NLP Meets the World: Toward Improving Conversations With News Media About Natural Language Processing Research

Shomir Wilson
Department of Human-Centered Computing and Social Informatics
College of Information Sciences and Technology
Pennsylvania State University
shomir@psu.edu

Abstract

Recent advances in large language models (LLMs) have rapidly increased the public's interest in natural language processing (NLP). This attention is reflected by news outlets inviting NLP researchers to share their knowledge and views with public audiences. Recognizing the opportunities for the research field and for individual researchers, this paper shares recommendations for speaking with news media about NLP's capabilities and limitations. Drawing on the topical structure of research communication frameworks published by the US National Academies of Sciences, Engineering, and Medicine and the European Union's QUEST project, recommendations are provided for three field-appropriate communication challenges: cognitive terminology, unsustainable expectations, and ethical failures. Published NLP research and popular news coverage are cited to provide examples. The recommendations promote effective, transparent communication with the general public about NLP, in order to strengthen public understanding and encourage support for research.

Note

*Because of the subject matter, this paper cites a variety of types of sources, including news articles, unpublished essays, preprints, and whitepapers. Citations and references marked with an asterisk (i.e., *) represent sources that are not known to be peer-reviewed.*

1 Introduction

Public communication is a vital task for the natural language processing (NLP) research community, which faces a unique confluence of circumstances. Interest in LLMs has led news venues to publish non-technical explanations of how LLMs work (Roose, 2023*; Clarke et al., 2023*), predictions about LLMs' future abilities (Waters, 2025*), explorations of LLMs' social biases (Rogers, 2025*), and concerns about their impacts on employment (Hayes, 2025*). Journalists also have interviewed researchers who critique claims about artificial general intelligence (AGI) (Heikkilä, 2025*; Piquard, 2024*). Within scholarly literature, LLMs are sometimes claimed to be an existential threat to humanity (Tait and Bensemann, 2024; Grozdanoff, 2023), while some critique the plausibility or meaning of AGI (Blili-Hamelin et al., 2025; Mueller, 2024*). Regardless of stance, these activities reflect an unprecedented surge of public attention.

The public discussion of NLP extends into culture and language, as NLP researchers contend with portrayals of AI in science fiction (Osawa et al., 2022) and ambiguity over cognitive terms such as *reason* and *understand* (Mitchell and Krakauer, 2023). While research communication is a challenging task for many scholarly disciplines (Simis et al., 2016), little support is available for the acute challenges that NLP researchers currently face. This gap in support contrasts with the importance of research communication: it encourages public support for research funding, influences public policy, and engages students who may become future researchers.

This paper serves as a referenceable set of recommendations for NLP researchers interacting with news media, including interviewing with journalists and publishing commentary. It derives thematic and structural inspiration from research communication frameworks published by the US National Academies of Sciences, Engineering, and Medicine (2017*) and the European Union's QUEST project (2020*). Their guidance is translated into three challenges for communicating NLP research: *cognitive terminology* (§3), *unsustainable expectations* (§4), and *ethical failures* (§5). Rather than replacing existing guidance (as reviewed in §2), this paper fills a notable gap by

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 32910–32917
October 24-29, 2026 ©2026 Association for Computational Linguistics

attending to the contemporary challenges specifically relevant to NLP.

2 Background and Related Work

The US National Academies created a framework for exploring the task of research communication (2017*), but its guidance is intentionally broad (i.e., for all disciplines at once). This paper translates the three core challenges in their framework into challenges for NLP: *the complexities of communicating science* (cognitive terminology, §3), *communicating science in a complex, competitive communication environment* (unsustainable expectations, §4), and *the nature of science-related public controversies* (ethical failures, §5). Similarly, the European Union-funded QUEST (QUality and Effectiveness in Science and Technology communication) project identified 12 quality indicators for science communication and clustered them into three themes (QUEST Project, 2020*). Their themes are aligned with the same challenges: *presentation* (cognitive terminology), *connection* (unsustainable expectations), and *trustworthiness* (ethical failures). No such mapping or guidance can be comprehensive, which is discussed further in Limitations, but these prior frameworks lend focus and structure to the present work.

Prior published literature on research communication provides little guidance tailored to NLP (i.e., in contrast with general guidance for researchers in any discipline). Notably, Hudson and Franklin (2023*) observe that using performance figures to describe explainable AI research alienates public audiences. A series of AAAI tutorials have focused on science communication (e.g., Smith (2025)), but their reach and accessibility differ from a published paper. Wyrich and Wagner (2023) designed a course-based exercise for computer science undergraduates to practice communicating software engineering research, but they focus on describing the exercise itself. Beyond computing, Fontaine et al. (2019) surveyed the literature on public communication by researchers within the health sciences. They noted wide use of social media and proposed a typology of strategies. These include minimizing the use of jargon, encouraging public discussion, and making information actionable, among others. Kuncel and Rigdon (2013*) address science communication for psychology, with a focus on making statistical methods understandable.

Finally, this paper follows an established practice of NLP conference papers that reflect on community challenges. Examples include Rogers and Augenstein's recommendations for improving peer review (2020), Blodgett et al.'s critique of how *bias* is discussed (2020), and Mosbach et al.'s discussion of interpretability research (2024).

3 Cognitive Terminology

NLP researchers often describe their work using terms that public audiences associate with cognition, and the cognitive claims those terms imply might be intended or unintentional. Additionally, the research community lacks an agreed-upon stance on some cognitive claims, partly because of a lack of agreement for what they mean (Bender and Koller, 2020). These conditions are difficult for non-experts to navigate, limiting their abilities to interpret or support NLP research. This section illustrates with terms used in published research.

3.1 It's Complicated: NLP's Relationship With Cognitive Terms[1]

Terms with cognitive implications (among their other implications) have a long history in NLP, and some are widely interpreted by researchers as non-cognitive. *Predict*, for example, has been associated with text classification at least as early as the 1990s (Riloff and Lehnert, 1992; Liddy et al., 1994) and possibly earlier. *Read* also has a long history (Srihari, 1992; Govindan and Shivaprasad, 1990). *Learn* (as in *machine learning*) originates from a 1959 paper by Arthur Samuel about programs that play the game of checkers (Samuel, 1959; Foote, 2021*). However, a non-expert may associate all three of those terms with the ability to *think* or *understand*, broadly without definitions. Observe that "a *human* must understand to predict" is intuitively reasonable, but several papers cited in §1 show that "a *language model* must understand to predict" is not universally agreed upon by the NLP research community.

LLMs' capabilities and limitations have amplified attention to cognitive terms. *Hallucinate* is widely used by the research community (Huang et al., 2025), even if a precise meaning is not agreed upon (Narayanan Venkit et al., 2024). However, the relevance of *understand*, *reason*, and *think* are controversial (Bender and Koller, 2020; Shojaee et al.,

[1]Note this manuscript does not evaluate cognitive claims about LLMs. Instead, it shows how the NLP research community uses cognitive terms differently from the general public, including examples that predate LLMs.

2025*). Again, the relationships between these terms within NLP may not match the relationships intuited by an outside audience. For example, a human who is *hallucinating* is presumably *thinking* (albeit aberratively), but a *hallucinating* LLM is not agreed upon by the research community to be *thinking*. Additionally, researchers in the health sciences have criticized using *hallucinate* to refer to LLM behavior as inappropriate and misleading (Smith et al., 2023; Østergaard and Nielbo, 2023).

A further complication occurs when arguments against LLMs' cognitive abilities nominally exclude humans from having those same abilities. For example, Shojaee et al. (2025*) argue that language reasoning models (LRMs, closely related to LLMs) cannot *think* because of "a complete accuracy collapse" for complex tasks. The human brain also has task complexity limits (Marois and Ivanoff, 2005), but in contrast with LLMs, the human brain is widely recognized as capable of *thinking*.

3.2 Recommendations

NLP researchers speaking to the public should provide scope for cognitive terms they use. Given the above obstacles, it is helpful to tell an audience what terms such as *predict* or *reason* mean in a context. If a non-cognitive term can replace a cognitive one, that substitution may provide greater clarity. When benchmarks are relevant, describing them can demystify how a cognitive or quasi-cognitive claim is justified. A general goal is to avoid the possibility of unintended cognitive implications for LLMs or other human language technologies.

NLP researchers should balance correcting cognitive misconceptions about LLMs with tolerating a general audience's verbal shorthand and metaphor. Humans apply cognitive anthropomorphism to a variety of objects (Boyer, 1996) such as lamps and umbrellas (Edwards and Shafer, 2022), and LLMs display far more sophisticated behavior than those objects. When cognitive terms surface in discussions about LLMs, a diplomatic strategy is to point out the verbal shorthand or metaphor, acknowledge its naturalness, and explain the actual limitations of the technology. This strategy validates a non-expert's experiences with technology while illuminating how cognitive terms can mislead.

4 Unsustainable Expectations

Predictions of the impacts of LLMs include radical improvements in biology, neuroscience, economic development, and human welfare, and the onset of global peace (Amodei, 2024*). The possibility of an "AI bubble"[2] (i.e., an unsustainable level of public interest in AI that exceeds its capabilities) is widely acknowledged (Horowitch, 2025*; Press, 2025*). A bubble implies short-term benefits, such as publicity and abundant resources. However, its collapse—implied by the metaphor—creates harm through an unpredictably sharp decline in resources. This section provides examples from the history of AI and recommendations for how NLP researchers can encourage realistic expectations about the field.

4.1 Remembering AI Winters

Multiple AI winters are said to have occurred: Roger Schank and Marvin Minsky referred to one in the 1970s (Crevier, 1993*), and James Hendler referred to one in the 1980s (Hendler, 2008*). This section uses Hendler's more recent perspective to structure the discussion.

Hendler described a confluence of three conditions that led to the AI winter of the 1980s. The first was a decrease in government funding for AI research in the 1970s, specifically in the USA, which dominated AI research at the time. He argues the impact was delayed by five to ten years, as already-approved projects ran their course. For the rationale of the cuts, Hendler quotes a 1973 report by James Lighthill in the United Kingdom that concluded "in no part of the field have discoveries made so far produced the major impact that was then promised." (Lighthill et al., 1973*) The second condition was an erosion of recognition for AI's successes. Hendler argues that expert systems (Buchanan and Smith, 1988), which were successful in industry, were "disowned" by AI researchers in favor of topics where less progress had been made. The third condition was the rapid democratization of hardware needed to run AI models, in the sense that the hardware became dramatically cheaper and easier to own, thus disrupting sociotechnical power structures within research (Hendler, 2008*).

These conditions require some subjectivity to assess, making it difficult to firmly state whether they exist now or will appear in the near future. However, there are discernible lessons in them for how NLP researchers communicate their work.

[2]There exists a lack of agreement on the distinction between contemporary AI and NLP, with some common ground and some differences. To avoid misattribution, this paper uses "AI" for topics where the term prevails in popular discussion.

4.2 Recommendations

For the NLP research community as a whole, the downsides of setting realistic expectations are modest when compared to the downsides of a sudden collapse in public interest. At the individual level, one researcher's careful management of expectations when speaking about their work is unlikely to change public perception of the field, but strategies exist that fulfill that goal while creating individual benefit as well. Such a strategy is described below.

Public engagement is an opportunity to support recognition of NLP applications that may have been inadvertently "disowned" (borrowing Hendler's description) for their success. NLP played a formative role in early search engines (Schwartz, 1998) and enabled early spam detection (Cormack, 2008). Optical character recognition made a wealth of pre-digital texts searchable (Hill and Hengchen, 2019), and authorship identification solved historical mysteries (Tweedie et al., 1996). Those examples and others show that NLP can enable *calm technologies*, supporting human work rather than monopolizing human attention (Weiser and Brown, 1996). They also may serve as an antidote to hype by illustrating that research products need time to mature. Dialog systems are a strong public-facing example of the long road to practical use: early systems such as Eliza (Weizenbaum, 1966) are distant ancestors of modern virtual assistants, such as Amazon Alexa, Siri, and Google Assistant.

5 Ethical Failures

Ethically dubious practices are known to exist within the NLP research community, and external to it, problematic NLP applications also exist. Both are obstacles to how NLP research is received by a public audience. This section gives examples of relevant ethical failures and provides suggestions for how NLP researchers can speak frankly and constructively when called upon to discuss them.

5.1 Reckoning With Problematic NLP

Research institutions in many countries have procedures for reviewing experimental protocols that involve human subjects. However, sometimes an experiment complies with requirements while violating ethical norms, or researchers fail to seek approval. A recent example is an experiment in which researchers studied the persuasiveness of an LLM-powered chatbot by covertly deploying it on the Reddit forum *r/changemyview*[3], concealing that it was an LLM. When forum moderators were informed about the deception, they expressed shock, asked for an apology from the university, and requested that the research not be published (O'Grady, 2025). The experiment also reflects a mismatch between researchers' expectations and a studied population's expectations for ethical uses of NLP. Another example of mismatched ethical expectations is the use of deidentified conversation data from Crisis Text Line, a free nonprofit service for people in crisis in the US, by a for-profit company. Although users of Crisis Text Line were provided with a link to terms of service, informing them how their data could be shared, public backlash led to the termination of the data sharing agreement (Eysenbach, 2025).

A comprehensive list of ethical failures or misuses of NLP is impractical, but examples are readily available. LLMs have been used to generate disinformation, potentially manipulating geopolitics (Mejias and Vokuev, 2017). Chatbots have demonstrated the capacity to respond harmfully in conversations with people suffering from mental illness (Coghlan et al., 2023), and more broadly, LLMs' sociodemographic biases have been shown to harm individuals from minoritized groups (Ghosh et al., 2024). The broad-based collection of training data for LLMs raises privacy concerns and intellectual property concerns (Novelli et al., 2024). NLP and computing security are increasingly intertwined, obligating NLP researchers to learn more about the nuanced ethics of security research (Lent et al., 2025). LLM-assisted writing has been claimed to produce "cognitive debt" and reduce critical thinking (Kosmyna et al., 2025*). LLM training and usage consume significant resources, raising questions about environmental sustainability (Kogkalidis and Chatzikyriakidis, 2025). Finally, NLP research is vulnerable to distortion by monied interests, potentially influencing individual researchers and entire institutions (Abdalla and Abdalla, 2021).

The above harms and others receive public attention, to the extent that a 2023 Pew Research survey found 52% of Americans were more concerned than excited about AI in their daily lives (Faverio and Tyson, 2023*). This level of concern is demonstrably relevant to NLP researchers.

[3]https://www.reddit.com/r/changemyview/

5.2 Recommendations

Several popular news articles cited in this paper show that NLP researchers are called upon to discuss problematic uses of NLP. Open dialog about these problems is obligatory and familiar as part of open science, but further steps can help to sustain public support. Researchers speaking with the general public about problematic NLP should be prepared to explain what should have been done differently to achieve the same goals, or why the goals never should have been undertaken. Public interaction is also an opportunity to point to research directly aimed at solving ethical issues in NLP, showing how human-centered research is a valuable part of the community. Such research is a marked contrast with a hypothetical public image of AI research as solely focused on computation.

Within the research community, researchers can lay the groundwork for strong public support by recognizing that research that critiques NLP fully belongs in the community. Such research does not necessarily create human language technologies, but it guides their ethical development. This inclusion also encourages grounding NLP research in human benefit, which demonstrates responsiveness to the concerns of the general public.

6 Conclusion

Beyond this paper, other strategies are possible for encouraging NLP researchers to engage productively with news media. Among them are holding media training sessions at conferences and developing curricular modules for students to learn about research communication. A multi-pronged strategy, including these activities or others, has the greatest potential to reach a wide audience of researchers.

Public attention to NLP research comes with benefits, obstacles, and responsibilities. This paper presented recommendations to help NLP researchers communicate effectively with the public, focusing on obstacles that are distinctive to our research community. It drew upon a variety of sources to illustrate those obstacles and how to navigate them, for the prosperity of individual researchers and for the research community at large.

Acknowledgments

The reflections in this manuscript benefited from several experiences discussing research with news media, and I gratefully acknowledge the work of Penn State's Office of Strategic Communications and (also within Penn State) the College of IST's Office of Marketing and Communications, who made those media engagements possible. I also gratefully acknowledge Kenneth Huang and Natalie Parde for their feedback on ideas in this manuscript. This manuscript is based upon work supported by the National Science Foundation under Grant #2237574.

Limitations

A comprehensive guide to research communication, examining all possible scenarios, is infeasible. Instead, this manuscript augments existing discipline-nonspecific guides to focus on NLP. In §1 and §2 readers are reminded of that, and in §2 more sources are cited for the reader to consult. Additionally, the complexity and subjectivity of research communication mean that no one guide (or even the union of all guides) will be comprehensive for the problems researchers face. However, other fields appear to find research communication guides helpful, as shown by their acceptance for publication, the citations to them, and reflections on them (de Bruin and Bostrom, 2013).

NLP researchers may face public communication challenges not covered by this paper's guidance. This paper identifies challenges of particular concern by interpreting established research communication frameworks, but other challenges may exist. The subjective nature of the topic space obviates covering it in an exhaustive way, but sharing guidance grounded in recognized prior frameworks (as this manuscript does) is deemed more likely to lead to positive outcomes than the relative lack of guidance in the status quo.

Ethical Considerations

This paper is reflective in purpose and does not represent experiments, and its ethical ramifications are limited to the guidance it provides. Existing research communication frameworks and other cited literature provide evidence that the topics in this manuscript are relevant and have received little or no prior attention for NLP. The reader is encouraged to consult additional guidance as necessary.

References

Mohamed Abdalla and Moustafa Abdalla. 2021. The grey hoodie project: Big tobacco, big tech, and the

threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 287–297, New York, NY, USA. Association for Computing Machinery.

Dario Amodei. 2024*. Machines of loving grace: How AI could transform the world for the better. Online essay. https://www.darioamodei.com/essay/machines-of-loving-grace.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Borhane Blili-Hamelin, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A Heller, Jacob Metcalf, Fabricio Murai, Eryk Salvaggio, Andrew J Smart, Todd Snider, Mariame Tighanimine, Talia Ringer, Margaret Mitchell, and Shiri Dori-Hacohen. 2025. Position: Stop treating 'AGI' as the north-star goal of AI research. In *Forty-Second International Conference on Machine Learning Position Paper Track*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Pascal Boyer. 1996. What makes anthropomorphism natural: Intuitive ontology and cultural representations. *The Journal of the Royal Anthropological Institute*, 2(1):83–97.

Bruce G Buchanan and Reid G Smith. 1988. Fundamentals of expert systems. *Annual Review of Computer Science*, 3(1):23–58.

Seán Clarke, Dan Milmo, and Garry Blight. 2023*. How AI chatbots like ChatGPT or Bard work – visual explainer. *The Guardian*. Accessed: 2025-06-19.

Simon Coghlan, Kobi Leins, Susie Sheldrick, Marc Cheong, Piers Gooding, and Simon D'Alfonso. 2023. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digital health*, 9.

Gordon V. Cormack. 2008. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455.

Daniel Crevier. 1993*. *AI: The Tumultuous History of the Search for Artificial Intelligence*, 1st edition. Basic Books, New York, NY.

Wändi Bruine de Bruin and Ann Bostrom. 2013. Assessing what to address in science communication. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14062–14068.

Alyssa D Edwards and Daniel M Shafer. 2022. When lamps have feelings: Empathy and anthropomorphism toward inanimate objects in animated films. *Projections*, 16(2):27–52.

Gunther Eysenbach. 2025. Crisis text line and Loris.ai controversy highlights the complexity of informed consent on the internet and data-sharing ethics for machine learning and research. *Journal of medical Internet research*, 27.

Michelle Faverio and Alec Tyson. 2023*. What the data says about Americans' views of artificial intelligence. Pew Research Center – Short Reads.

Guillaume Fontaine, Marc-André Maheu-Cadotte, Andréane Lavallée, Tanya Mailhot, Geneviève Rouleau, Julien Bouix-Picasso, and Anne Bourbonnais. 2019. Communicating science in the digital and social media ecosystem: Scoping review and typology of strategies used by health scientists. *JMIR Public Health and Surveillance*, 5(3):e14447.

Keith D. Foote. 2021*. A brief history of machine learning. Online article. https://www.dataversity.net/a-brief-history-of-machine-learning/.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do generative AI models output harm while representing non-western cultures: Evidence from a community-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 7 of *AIES '24*, pages 476–489.

V.K Govindan and A.P Shivaprasad. 1990. Character recognition — a review. *Pattern Recognition*, 23(7):671–683.

Boris D Grozdanoff. 2023. The looming shadow: Taking seriously potential existential threats brought about by artificial intelligence. *Ethical Studies*, pages 100–106.

Adam Hayes. 2025*. Is AI coming for your job? Here's how to tell. *Investopedia*. Accessed: 2025-06-19.

Melissa Heikkilä. 2025*. Margaret Mitchell: Artificial general intelligence is 'just vibes and snake oil'. *Financial Times*. Accessed: 2025-06-19.

James Hendler. 2008*. Avoiding another AI winter. *IEEE Intelligent Systems*, 23(2):2–4.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Rose Horowitch. 2025*. The computer-science bubble is bursting. *The Atlantic*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting

Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Simon Hudson and Matija Franklin. 2023*. Science communications for explainable artificial intelligence. *Preprint*, arXiv:2308.16377.

Konstantinos Kogkalidis and Stergios Chatzikyriakidis. 2025. On tables with numbers, with numbers. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 104–115, Albuquerque, New Mexico. Association for Computational Linguistics.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025*. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *Preprint*, arXiv:2506.08872.

Nathan R. Kuncel and J. Rigdon. 2013*. Communicating research findings. In Neal W. Schmitt, Scott Highhouse, and Irving B. Weiner, editors, *Handbook of Psychology: Industrial and Organizational Psychology*, 2nd edition, volume 12, pages 43–58. John Wiley & Sons, Inc., Hoboken, NJ.

Heather Lent, Erick Galinkin, Yiyi Chen, Jens Myrup Pedersen, Leon Derczynski, and Johannes Bjerva. 2025. NLP security and ethics, in the wild. *Transactions of the Association for Computational Linguistics*, 13:709–743.

Elizabeth D. Liddy, Woojin Paik, and Edmund S. Yu. 1994. Text categorization for multiple users based on semantic features from a machine-readable dictionary. *ACM Trans. Inf. Syst.*, 12(3):278–295.

James Lighthill, Sutherland Stuart, Needham Roger, and Longuet-Higgins Christopher. 1973*. Artificial intelligence: A general survey. *Science Research Council*.

René Marois and Jason Ivanoff. 2005. Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6):296–305.

Ulises A Mejias and Nikolai E Vokuev. 2017. Disinformation and the media: the case of russia and ukraine. *Media, Culture & Society*, 39(7):1027–1042.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105, Miami, Florida, USA. Association for Computational Linguistics.

Milton Mueller. 2024*. The myth of AGI. *Internet Governance Project*.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An audit on the perspectives and challenges of hallucinations in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6528–6548, Miami, Florida, USA. Association for Computational Linguistics.

National Academies of Sciences, Engineering, and Medicine. 2017*. Communicating science effectively: A research agenda.

Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law & Security Review*, 55:106066.

Cathleen O'Grady. 2025. 'Unethical' AI research on Reddit under fire. *Science*, 388(6747):570–571. Perspective.

Hirotaka Osawa, Dohjin Miyamoto, Satoshi Hase, Reina Saijo, Kentaro Fukuchi, and Yoichiro Miyake. 2022. Visions of artificial intelligence and robots in science fiction: a computational analysis. *International Journal of Social Robotics*, 14(10):2123–2133.

Søren Dinesen Østergaard and Kristoffer Laigaard Nielbo. 2023. False responses from artificial intelligence models are not hallucinations. *Schizophrenia bulletin*, 49(5):1105–1107.

Alexandre Piquard. 2024*. Chatbots are like parrots, they repeat without understanding. *Le Monde (English edition)*. Interview with Emily M. Bender; Accessed: 2025-06-19.

Gil Press. 2025*. Are we at peak AI bubble and the cusp of AI moment? *Forbes*.

QUEST Project. 2020*. 12 quality indicators for science communication: Guide for science communication. Accessed: 2026-05-13.

Ellen Riloff and Wendy Lehnert. 1992. Classifying texts using relevancy signatures. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Reece Rogers. 2025*. AI is spreading old stereotypes to new languages and cultures. *WIRED*. Accessed: 2025-06-19.

Kevin Roose. 2023*. How does ChatGPT really work? *New York Times*, 28.

A. L. Samuel. 1959. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229.

Candy Schwartz. 1998. Web search engines. *Journal of the American Society for Information Science*, 49(11):973–982.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025*. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Preprint*, arXiv:2506.06941.

Molly J Simis, Haley Madden, Michael A Cacciatore, and Sara K Yeo. 2016. The lure of rationality: Why does the deficit model persist in science communication? *Public understanding of science*, 25(4):400–414.

Andrew L Smith, Felix Greaves, and Trishan Panch. 2023. Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health*, 2(11):e0000388.

Lucy Smith. 2025. Science communication for AI researchers: Introductory training. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI-25), Tutorial Track*.

S.N. Srihari. 1992. High-performance reading machines. *Proceedings of the IEEE*, 80(7):1120–1132.

Izak Tait and Joshua Bensemann. 2024. Clipping the risks: Integrating consciousness in AGI to avoid existential crises. In *International Conference on Artificial General Intelligence*, pages 176–182. Springer.

Fiona J Tweedie, Sameer Singh, and David I Holmes. 1996. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, 30:1–10.

Richard Waters. 2025*. The diverging future of AI. *Financial Times*. Accessed: 2025-06-19.

Mark Weiser and John Seely Brown. 1996. Designing calm technology. *PowerGrid Journal*, 1(1):75–85.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Marvin Wyrich and Stefan Wagner. 2023. Teaching computer science students to communicate scientific findings more effectively. In *Proceedings of the 45th International Conference on Software Engineering: Software Engineering Education and Training*, ICSE-SEET '23, page 107–114. IEEE Press.