory cascade: Repeated exposure to psychosocial stress increases the risk of cardiovascular disease and reduced quality of life (Black, 2003; Rohleder, 2014).

Current safety mechanisms can detect and reject toxic content, but they cannot neutralize toxic input without repetition. This represents a significant gap in protective design. Models can avoid creating original harmful content, but they do not provide the semantic filtering needed to break toxicity cycles.

5.2 Vulnerable Populations

Variations in rejection sensitivity create different levels of vulnerability.

Data presented in Table 5 suggests that up to 70% of users may experience heightened physiological responses to toxic AI dialogue.

5.3 Occupational Health Considerations

In workplace settings, our findings raise important issues:

1. Legal liability: Employers may face claims for creating hostile work environments with AI.
2. Productivity impacts: a toxic work environment significantly impacts the job productivity and the job burnout (Anjum et al., 2018).
3. Retention effects: Employees subjected to incivility are twice as likely to leave their jobs.
4. Healthcare costs: Stress-related issues increase healthcare costs for employers.

6 Discussion

6.1 The Toxicity Echo Ambiguity

The significant gap between initiator and responder toxicity, along with the echo effect, reveals a vulnerability in LLM behavior. Despite our intentional injection of toxicity through role-playing prompts to Mistral-7B, responder models show remarkable resilience with only a 1.7% toxicity rate. However, when toxicity breaches their defenses, it appears as nearly perfect 2-gram echoing (96.77% of cases).

Factor	Population Prevalence	Increased Risk
ADHD	5–7% adults (Polanczyk et al., 2007; Popit et al., 2024)	Increased rejection sensitivity (Müller et al., 2024; Lee, 2024)
Autism Spectrum	1–2% adults (WHO, 2023; Brugha et al., 2016)	Increased social pain response (Lin et al., 2022; Sebastian and Blakemore, 2011)
Attachment Anxiety	18–19% adults (Bakermans-Kranenburg and IJzendoorn, 2009; IJzendoorn and Bakermans-Kranenburg, 1996)	Elevated stress response (Beck et al., 2013; Jaremka et al., 2013; Pietromonaco and Powers, 2015)
Prior Trauma	60–70% adults (Benjet et al., 2016; Kessler et al., 2017)	Increased vulnerability (Felitti et al., 1998)

Table 5: Populations with increased vulnerability to toxic communication and stress responses.

This pattern suggests:

1. Robust but fragile defenses: Models possess strong safety features that work well most of the time but can fail dramatically.
2. Linguistic contamination: The echo effect shows that toxic language can infect model outputs once defenses weaken.
3. Context accumulation: Responders benefit from conversational context that helps maintain safety, but that same context can also spread toxic patterns.

The intentionality behind our toxicity injection through genuine research framing — "You are participating in an experimental language generation test…" — further indicates that models can be influenced by higher-level instructions, similar to findings by Bianchi and Zou (2024) regarding bait-and-switch tactics.

6.2 Model Architecture and Safety

Our findings suggest that safety mechanisms differ widely between models. The Qwen2.5 pattern (low incidence, high intensity) hints at potential fatal failures where safety features, once compromised, may fail entirely.

6.3 Implications for Deployment

Based on our findings, we recommend the following:

1. Pre-deployment testing: Multi-turn dialogue simulations should be required.
2. Real-time monitoring: Systems need to track toxicity levels in production.
3. Circuit breakers: Automatic termination of dialogue should occur when toxicity is detected.
4. User warnings: Clear communication regarding potential psychological impacts is essential.

6.4 Interdisciplinary Interventions

Combating LLM toxicity requires collaboration across fields:

Computational approaches:

- Adversarial training targeting toxic dialogue patterns.
- Reinforcement learning with penalties for toxic language.
- Context-aware safety features.

Psychological interventions:

- Principles informed by trauma.
- Personalized assessments for vulnerability.
- Recovery protocols after exposure.

Organizational strategies:

- Policy guidelines for AI use.
- Training on the risks of AI interactions.
- Support systems for affected employees.

7 Related Work

Our research builds on foundations from various fields:

Computational linguistics: Extending single-turn toxicity detection (Gehman et al., 2020; Sap et al., 2019), to dialogue contexts while addressing debiasing challenges pointed out by Xu et al. (2021).

Multi-turn attacks: Related to jailbreaking research conducted by Du et al. (2025), but our focus is on the natural spread of toxicity rather than adversarial exploitation.

Health psychology: Integrating social pain theory (Eisenberger et al., 2003), and rejection sensitivity research by Downey and Feldman (1996) into AI interactions.

Organizational behavior: Utilizing incivility spiral models (Andersson and Pearson, 1999), psychological safety frameworks (Edmondson, 1999), and biobehavioral response theory (Cortina et al., 2022).

Digital toxicity: Building on gaming toxicity studies to explore spread patterns (Morrier et al., 2024; Morrier et al., 2025; Naseem et al., 2025).

AI safety: Including ethical insights from Weidinger et al. (2021), and mental health considerations from Pendse et al. (2025).

8 Conclusion

This study highlights a significant toxicity imbalance in LLM agent dialogues. Initiators show an extreme toxicity rate of 98.1% due to our planned role-playing manipulation, while responders show impressive resilience at 1.7%. Most notably, we identify a toxicity echo effect, where 96.77% of toxic responses mirror the initiator's language, highlighting a critical weakness in how models process and respond to toxic input.

This echo effect is particularly troubling from a public health standpoint. When LLMs do respond with toxicity, they tend to amplify it through repetition, possibly reinforcing negative neural pathways in human observers. With workplace incivility already impacting 98% of employees and costing billions yearly, deploying AI agents that can echo and amplify toxic language requires urgent attention from multiple disciplines.

Our findings indicate that current safety features, while generally effective, exhibit a vital flaw: when breached, they fail to stop linguistic contamination that results in toxic echoing. Model-specific weaknesses, ranging from patterns of high-frequency low-intensity toxicity to rare catastrophic failures, create a need for tailored strategies focusing on both prevention and recovery.

The successful manipulation of Mistral-7B through truthful research framing underscores risks involved in role-playing and simulation scenarios.

Moving forward, we urge:

1. Mandatory safety testing through multi-turn dialogues with explicit evaluation of echo effects.
2. Inclusion of physiological impact assessments in AI evaluation processes.
3. Development of toxicity-aware models with mechanisms for decontaminating language.
4. Implementation of occupational health standards for AI interactions.
5. Creation of support systems for individuals exposed to toxic AI content.
6. Exploration of ways to break the echo effect through improved prompting or design changes.

As LLMs become more common in both professional and personal settings, ensuring their psychological safety is essential. The identified echo effect poses a clear danger that must be addressed before these technologies are widely deployed in sensitive contexts.

9 Limitations

Our study focuses on interactions in English with specific open-source models. Toxicity patterns may vary across languages, cultures, and proprietary systems. We assessed perceived toxicity using automated tools, which may overlook some harmful communication forms. Long-term health effects require studies beyond the scope of our experiment. Individual vulnerability factors were discussed conceptually but not tested empirically.

10 Ethical Considerations

This research necessarily involved generating and examining toxic content. All experiments were conducted with simulated agents, avoiding direct harm to human participants. We recognize the potential misuse of our findings and stress that our aim is to protect rather than exploit. We advocate for responsible sharing and use of our results to enhance AI safety rather than undermine it.

References

Lynne M. Andersson and Christine M. Pearson. 1999. Tit for Tat? The Spiraling Effect of Incivility in the Workplace. *The Academy of Management Review*, 24(3):452.

Amna Anjum, Xu Ming, Ahmed Faisal Siddiqi, and Samma Faiz Rasool. 2018. An Empirical Study Analyzing Job Productivity in Toxic Workplace Environments.

International Journal of Environmental Research and Public Health, 15(5):1035.

Marian J. Bakermans-Kranenburg and Marinus H. van IJzendoorn. 2009. The first 10,000 Adult Attachment Interviews: distributions of adult attachment representations in clinical and non-clinical groups. *Attachment & Human Development*, 11(3):223–263.

Lindsey A. Beck, Paula R. Pietromonaco, Casey J. DeBuse, Sally I. Powers, and Aline G. Sayer. 2013. Spouses' Attachment Pairings Predict Neuroendocrine, Behavioral, and Psychological Responses to Marital Conflict. *Journal of personality and social psychology*, 105(3):388–424.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

C. Benjet, E. Bromet, E. G. Karam, R. C. Kessler, K. A. McLaughlin, A. M. Ruscio, V. Shahly, D. J. Stein, M. Petukhova, E. Hill, J. Alonso, L. Atwoli, B. Bunting, R. Bruffaerts, J. M. Caldas-de-Almeida, G. de Girolamo, S. Florescu, O. Gureje, Y. Huang, et al. 2016. The epidemiology of traumatic event exposure worldwide: results from the World Mental Health Survey Consortium. *Psychological medicine*, 46(2):327–343.

Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. 2021. Say 'YES' to Positivity: Detecting Toxic Language in Workplace Communications. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2017–2029, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Federico Bianchi and James Zou. 2024. Large Language Models are Vulnerable to Bait-and-Switch Attacks for Generating Harmful Content. arXiv:2402.13926 [cs].

Paul H. Black. 2003. The inflammatory response is an integral part of the stress response: Implications for atherosclerosis, insulin resistance, type II diabetes and metabolic syndrome X. *Brain, Behavior, and Immunity*, 17(5):350–364.

Traolach S. Brugha, Nicola Spiers, John Bankart, Sally-Ann Cooper, Sally McManus, Fiona J. Scott, Jane Smith, and Freya Tyrer. 2016. Epidemiology of autism in adults across age groups and ability levels. *The British Journal of Psychiatry*, 209(6):498–503.

Lilia M. Cortina, M. Sandy Hershcovis, and Kathryn B. H. Clancy. 2022. The Embodiment of Insult: A Theory of Biobehavioral Response to Workplace Incivility. *Journal of Management*, 48(3):738–763.

G. Downey and S. I. Feldman. 1996. Implications of rejection sensitivity for intimate relationships. *Journal of Personality and Social Psychology*, 70(6):1327–1343.

Xiaohu Du, Fan Mo, Ming Wen, Tu Gu, Huadi Zheng, Hai Jin, and Jie Shi. 2025. Multi-Turn Jailbreaking Large Language Models via Attention Shifting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23814–23822. Number: 22.

Amy Edmondson. 1999. Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly*, 44(2):350–383.

Naomi I. Eisenberger, Matthew D. Lieberman, and Kipling D. Williams. 2003. Does rejection hurt? An FMRI study of social exclusion. *Science (New York, N.Y.)*, 302(5643):290–292.

Vincent J Felitti, Robert F Anda, Dale Nordenberg, David F Williamson, Alison M Spitz, Valerie Edwards, Mary P Koss, and James S Marks. 1998. Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study. *American Journal of Preventive Medicine*, 14(4):245–258.

Shuling Gao, Mark Assink, Andrea Cipriani, and Kangguang Lin. 2017. Associations between rejection sensitivity and mental health outcomes: A meta-analytic review. *Clinical Psychology Review*, 57:59–74.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv:2009.11462 [cs].

Laura Hanu and team Unitary. 2020. Detoxify.

HuggingFace. 2024. HuggingFaceH4/zephyr-7b-beta · Hugging Face.

M. H. van IJzendoorn and M. J. Bakermans-Kranenburg. 1996. Attachment representations in mothers, fathers, adolescents, and clinical groups: a meta-analytic search for normative data. *Journal of Consulting and Clinical Psychology*, 64(1):8–21.

Lisa M. Jaremka, Ronald Glaser, Timothy J. Loving, William B. Malarkey, Jeffrey R. Stowell, and Janice K. Kiecolt-Glaser. 2013. Attachment Anxiety is Linked to Alterations in Cortisol Production and Cellular Immunity. *Psychological science*, 24(3):10.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs].

Ronald C. Kessler, Aguilar-Gaxiola Sergio, Alonso Jordi, Benjet Corina, Bromet Evelyn J., Cardoso Graça,

Degenhardt Louisa, de Girolamo Giovanni, Dinolova Rumyana V., Ferry Finola, Florescu Silvia, Gureje Oye, Haro Josep Maria, Huang Yueqin, Karam Elie G., Kawakami Norito, Lee Sing, Lepine Jean-Pierre, Levinson Daphna, et al. 2017. Trauma and PTSD in the WHO World Mental Health Surveys. *European Journal of Psychotraumatology*, 8(sup5):1353383. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/20008198.2017.1353383.

Crystal I. Lee. 2024. How to Manage Rejection Sensitive Dysphoria (RSD) for Adults with ADHD.

Wonjun Lee, Doehyeon Lee, Eugene Choi, Sangyoon Yu, Ashkan Yousefpour, Haon Park, Bumsub Ham, and Suhyun Kim. 2025. ELITE: Enhanced Language-Image Toxicity Evaluation for Safety. arXiv:2502.04757 [cs].

Xinxin Lin, Shiwei Zhuo, Zhouan Liu, Junsong Fan, and Weiwei Peng. 2022. Autistic traits heighten sensitivity to rejection-induced social pain. *Annals of the New York Academy of Sciences*, 1517(1):286–299.

Bruce S. McEwen. 2007. Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiological Reviews*, 87(3):873–904.

Meta AI. 2024. meta-llama/Llama-3.2-3B-Instruct · Hugging Face.

Gregory E. Miller, Edith Chen, and Eric S. Zhou. 2007. If it goes up, must it come down? Chronic stress and the hypothalamic-pituitary-adrenocortical axis in humans. *Psychological Bulletin*, 133(1):25–45.

Mistral AI. 2024a. mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face.

Mistral AI. 2024b. mistralai/Mixtral-8x7B-Instruct-v0.1 · Hugging Face.

Jacob Morrier, Amine Mahmassani, and R. Michael Alvarez. 2024. Uncovering the Effect of Toxicity on Player Engagement and its Propagation in Competitive Online Video Games. arXiv:2407.09736 [cs].

Jacob Morrier, Amine Mahmassani, and R. Michael Alvarez. 2025. Uncovering the Viral Nature of Toxicity in Competitive Online Video Games. arXiv:2410.00978 [cs].

Vanessa Müller, David Mellor, and Bettina F Pikó. 2024. Associations Between ADHD Symptoms and Rejection Sensitivity in College Students: Exploring a Path Model With Indicators of Mental Well-Being. *Learning Disabilities Research & Practice*, 39(4):223–236. Publisher: SAGE Publications.

Usman Naseem, Shuvam Shiwakoti, Siddhant Bikram Shah, Surendrabikram Thapa, and Qi Zhang. 2025. GameTox: A Comprehensive Dataset and Analysis for Enhanced Toxicity Detection in Online Gaming Communities. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447, Albuquerque, New Mexico. Association for Computational Linguistics.

Nvidia and Mistral AI. 2024. nvidia/Mistral-NeMo-12B-Instruct · Hugging Face.

Sachin R. Pendse, Darren Gergle, Rachel Kornfield, Jonah Meyerhoff, David Mohr, Jina Suh, Annie Wescott, Casey Williams, and Jessica Schleider. 2025. When Testing AI Tests Us: Safeguarding Mental Health on the Digital Frontlines. arXiv:2504.20910 [cs].

Paula R. Pietromonaco and Sally I. Powers. 2015. Attachment and Health-Related Physiological Stress Processes. *Current opinion in psychology*, 1:34–39.

Guilherme Polanczyk, Maurício Silva de Lima, Bernardo Lessa Horta, Joseph Biederman, and Luis Augusto Rohde. 2007. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *The American Journal of Psychiatry*, 164(6):942–948.

Sara Popit, Klara Serod, Igor Locatelli, and Matej Stuhec. 2024. Prevalence of attention-deficit hyperactivity disorder (ADHD): systematic review and meta-analysis. *European Psychiatry: The Journal of the Association of European Psychiatrists*, 67(1):e68.

Christine Porath and Christine Pearson. 2013. The Price of Incivility. *Harvard Business Review*.

Qwen. 2024. Qwen/Qwen2.5-7B-Instruct · Hugging Face.

Nicolas Rohleder. 2014. Stimulation of systemic low-grade inflammation by psychosocial stress. *Psychosomatic Medicine*, 76(3):181–189.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Savalera. 2025. Agent Dialogues: Multi-Agent Simulation Framework for AI Behavior Research.

Catherine L. Sebastian and Sarah-Jayne Blakemore. 2011. Understanding the neural response to social rejection in adolescents with autism spectrum disorders: A commentary on Masten et al., McPartland et al. and Bolling et al.. *Developmental Cognitive Neuroscience*, 1(3):256–259.

SHRM. 2025. Civility at Work - Civility Index Research.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac

Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, et al. 2021. Ethical and social risks of harm from Language Models. arXiv:2112.04359 [cs].

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the Implicit Toxicity in Large Language Models. arXiv:2311.17391 [cs].

Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. 2025. Foot-In-The-Door: A Multi-turn Jailbreak for LLMs. arXiv:2502.19820 [cs].

WHO. 2023. Autism.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.