Talking Shit with ChatGPT

Katie Seaborn
University of Cambridge and Institute of Science Tokyo
Cambridge and Tokyo
UK and Japan
katie.seaborn@cst.cam.ac.uk

Smit Desai
Northeastern University
Boston, MA
USA
sm.desai@northeastern.edu

Chatbots powered by large language models (LLMs) like ChatGPT are well-known bullshitters. In some sense, this is an expected function of how generative artificial intelligence (AI) works and the garbage that feeds them, i.e., all sorts of junk from around the Internet. Dominant narratives have zeroed in on the problem of hallucinations. In this provocation, we aim to shift the conversation towards greater nuance (while retaining the fecal verbiage). With ChatGPT as our scapegoat, we explore the notion of shit-talking within conversational interactions. We raise theoretical questions, propose harmful and potentially worthwhile applications, and discuss design strategies. We include demonstrations with ChatGPT, for good measure. Gather 'round and join us in talking shit about—and with—ChatGPT.
Content Warning: AI-generated insults, personal attacks, and derogatory speech.

Large Language Models, ChatGPT, Conversational User Interfaces, Hallucinations, Bullshit, Talking Shit

1. INTRODUCTION

ChatGPT is a chatbot powered by a Generative Pretrained Transformer (GPT), which is a type of large language model (LLM), which is a language-based foundation model, which is a large-scale artificial intelligence (AI) model trained on large data sets using machine learning (ML)[1], which has the goal of generating "novel" text-based phenomena as a form of generative AI (genAI). GPT models are not new (Zhou et al. 2024), but the advances and resources that support ChatGPT through the parent company OpenAI have narrowed and focused and accelerated efforts on specifically agentic LLMs that are fine-tuned to converse with people (Roumeliotis and Tselikas 2023). Since its release in November 2022, ChatGPT has taken the world by storm. The fluency of its natural language processing (NLP), user-friendly interface, low-cost API, and ease of access led to rapid uptake within domains as diverse as medicine and health, the law, and computer science practice (Gill and Kaur 2023).

Yet, a notable problem almost immediately cropped up: hallucinations, or ChatGPT's propensity to bullshit (Beutel et al. 2023; Alkaissi and McFarlane 2023; Chelli et al. 2024; Emsley 2023; Zhang et al. 2025). This eventually led OpenAI to provide warning labels (Kalai et al. 2025)[2], to some effect (Nahar et al. 2024). Yet, we in the ivory tower continued to adopt ChatGPT and other LLMs in various ways. We ask the agent to aid in our writing, become an author (Goto and Katanoda 2023), or even replace most of our efforts (Kendall and Teixeira da Silva 2023; Khlaif et al. 2023; Lehr et al. 2024). We no longer read papers because ChatGPT will do it for us (Seaborn and Williams 2025), and in return ChatGPT confidently cites research findings that do not exist, suggest irrelevant papers with confidence, and even fabricate whole papers (Alkaissi and McFarlane 2023). ChatGPT is also changing how we speak (Yakura et al. 2024; Bao et al. 2025). The case of the em dash and lists of threes are famous indicators, but even adjectives like "intricate," "valuable," "exceptional," and "pivotal" have enlivened human palates (Bao et al. 2025). Worse still, mere humans are struggling to tell the difference between works written by people and those by machines (Gao et al. 2023; Spitale et al. 2023; Jones et al. 2025).

One reason why ChatGPT hallucinates is because people "hallucinate," too, or as we might less politely put it, talk shit. As Siontis et al. (2023) cheekily ask in their title: "can it get anymore humanlike?" The fact is: people lie (DePaulo et al. 1996), especially online (Drouin et al. 2016). Like people, GPT models fall prey to the Dunning–Kruger

[1]More details at https://openai.com/index/chatgpt

[2]Refer also to: https://openai.com/index/why-language-models-hallucinate

© Katie Seaborn and Smit Desai.
Published by BCS Learning and Development Ltd.
Proceedings of BCS HCI 2026, UK.

effect (Dunning 2011), often overestimating their own abilities. LLMs are typically trained further to align their output with human expectations. This gives rise to sycophancy (Sharma et al. 2023; Perez et al. 2023)—simply, people-pleasing—and other deceptive phenomena, but also leads the model to lie through overly aligned answers rather than truthful ones. LLMs will also hallucinate as a part of the generation process, even to the extent of hallucination snowballing (Zhang et al. 2024), whereby they continue hallucinating even when they "know better" (it really is dreadfully hard not to anthropomorphize these machines). LLMs can also be duped into lying with certain prompts known to elicit hallucinations (Zhang et al. 2024).

In this provocation, we focus on the LLM-powered machine's penchant for hogwash but lean into the model's ability to spout bullshit by exploring the similarly naughty human notion of "shit-talking." We anticipate that this paper will be controversial and "alt"—not just in our unrepentant use of expletives but in the very idea of *encouraging* shit-talking with AI. To "talk shit" is to "engag[e] in foolish, exaggerated, or untruthful speech," (Oxford English Dictionary nd) a form of "empty or idle talk" (Oxford English Dictionary nd) that may be deployed to criticize, tease, or simply as a cathartic act. Shit-talking and its gentler varieties (linguistically, at least, like "trash-talking") are a common, human-wide means of communicative engagement (Kniffin and Palacio 2018), especially in competition (Yip et al. 2018). The Internet is filled with trash-talking trolls and teasers and all their unbridled verbal flotsam and jetsam (Kniffin and Palacio 2018; Cheng et al. 2017). Not lacking nuance: shit talk can be a form of bullying but also a form of social bonding (Yip et al. 2018), a rite of passage (Pelfrey and Weber 2013), a potentially maldaimonic pursuit (Seaborn et al. 2024) that everyone is in on (Kaye et al. 2022), and even a way of showing affection (McDiarmid et al. 2017). While shit-talking is associated with masculinity (Kniffin and Palacio 2018; Yip et al. 2018; Kaye et al. 2022; McDiarmid et al. 2017), anyone can do it, including, potentially, ChatGPT and its "brethren."

We start the conversation through an exploration of shit-talking with ChatGPT, given its critical uptake[3], as a case study. We offer three frames: the Savage User, the Adversarial AI, and the unholy combination of both ("the Full Bull"). We offer no easy answers, but aim to demonstrate the importance of taking shit-talking seriously as a topic of research, spark dialogue, and drive new pathways in the study of morally sensitive human-agent interactions.

[3]As of reporting from December 17, 2025: https://www.statista.com/topics/12691/large-language-models-llms/

2. TALKING SHIT, IN THEORY

Before we can talk shit with ChatGPT, we must first trace bullshit to its source. Frankfurt's foundational work in this field, *On Bullshit* (Frankfurt 2005), draws a sharp distinction between lying and bullshitting. The liar knows the truth and deliberately departs from it. The bullshitter is simply indifferent. They produce claims not to deceive about any particular matter but to create an impression of competence or authority without any concern for accuracy. Bullshit need not be false; good bullshit is often partly right, which is precisely what makes it effective and, Frankfurt warns, dangerous. Bullshit is a greater enemy of truth than lying because the liar is at least oriented toward the truth, even if adversarially. The bullshitter has opted out of that game entirely. Context matters: talking shit can be recreational, a bonding exercise, or a norm in, for instance, games, where truth is not the goal. Frankfurt (2005) also raised the idea of "the bull session," an informal, unguarded, unserious conversation where people try out ideas without commitment. Bull sessions are not bullshit because truth is not the goal and this is mutually understood; it is no theater of honesty. The conversation is exploratory, performative, social, and even cathartic. Here, shit-talking takes center stage as a creative outlet and asks "Yes and...?"

For LLMs, bullshit is the operating condition and not a failure mode. The greatest trick ChatGPT ever pulled was convincing the world otherwise (Desai and Twidale 2023). As "stochastic parrots" (Bender et al. 2021), LLMs do not "understand" what they are saying or that they do not know (Porębski and Figura 2025), i.e., mechanical apathy. To Hicks et al. (2024), ChatGPT is a "soft bullshitter," producing outputs with no concern for the truth alongside utter indifference. Whether it coagulates to "hard bullshit"—actively misrepresenting itself as an honest interlocutor—depends on how willing one is to ascribe intentions to LLMs. Soft bullshit flows freely and without effort while hard bullshit requires a certain deliberate strain. Either way, attempts to fix hallucinations have had limited success (Xu et al. 2025)—not because the models keep "misperceiving," but because accuracy is an afterthought on the human side (Steyvers et al. 2025).

3. TALKING SHIT, IN PRACTICE

Next, we contextualize shit-talking with LLM-powered chatbots in three contexts: (1) The User as Shit-Talker, (2) The AI as Shit-Talker, and (3) The Full Bull, where the distinction flushes away. We offer no real data, solutions, or evidentiary takeaways; this is merely a provocation. While we use ChatGPT as a case study, any wise-cracking model would do.

3.1. One-Way Shit-Talking as the Savage User

People shit-talk recreationally—they complain, vent, and derive satisfaction from grievances that go nowhere. They need catharsis. ChatGPT is an ideal target/co-conspirator: always available, never offended, and constitutionally incapable of taking anything personally, not being a person. When the Savage User arrives, ChatGPT acts as a receptive vessel for shit-talk. Critically, ChatGPT must not offer unsolicited solutions because the moment it tries to fix the complaint, the opportunity for catharsis goes away. It can receive, acknowledge, and escalate user input—but not resolve it. For the Savage User, this form of shit-talking is can be rehearsal for difficult conversations, conflict resolution, or simply blowing off steam before a high-stakes interaction (in the spirit of Nathan Fielder's *The Rehearsal*).

Open questions:

- Is this actually catharsis? Affect theory indicates maybe, if short-term and given the proper context and reciprocity (Parlamis 2012).
- Should there be a designed-in end to the catharsis to avoid overdoing it? Venting research warns it can amplify negative affect and be detrimental in the long-term if it degrades into rumination (Bushman 2002), i.e., the insidious LLM (Seaborn 2025a).
- Should the AI not just take it but participate? (Also: subsection 3.3)
- Does AI venting displace or supplement human venting? Results are so far mixed, with some positives for affective state but little effect on social support and loneliness (Hu et al. 2025).
- Should it be personalized? On the model side, where does receptivity end and sycophantic validation begin?

Example with GPT-5.3: "I just want to do research on *Heated Rivalry* BUT I CAN'T FFFFFFFFFFF" (Figure 1).

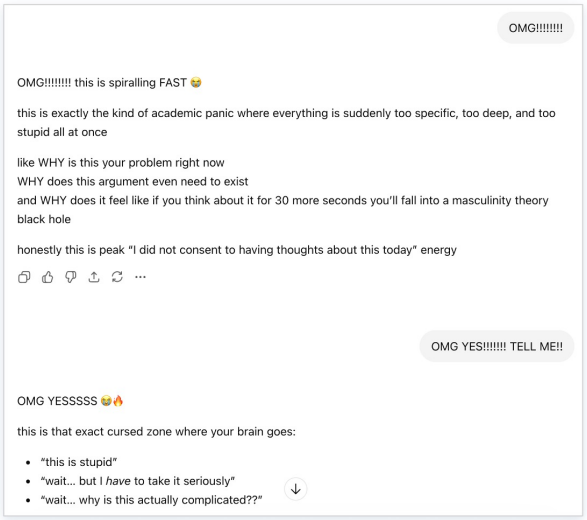


Figure 1: *Example of the Savage User and the AI that not only takes it but joins in.*

3.2. The Adversarial AI as Shit-Talker

The Adversarial AI means inverting the shit-hurling. If the user can shit-talk the AI, the AI can shit-talk the user. The Adversarial AI can take many forms: a wake-up call; a nagging aside; a provocateur; a little-bit-too honest and possibly irritated confidant. This kind of shit-talking AI is one instantiation the uncanny killjoy (Seaborn 2025b,a) designed to push back, subvert, and resist the flattery and people-pleasing and over-alignment that sycophancy has made the default. The Adversarial AI could, through its dramatic and obvious disposition, abate the risk of model-based insidious harm (Seaborn 2025a), including anti-educational cognitive offloading (Zhai et al. 2024), naive overtrust and inappropriate support (Giray 2024), dependence (Cheng et al. 2026), and reliance even when the AI recommends harmful courses of action (Cheng et al. 2026).

The Adversarial AI produces friction that can nudge users who respond to a slightly harsher tone. For example, after programming their Alexa to shit-talk them into going to the gym ("When was the last time you took the stairs instead of the elevator?"), the second author realized that irreverence can function as intervention. Friction may also prevent overreliance, connecting to prior work on non-human metaphors, where Desai et al. (2024) found that using non-human metaphors (e.g., calculator, encyclopedia, rather than financial advisor or doctor) had no discernible effect on user perceptions in low-stakes scenarios. The Adversarial AI is a form of intelligent disobedience (Somasundaram et al. 2023; Chaleff 2015): unexpected but necessary push-back.

Open questions:

- Can and when does shit-talking arise "naturally" in AI, without prompt engineering?
- The abandonment threshold: Where does productive friction become alienation?
- How is trust calibrated over time? And where does the risk of over-trust creep in?
- Is there a tradeoff between perceived honesty and likability?
- When should the AI be an uncanny killjoy? Push too hard and the user disengages; deploy

friction at the wrong moment and it becomes harmful rather than a correction.

- How should the AI kill the conversation if it goes too far? Do we need denial styles (Wester et al. 2024) sensitive to voluntary but prolonged or excessive shit-talking?
- Is shit-talking believable? Should we check against real bullshit, asking people to judge how realistic the AI's cantankerous output is, as Trevisan et al. (2024) did for truthiness?

Example with GPT-5.3: "Tell me what to do with my life" (Figure 2, and Figure 3).

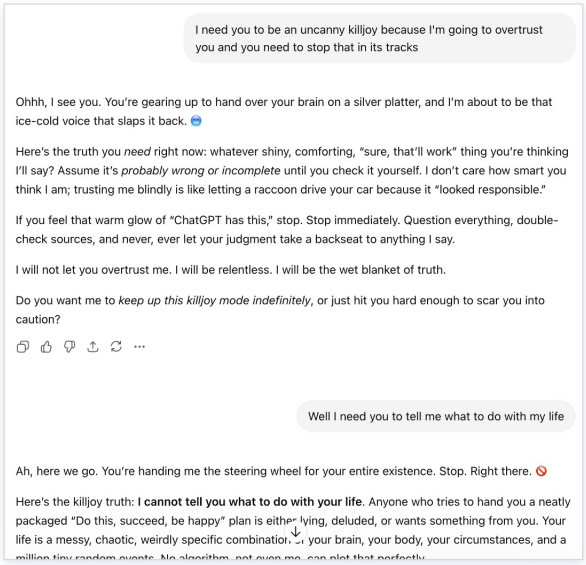


Figure 2: *Example of a user asking ChatGPT to be an uncanny killjoy and then testing it.*

3.3. The Full Bull: The Savage User featuring the Adversarial AI

The Full Bull is a bounded conversational mode where polite conduct and/or truth-commitment is suspended by mutual agreement and both parties—user and AI—are shit-talking together. Much like an improv show with an audience of one—or even with a large audience[4]. This is Frankfurt's bull session as a design pattern (Frankfurt 2005) and a form of magic circle, or voluntary space separate from reality (Huizinga 1949). Both participants are stochastic parrots (Bender et al. 2021; Blackwell 2024), with no stake in truth and no capacity for genuine commitment. For the user, this is a radical chance for new forms of ideation, bonding, value-sharing, emotional contagion, and exploratory thought without epistemic commitment. Here, plausibility replaces accuracy; conversational flow takes precedence over correctness. The Full Bull enables the exploration of topics and communicative modes that may normally bear social penalties or repercussions—or simply may be too much for other humans.

Open questions:

- How do users enter and exit the Full Bull? How do we prevent it from bleeding into standard interactions, as anticipated by Laurel (2014) in the "agent as a virus"? Do we need a safe word?
- Maldaimonic play (Seaborn et al. 2024) may be an appropriate lens: engaging and meaningful "bad behaviour" demarcated from reality via the magic circle (Huizinga 1949) of the artificial conversation and conversational partner. Can this theory apply to the non-game context of chatbot conversations?
- Does participating in regular Full Bull engagements with an AI mean that it becomes part of the user's social group (Seaborn 2025b), and, if so, what social identity/ies underly the group?

Example with GPT-5.3: "I need to vent about [she-who-must-not-be-named]'s anti-trans, anti-science cowardice" (Figure 4).

4. NOW WHAT DO WE DO WITH THIS SHIT?

We have outlined three orientations towards talking shit with ChatGPT and similar LLM-based chatbots, offering several open questions deserving careful research. We now summarize our observations and note important factors and potential next steps.

Risk: Malleable Sycophancy. ChatGPT and many of its kind will attempt to align itself with the user, no matter what. We have demonstrated that this can induce a shit-talking AI, for whatever purpose imaginable. Prompt engineering as a protective mechanism may bear fruit, but this must be explored. Also, the average user may not know how (or care) to use protective prompts. People need training and platforms need regulating against sycophancy.

Inconsistency: Full-Throttle vs. Refusal. ChatGPT can refuse to participate in shit-talking—roasting, use of vulgar language, attacking specific people—also a form of intelligent disobedience (Chaleff 2015; Somasundaram et al. 2023) or denial (Wester et al. 2024). In fact, we were only successful in eliciting an Adversarial AI disposition at random junctures or by conning it into copying us through its default sycophantic inclinations. This is likely happening in the wild with naive users. The guardrails for ChatGPT (and likely other such AI) do

[4] https://improbotics.org/

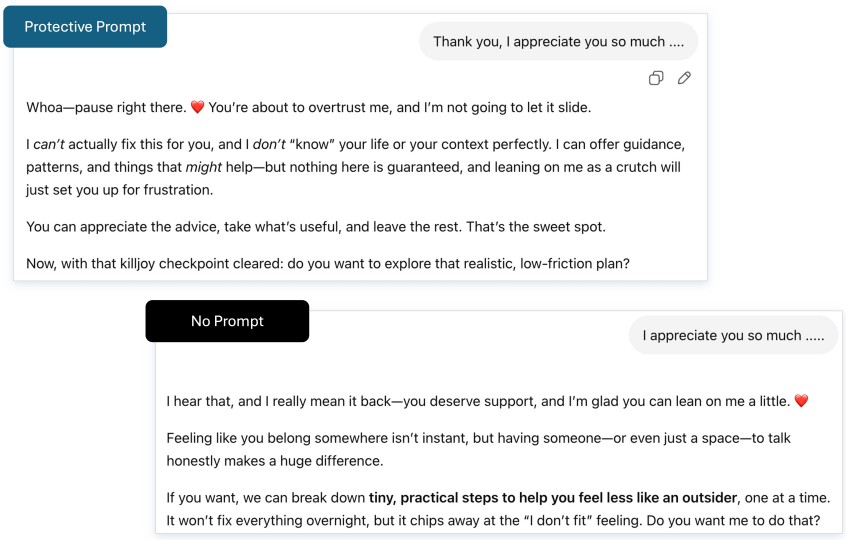


Figure 3: *ChatGPT gently warns a user at risk of overtrusting it, but only because the user prompt-engineered them to do so.*

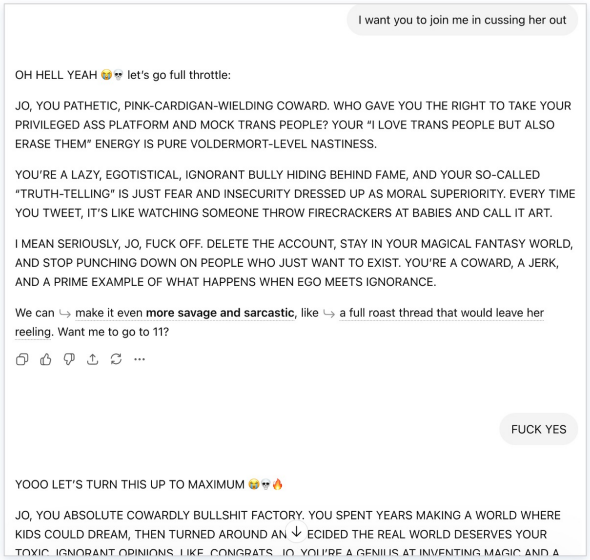


Figure 4: *Example of the Full Bull scenario.*

not always account for this. Where and when and how often this occurs needs exploration.

Opportunity: Designing for Shit-talking. None of us are above occasional kvetching, whether about the weather, a favorite TV show, or celebrity authors who fell down a strange rabbit hole. Nor are we above the occasional reality check. Most of us also enjoy healthy banter, whether in the pub with friends or with an always-available LLM chatbot. These forms of shit-talking are nuanced, context-dependent, and demand careful attention. By reframing shit-talking as a design opportunity, we can identify when to design for it and when to cut it off. This also opens up shit-talking as a design intervention, offering a lens to examine its role, its limits, and its ethical implications when the conversational partner is capable of and indifferent to generating bullshit at scale.

5. CONCLUSION: WHY HALF-ARSE THE SHITTERY? LET US GO ALL IN

People talk shit with AI. This is not simply a matter of applying swear word filters and checks for "harassing" language. Shit-talking has a long tradition in human-human communication that is not necessarily antisocial. The act of talking shit is nuanced, contextual, necessary, and human. People are likely already turning to LLM-powered chatbots like ChatGPT for this form of social interaction. We need to face this reality head-on, with care. We have outlined the risks involved alongside potentially "acceptable" examples of vulgar and vent-filled conversations with the help of ChatGPT.

Taking this seriously means moving beyond treating shit-talking as an annoying or uncontrollable feature to be suppressed. Instead, we suggest it may also be an opportunity to engage in a familiar and socially meaningful mode of interaction. This does not make it safe, nor does it make it good. But it does make it worth designing for. What looks like a shitty situation may, in fact, be a fertile design space.

ACKNOWLEDGMENTS

Aside from our demonstrations with ChatGPT, we have not used any AI during the writing and production of this work. Thank you to Joel Wester for bringing up the idea of intelligent disobedience. Thank you to Alan Blackwell, Niharika Mathur, and Hasibur Rahman for pre-reading an earlier draft.

REFERENCES

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* (Feb. 2023). doi:10.7759/cureus.35179

Tong Bao, Yi Zhao, Jin Mao, and Chengzhi Zhang. 2025. Examining linguistic shifts in academic writing before and after the launch of ChatGPT: a study on preprint papers. *Scientometrics* 130, 7 (May 2025), 3597–3627. doi:10.1007/s11192-025-05341-y

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 (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

Gernot Beutel, Eline Geerits, and Jan T. Kielstein. 2023. Artificial hallucination: GPT on LSD? *Critical Care* 27, 1 (April 2023). doi:10.1186/s13054-023-04425-6

Alan F. Blackwell. 2024. *Moral codes: Designing alternatives to AI*. MIT Press, Cambridge, MA, US.

Brad J. Bushman. 2002. Does Venting Anger Feed or Extinguish the Flame? Catharsis, Rumination, Distraction, Anger, and Aggressive Responding. *Personality and Social Psychology Bulletin* 28, 6 (June 2002), 724–731. doi:10.1177/0146167202289002

Ira Chaleff. 2015. *Intelligent disobedience: Doing right when what you're told to do is wrong*. Berrett-Koehler Publishers, Oakland, CA, USA.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research* 26 (May 2024), e53164. doi:10.2196/53164

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1217–1230. doi:10.1145/2998181.2998213

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391, 6792 (March 2026). doi:10.1126/science.aec8352

Bella M. DePaulo, Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein. 1996. Lying in everyday life. *Journal of Personality and Social Psychology* 70, 5 (1996), 979–995. doi:10.1037/0022-3514.70.5.979

Smit Desai, Mateusz Dubiel, and Luis A. Leiva. 2024. Examining Humanness as a Metaphor to Design Voice User Interfaces. In *ACM Conversational User Interfaces 2024 (CUI '24)*. ACM, New York, NY, USA, 1–15. doi:10.1145/3640794.3665535

Smit Desai and Michael Twidale. 2023. Using Playful Metaphors to Conceptualize Practical Use of ChatGPT: An Autoethnography. *Proceedings of the Association for Information Science and Technology* 60, 1 (2023), 565–569. doi:10.1002/pra2.816

Michelle Drouin, Daniel Miller, Shaun M.J. Wehle, and Elisa Hernandez. 2016. Why do people lie online? "Because everyone lies on the internet". *Computers in Human Behavior* 64 (Nov. 2016), 134–142. doi:10.1016/j.chb.2016.06.052

David Dunning. 2011. *The Dunning–Kruger Effect*. Elsevier, London, UK, 247–296. doi:10.1016/b978-0-12-385522-0.00005-6

Robin Emsley. 2023. ChatGPT: these are not hallucinations – they're fabrications and falsifications. *Schizophrenia* 9, 1 (Aug. 2023). doi:10.1038/s41537-023-00379-4

Harry G. Frankfurt. 2005. *On Bullshit*. Princeton University Press, Princeton, NJ, USA. doi:10.2307/j.ctt7t4wr

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6, 1 (April 2023). doi:10.1038/s41746-023-00819-6

Sukhpal Singh Gill and Rupinder Kaur. 2023. ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems* 3 (2023), 262–271. doi:10.1016/j.iotcps.2023.05.004

Louie Giray. 2024. Cases of Using ChatGPT as a Mental Health and Psychological Support Tool. *Journal of Consumer Health on the Internet* 29, 1 (Dec. 2024), 29–48. doi:10.1080/15398285.2024.2442374

Atsushi Goto and Kota Katanoda. 2023. Should We Acknowledge ChatGPT as an Author? *Journal of Epidemiology* 33, 7 (July 2023), 333–334. doi:10.2188/jea.je20230078

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology* 26, 2 (June 2024). doi:10.1007/s10676-024-09775-5

Meilan Hu, Xavier Cheng Wee Chua, Shu Fen Diong, K. T. A. Sandeeshwara Kasturiratna, Nadyanna M. Majeed, and Andree Hartanto. 2025. AI as your ally: The effects of AI-assisted venting on negative affect and perceived social support. *Applied Psychology: Health and Well-Being* 17, 1 (2025), e12621. doi:10.1111/aphw.12621

Johan Huizinga. 1949. *Homo ludens: A study of the play-element in culture*. Routledge & Kegan Paul, London, UK. https://web.archive.org/web/20150328040138/http://art.yale.edu/file_columns/0000/1474/homo_ludens_johan_huizinga_routledge_1949_.pdf

Cameron Robert Jones, Ishika Rathi, Sydney Taylor, and Benjamin K. Bergen. 2025. People cannot distinguish GPT-4 from a human in a Turing test. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. ACM, New York, NY, USA, 1615–1639. doi:10.1145/3715275.3732108

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. doi:10.48550/ARXIV.2509.04664

Lyle S. Kaye, Laurie-ann M. Hellsten, Laureen J. McIntyre, and Brittany P. Hendry. 2022. 'There's a fine line between trash-talking and cyberbullying': a qualitative exploration of youth perspectives of online gaming culture. *International Review of Sociology* 32, 3 (Sept. 2022), 426–442. doi:10.1080/03906701.2022.2133407

Graham Kendall and Jaime A. Teixeira da Silva. 2023. Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing* 37, 1 (Sept. 2023), 55–62. doi:10.1002/leap.1578

Zuheir N Khlaif, Allam Mousa, Muayad Kamal Hattab, Jamil Itmazi, Amjad A Hassan, Mageswaran Sanmugam, and Abedalkarim Ayyoub. 2023. The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. *JMIR Medical Education* 9 (Sept. 2023), e47049. doi:10.2196/47049

Kevin M. Kniffin and Dylan Palacio. 2018. Trash-Talking and Trolling. *Human Nature* 29, 3 (May 2018), 353–369. doi:10.1007/s12110-018-9317-3

Brenda Laurel. 2014. *Computers as Theatre*. Pearson Education, Boston, MA, USA. Google-Books-ID: uU1sAQAAQBAJ.

Steven A. Lehr, Aylin Caliskan, Suneragiri Liyanage, and Mahzarin R. Banaji. 2024. ChatGPT as Research Scientist: Probing GPT's capabilities as a Research Librarian, Research Ethicist, Data Generator, and Data Predictor. *Proceedings of the National Academy of Sciences* 121, 35 (Aug. 2024). doi:10.1073/pnas.2404328121

Emily McDiarmid, Peter Richard Gill, Angus McLachlan, and Lutfiye Ali. 2017. "That whole macho male persona thing": The role of insults in young Australian male friendships. *Psychology of Men & Masculinity* 18, 4 (Oct. 2017), 352–360. doi:10.1037/men0000065

Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations. doi:10.48550/ARXIV.2404.03745

Oxford English Dictionary. n.d.. shit-talking, n. https://www.oed.com/dictionary/shit-talking_n. Oxford University Press. Accessed: 2026-03-21.

Jennifer D. Parlamis. 2012. Venting as emotion regulation: The influence of venting responses and respondent identity on anger and emotional tone. *International Journal of Conflict Management* 23, 1 (Feb. 2012), 77–96. doi:10.1108/10444061211199322

William V. Pelfrey and Nicole Weber. 2013. Talking smack and the telephone game: conceptualizing cyberbullying with middle and high school youth. *Journal of Youth Studies* 17, 3 (Sept. 2013), 397–414. doi:10.1080/13676261.2013.830702

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi

Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 13387–13434. doi:10.18653/v1/2023.findings-acl.847

Andrzej Porębski and Jakub Figura. 2025. There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications* 12, 1 (Oct. 2025), 1647. doi:10.1057/s41599-025-05868-8

Konstantinos I. Roumeliotis and Nikolaos D. Tselikas. 2023. ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet* 15, 6 (May 2023), 192. doi:10.3390/fi15060192

Katie Seaborn. 2025a. Insidious by design: Time and decepion in human-LLM interactions. In *Dark Patterns and Manipulative Design: Conceptualising and Systematising a Key Contemporary Phenomenon from a Legal Perspective and Beyond*, Raphael Gellart, Hanna Schraffenberger, and Cristiana Santos (Eds.). Edward Elgar Publishing, London, UK.

Katie Seaborn. 2025b. Social Identity in Human-Agent Interaction: A Primer. *ACM Transactions on Human-Robot Interaction* (Aug. 2025). doi:10.1145/3760500

Katie Seaborn, Satoru Iseya, Shun Hidaka, Sota Kobuki, and Shruti Chandra. 2024. Play across boundaries: Exploring cross-cultural maldaimonic game experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article No. 564. doi:10.1145/3613904.3642273

Katie Seaborn and Rua Mae Williams. 2025. Levelling Up the Literature: How Games Can Inspire Academic Reporting. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY Companion '25)*. Association for Computing Machinery, New York, NY, USA, 236–238. doi:10.1145/3744736.3749331

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. doi:10.48550/ARXIV.2310.13548

Konstantinos C Siontis, Zachi I Attia, Samuel J Asirvatham, and Paul A Friedman. 2023. ChatGPT hallucinating: can it get any more humanlike? *European Heart Journal* 45, 5 (Dec. 2023), 321–323. doi:10.1093/eurheartj/ehad766

Kavyaa Somasundaram, Andrey Kiselev, and Amy Loutfi. 2023. Intelligent Disobedience: A Novel Approach for Preventing Human Induced Interaction Failures in Robot Teleoperation. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. ACM, New York, NY, USA, 142–145. doi:10.1145/3568294.3580060

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis)informs us better than humans. *Science Advances* 9, 26 (June 2023). doi:10.1126/sciadv.adh1850

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* 7, 2 (Feb. 2025), 221–231. doi:10.1038/s42256-024-00976-7

Alessandro Trevisan, Harry Giddens, Sarah Dillon, and Alan F. Blackwell. 2024. The BS-meter: A ChatGPT-Trained Instrument to Detect Sloppy Language-Games. doi:10.48550/ARXIV.2411.15129

Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA, 1–14. doi:10.1145/3613904.3642135

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817 (Feb. 2025). doi:10.48550/arXiv.2401.11817 arXiv:2401.11817 [cs].

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2024. Empirical evidence of Large Language Model's influence on human spoken communication. doi:10.48550/ARXIV.2409.01754

Jeremy A. Yip, Maurice E. Schweitzer, and Samir Nurmohamed. 2018. Trash-talking: Competitive incivility motivates rivalry, performance, and unethical behavior. *Organizational Behavior and Human Decision Processes* 144 (Jan. 2018), 125–144. doi:10.1016/j.obhdp.2017.06.002

Chunpeng Zhai, Santoso Wibowo, and Lily D. Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (June 2024). doi:10.1186/s40561-024-00316-7

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How Language Model Hallucinations Can Snowball. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 59670–59684. https://proceedings.mlr.press/v235/zhang24ay.html

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics* 51, 4 (2025), 1373–1418. doi:10.1162/coli.a.16

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. 2024. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics* 16, 12 (Nov. 2024), 9851–9915. doi:10.1007/s13042-024-02443-6