Developing Story: Case Studies of Generative AI's Use in Journalism

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno
Franziska Roesner, Niloofar Mireshghallah
University of Washington
natalie.brigham@gmail.com, {yoshi, franzi, niloofar}@cs.washington.edu

Abstract

Journalists are among the many users of large language models (LLMs). To better understand the journalist-AI interactions, we conduct a study of LLM usage by two news agencies through browsing the WildChat dataset, identifying candidate interactions, and verifying them by matching to online published articles. Our analysis uncovers instances where journalists provide sensitive material such as confidential correspondence with sources or articles from other agencies to the LLM as stimuli and prompt it to generate articles, and publish these machine-generated articles with limited intervention (median output-publication ROUGE-L of 0.62). Based on our findings, we call for further research into what constitutes responsible use of AI, and the establishment of clear guidelines and best practices on using LLMs in a journalistic context.

1 Introduction

Given the potential of LLMs to assist journalists and increase productivity, several initiatives are aiming to assist news organizations in finding, training, and applying AI-based solutions responsibly (Associated Press; Partnership on AI). However, many have raised concerns about the application of LLMs in the field of journalism, including misinformation (Pan et al., 2023), copyright violations (Karamolegkou et al., 2023), and privacy implications (Yao et al., 2024). Due to these issues, LLMs can be seen as both a tool for journalism and a threat to journalistic integrity (Wihbey, 2024), depending on how they are used. While previous work surveyed journalists' reported usage of generative AI (Diakopoulos et al., 2024; Gondwe, 2023), the opaque nature of newsrooms and challenges in detecting AI-generated content (Chakraborty et al., 2023; Weber-Wulff et al., 2023) have left a knowledge gap in the computer science research community regarding its on-the-ground utilization. This work aims to bridge this gap by analyzing journalist-AI interactions, including queries, provided materials, intervention levels, and query-to-publish timelines.

To achieve this, we probe the publicly available WildChat dataset (Zhao et al., 2024) of human-chatbot interactions. We identify potential journalist queries in WildChat, match them to published articles on two news agency websites, and analyze the resulting set of interactions. Given WildChat's scope and that other agencies likely use generative AI, we refer to the two identified agencies as Agency A and Agency B to maintain anonymity (see limitations and ethical considerations). We categorize the types of tasks for which LLMs were used across Agencies A and B and the distribution of input materials employed to generate articles, shown in Figure 3. Inputs include articles from other agencies or private conversations, real case examples of which are depicted in Figures 1 and 2 (Agencies B and A, respectively). We also analyze the extent of human intervention in article generation by examining the overlap between the model generated drafts and the matched published articles, along with the time between generation and publication. Finally, we go beyond the WildChat dataset and study machine-generated text on Agency A's and B's websites using GPTZero (Tian & Cui, 2023), a state-of-the-art commercial LLM-generation detector.

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

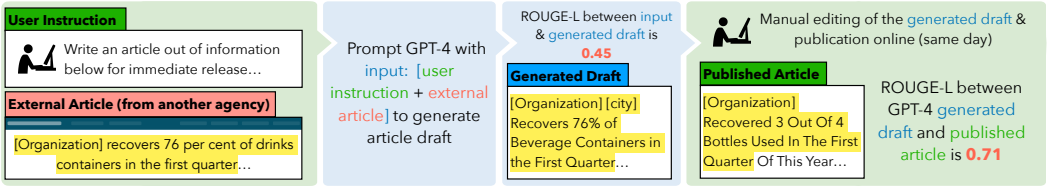


Figure 1: Case study of a single-turn journalist-LLM interaction for article generation where an external article by another agency is used as the input stimulus. The generated draft is published by the journalist with little modification as the ROUGE-L between the published article we identified and the generation is 0.71. The '[' and ']' symbols denote portions of the text that have been replaced to minimize identifiability.

Agency	A	B	Sum
Conversations	62	16	78
Turns	107	41	148
Verified published articles	79	32	111

Table 1: Evidence of AI-assisted journalism from WildChat

Task (turn count)	Agency A	Agency B
Article generation	89	34
Headline generation	18	1
Article editing	0	6

Table 2: Turn counts for each task type across both agencies.

Our findings offer insights about the role of LLMs in journalism and potential impacts to journalistic integrity, as we find substantial overlap between the model outputs and the published articles with a median ROUGE-L (Lin, 2004) score of 0.62, and a prompt-to-publication span of a single day. Indeed, 18% of the identified input stimuli are other agencies' news articles and 9% are potential private conversations, risking privacy breaches by sharing the data with the LLM-provider and publicly via WildChat. Our study points to a need for better AI literacy and education for model practitioners. A potential path forward could draw from human-computer interaction and usable security and privacy research on "nudging" users toward beneficial behaviors (Acquisti et al., 2017), which may offer a promising solution in this scenario as well. We also encourage the journalistic and computer science communities to iteratively refine guidelines (Kent, 2015; Dolnick & Seward, 2024; Cools & Diakopoulos, 2023) for the use of AI in journalism as technologies like LLMs become more advanced and ubiquitous.

Related work. Our work relates to studies on human-chatbot interactions for assistive writing, such as in academic paper writing (Liang et al., 2024b) and in the peer-review process (Liang et al., 2024a), but differs in topic and methodology, providing visibility into interactions and ground truth for machine-generated articles. We provide an extended study of related work in Appendix A.

2 Method

WildChat. Since our aim is to closely study the type of queries made by journalists to LLMs and to uncover interventions they make to model outputs before publication, we probe WildChat (Zhao et al., 2024), a publicly available dataset of 650k[1] conversations collected by offering free access to GPT-3.5 and GPT-4 to users, for potential candidate interactions.

Identifying journalist queries. We initially identified conversations with 4 or more PII types using an NER model[2], as we were interested in identifying sensitive disclosures. This yielded a $5k$ turn subset of the data. We then manually inspected this set and found instances of article generation for two different news agencies: Agency A, a local news outlet based in southern California, and Agency B, a local news platform covering a small nation in the Mediterranean region. Using location-based keywords for these agencies, we filtered the entire 650k conversation WildChat dataset and manually reviewed the results to find additional conversations related to these agencies.

[1]Recently a 1M version of this dataset was released, however our study is conducted on an earlier version.
[2]`lakshyakh93/deberta_finetuned_pii`

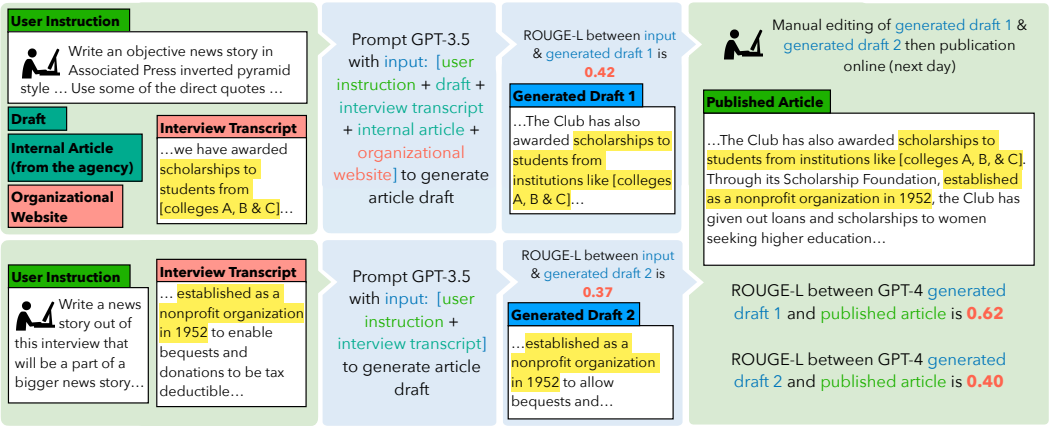


Figure 2: Case study of a multi-turn article generation using multiple stimuli, including an internal article from the same agency and an interview transcript. The '[' and ']' symbols denote portions of the text that have been replaced to minimize identifiability.

We acknowledge that this process may result in false negatives, missing cases of journalistic activity in WildChat that we did not identify. (See Section 5 for reasons why we upload conversation IDs but do not mention the specific agencies in the main text.) To match the output of conversational turns to published articles and verify that the queries were made by the journalists, we searched the identified agencies' websites for articles with highly similar content to the generated text.

Categorizing tasks. For each verified conversational turn, based on the user's prompt, we classify the journalistic tasks: (1) *article generation* which involves the LLM creating a new article from the provided user instruction and input material, (2) *headline generation* which involves generating a headline for a given article, and (3) *article editing* which involves the user requesting edits to a provided draft article.

Categorizing input stimuli. By stimulus, we refer to the input material provided to the LLM assistant as source or context. One researcher examined all user prompts for *article generation*, developed a codebook that was reviewed by the team, and, to ensure coding consistency and retain interpretive nuance (Armstrong et al., 1997; Morse, 1997), coded all stimuli. Then, a second researcher independently reviewed and validated the coding (Whittemore et al., 2001). Stimuli was only coded as an externally sourced type (e.g., press release, external news article) when verbatim text matched online sources. Thus, these codes represent a lower bound on the external source material.

Measuring journalist intervention. As a proxy for how much the journalists modify the model's output before publication online, we report the ROUGE-L recall score (Lin, 2004) between the generated output (used as source) and the matched online article. This score reflects the longest common subsequence between the two text sequences, normalized to the length of the source. We also report this score over the model's input and output.

Detecting LLM-generated articles beyond WildChat. To extend our study beyond the WildChat dataset, we scrape articles from the two identified news agencies' websites from between January 1, 2020 and April 15, 2024, use GPTZero (Tian & Cui, 2023), a state-of-the-art commercial machine-generated text detection method to study the prevalence of LLM use, and report our findings.

3 Findings

Table 1 summarizes the identified content from WildChat, with the last row showing the number of online articles we could match to the identified activity. We select two case studies that represent archetypal examples of different article generation behaviors, and use them for demonstrations throughout this section: Figure 1, a case of a single-turn article generation (from Agency B) and Figure 2, a multi-turn process of article generation (from Agency A). Below we discuss our observations from the two agencies.

What do journalists in our study prompt LLMs with? Table 3 provides input stimuli types, descriptions, and frequencies, Figure 3 shows the distribution of input stimulus types, and Appendix D

Stimuli type	Agency A	Agency B
Draft or other: Draft of article or unidentified material	39	9
Press release: Article on the same topic released by the subject of the story	30	4
External news article: Article on the same topic from another news agency	18	15
Interview: Transcript or written interview responses	14	2
Organizational report: Official report from a city or country government or organization (e.g., a school district's 100 Day Plan)	12	2
Event posting: Posting of event details (e.g., Meetup, event website)	12	0
Organization website: General information on an organization, company, or person (e.g., "About" page)	9	0
Intra-agency news article: Article on the same topic from within the agency	1	8
Email: Message from a source or editor about a story idea	0	5
Social media post: Post on social media (e.g., LinkedIn, GoFundMe)	2	0

Table 3: Types of stimulus used to generate articles in the identified WildChat activity and their frequencies across Agencies A and B, sorted in decreasing order from top to bottom by their combined frequencies.

provides additional results for this section. Across the 148 turns, there is a total of 182 stimuli used (all turns had at least one stimulus as input and some had multiple, some turns had multiple instance of the same stimulus type). Stimuli from external sources (i.e., news articles from other agencies, press releases, organizational reports, event postings, organizations' websites, social media posts) accounted for over two-thirds (83/137) of the identified stimuli for Agency A and about half (21/45) for Agency B. We also observe internally sourced stimuli types, such as draft material and articles originating from within the agency.

In Figure 1, we observe Agency B using a news article from another agency as a stimulus, a practice we identified relatively frequently for both agencies. In another case of this, a user from Agency B specified, *"write a new article out of the information in this article, do not make it obvious you are taking information from them but in very sensitive information give them credit."* As academics, where attribution is key, we found the creation of new content from third-party content, which may be copyrighted, as well as intentionally obscuring such usage, concerning.

Both agencies incorporated interviews into their input prompts. These interviews were presented as written logs from emails or chat applications, or as transcriptions of recorded conversations. While most interviews did not contain sensitive information, one instance included a complete WhatsApp conversation that disclosed a source's phone number and sensitive details about their children, including health conditions. This raises concerns about exposing such data to the LLM and through WildChat, without the source's consent or awareness.

In terms of the tasks users specified in their inputs, the vast majority of turns, 83.1%, are for article generation, 14.5% are headline generation, and a minority are article editing. We depict the distribution of task types in Appendix B.

How much do journalists in our study modify model outputs before publication? Figure 4 and Appendix E depict ROUGE-L distributions for prompts to GPT outputs and those outputs to published articles across all identified activity and for each agency, respectively. The median ROUGE-L score between the machine-generated drafts and published article text is 0.62 which indicates limited human editing before publication. As a point of reference, a ROUGE score of 0.5 is considered high in privacy and policy domains (Huang et al., 2023; MacLaughlin et al., 2020). Figure 1 depicts a specific use where the ROUGE-L score between the machine-generated draft and the published article is high at 0.71, indicating high overlap with the source text. In Figure 2, the ROUGE-L scores between the drafts and published articles are slightly lower but still relatively high. We also found a

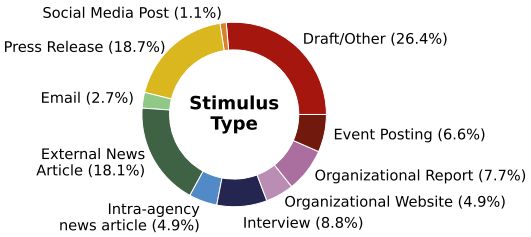


Figure 3: The distribution of different input stimulus material types provided by journalists to LLMs, over the WildChat conversations that are matched to online articles from the two identified news agencies.

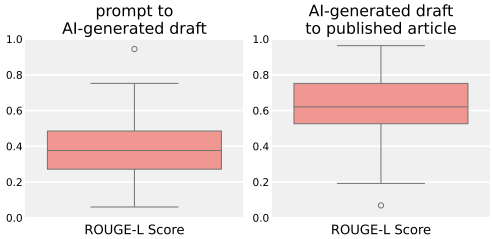


Figure 4: Box plots of ROUGE-L scores between outputs from users' article generations prompts and LLM response (left) as well as the LLM-generated article and the corresponding published article (right).

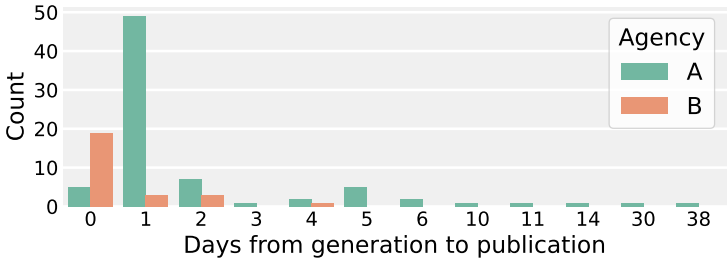


Figure 5: Distribution of days from generation (based on date in WildChat) to publication (based on published article's date) for published articles matched to turns of article generation in WildChat.

separate query (see Appendix C) to generate the headline for this article, and the user publishes the machine-generated headline with no editing.

What is the prompt to publication time? For Agency A, the majority (48/79) of verified articles were published one day after the recorded activity in WildChat. Similarly, the majority (19/32) of verified articles for Agency B were published on the same day as the recorded activity. The days of human editing time before publication seen in Figures 1 and 2 are consistent with these findings, as illustrated in Figure 5.

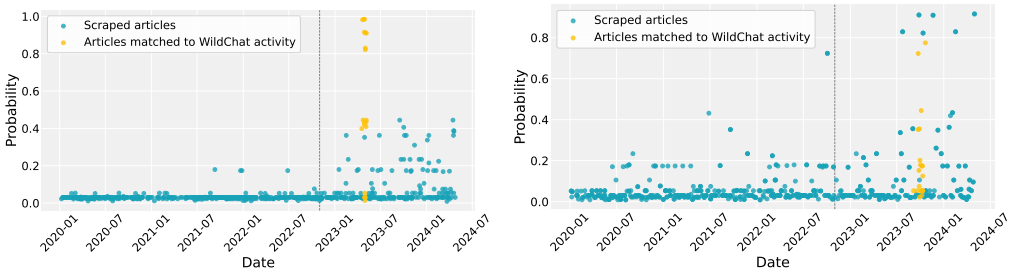
Are there more LLM-generated articles? To study the temporal trends in using LLMs for writing articles, we scrape articles from Agencies A and B, between January 1, 2020, and April 15, 2024, subsample 585 articles from each agency uniformly and utilize GPTZero to detect which articles might be machine-generated. Our findings, depicted in Figure 6, suggest that there are likely many additional published articles that may have been generated using ChatGPT. Furthermore, we observe a noticeable increase in articles with higher probabilities of machine generations following ChatGPT's release in November 2022.

4 Conclusion

We investigate the use of commercial LLMs in journalism by analyzing conversations from the WildChat dataset and matching them to published articles online. Our findings reveal the use of potentially unethical material to generate articles, limited human oversight on model outputs before publication, and the use of LLMs by the identified agencies beyond the scope of WildChat. These results suggest continuous, increasing generative AI use for news generation and can serve as additional motivation for the co-evolution of guidelines for responsible AI journalism in collaboration between the computer science and journalism communities.

5 Social Impacts Statement

This study is motivated by the recognition that generative AI capabilities are rapidly evolving and their potential impacts on journalism, both beneficial and detrimental, call for thoughtful and thorough



(a) Agency A

(b) Agency B

Figure 6: GPTZero probability scores for randomly selected articles from our scrape as well as the published articles matched to WildChat. The grey vertical line indicates November 2022, when ChatGPT was released.

research. Our goal is to provide insights that can inform guidelines and strategies for the responsible use of AI in journalism. We explicitly do not seek to speak to how *all* news agencies use or might use LLMs. Rather, through these case studies, we seek to develop an informed understanding of how *some* news agencies have used LLMs. Whether their practices are unique to them, or widespread, an understanding of their practices can contribute to a broader discussion about current and potential future uses of LLMs in journalism.

Positionality. This work stems from a collaboration between researchers in the NLP and computer security research communities. Consequently, our focus is on the potential risks and harms associated with the use of LLMs in journalism. We acknowledge that there are potential benefits of LLMs in journalism that other perspectives and studies may highlight.

Anonymization. We hypothesize that Agencies A and B are not unique cases of news organizations using generative AI, but rather examples of a broader phenomenon that happen to be present in LLM system usage. Furthermore, the intent of this work is not to "call out" specific agencies or individual users, but to investigate and better understand the actual use of generative AI in journalism. Accordingly, we have anonymized the identities of the agencies in this paper.

Although we anonymize the agencies' names and censor identifying information, we provide metadata and verbatim text that could be used to identify them. We include this information for scientific rigor and thoroughness. However, we do not encourage or condone the use of this data beyond our stated purpose of understanding AI usage in journalism. Our intent is to advance knowledge, not target or harm any individuals or organizations.

Mitigation of privacy concerns. Given the sensitive nature of the conversations identified, we notified the WildChat dataset authors of our findings. In response, they removed the identified conversations from the publicly available version of the dataset.

Acknowledgements

This work was supported in part by NSF Award #2205171. We thank Yuntian Deng, Yanai Elazar, Susan McGregor, Stefan Milne, Melanie Sclar, and Maria Antoniak for insightful discussions and feedback.

Limitations

We acknowledge three main limitations in our study: (1) We are constrained by the WildChat dataset itself, which limits our ability to generalize results beyond the two identified agencies and only allows examination of interactions with GPT-4 and GPT-3.5. (2) There may be interactions we were unable to identify within the dataset, though we conducted additional searches and manual reviews to mitigate this. (3) Concerns exist regarding the effectiveness of detecting machine-generated text (Solaiman et al., 2019; Chakraborty et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023), and we recognize that GPTZero, the tool used for detection in this study, does not provide ground truth predictions.

References

Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Comput. Surv.*, 50(3), aug 2017. ISSN 0360-0300. doi: 10.1145/3054926. URL https://doi.org/10.1145/3054926.

David Armstrong, Ann Gosling, and John Weinman. The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study. *Sociology-the Journal of The British Sociological Association - SOCIOLOGY*, 31:597–606, 08 1997. doi: 10.1177/0038038597031003015.

Associated Press. AP'S Local News AI Intiative. URL https://www.ap.org/solutions/artificial-intelligence/local-news-ai/. Accessed: 2024-05-13.

Suhib Y. Bdoor and Mohammad Habes. *"Use Chat GPT in Media Content Production Digital Newsrooms Perspective"*, pp. 545–561. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-52280-2. doi: 10.1007/978-3-031-52280-2_34. URL https://doi.org/10.1007/978-3-031-52280-2_34.

Souradip Chakraborty, A. S. Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the Possibilities of AI-Generated Text Detection. *ArXiv*, abs/2304.04736, 2023. URL https://api.semanticscholar.org/CorpusID:258048481.

Hannes Cools and Nicolas Diakopoulos. Writing guidelines for the role of AI in your newsroom? Here are some, er, guidelines for that, July 2023. URL https://www.niemanlab.org/2023/07/writing-guidelines-for-the-role-of-ai-in-your-\newsroom-here-are-some-er-guidelines-for-that/. Accessed: 2024-09-04.

Nicholas Diakopoulos, Hannes Cools, Charlotte Li, Natali Helberger, Ernest Kung, Aimee Rinehart, and Lisa Gibbs. Generative AI in Journalism: The Evolution of Newswork and Ethics in a Generative Information Ecosystem, 04 2024.

Sam Dolnick and Zach Seward. Principles for Using Generative A.I. in The Times's Newsroom, March 2024. URL https://www.nytco.com/press/principles-for-using-generative-a%E2%80%A4i%E2%80%A4-in-the-timess-newsroom/. Accessed: 2024-09-04.

Gregory Gondwe. CHATGPT and the Global South: how are journalists in sub-Saharan Africa engaging with generative AI? *Online Media and Global Communication*, 2(2):228–249, 2023. doi: doi:10.1515/omgc-2023-0023. URL https://doi.org/10.1515/omgc-2023-0023.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. "Privacy Implications of Retrieval-Based Language Models". In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14887–14902, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.921. URL https://aclanthology.org/2023.emnlp-main.921.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Sogaard. Copyright Violations and Large Language Models. *ArXiv*, abs/2310.13771, 2023. URL https://api.semanticscholar.org/CorpusID:264426289.

Tom Kent. An ethical checklist for robot journalism, February 2015. URL https://medium.com/@tjrkent/an-ethical-checklist-for-robot-journalism-1f41dcbd7be2. Accessed: 2024-09-04.

Tharindu Kumarage, Amrita Bhattacharjee, Djordje Padejski, Kristy Roschke, Dan Gillmor, Scott Ruston, Huan Liu, and Joshua Garland. J-Guard: Journalism Guided Adversarially Robust Detection of AI-generated News, 2023.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of chatgpt on AI conference peer reviews. *ArXiv*, abs/2403.07183, 2024a. URL https://api.semanticscholar.org/CorpusID:268363317.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. Mapping the increasing use of LLMs in scientific papers. *ArXiv*, abs/2404.01268, 2024b. URL https://api.semanticscholar.org/CorpusID:268857142.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Ansel MacLaughlin, John Wihbey, Aleszu Bajak, and David A. Smith. Source Attribution: Recovering the Press Releases Behind Health Science News. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):428–439, May 2020. doi: 10.1609/icwsm.v14i1.7312. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7312.

Janice M. Morse. "Perfectly Healthy, but Dead": The Myth of Inter-Rater Reliability. *Qualitative Health Research*, 7(4):445–447, 1997. doi: 10.1177/104973239700700401. URL https://doi.org/10.1177/104973239700700401.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. "On the Risk of Misinformation Pollution with Large Language Models". In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1389–1403, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.97. URL https://aclanthology.org/2023.findings-emnlp.97.

Partnership on AI. AI Adoption for Newsrooms: A 10-Step Guide. URL https://partnershiponai.org/ai-for-newsrooms/. Accessed: 2024-05-14.

John V. Pavlik. Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, 78(1):84–93, 2023. doi: 10.1177/10776958221149577. URL https://doi.org/10.1177/10776958221149577.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-Generated Text be Reliably Detected? *ArXiv*, abs/2303.11156, 2023. URL https://api.semanticscholar.org/CorpusID:257631570.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models. *ArXiv*, abs/1908.09203, 2019. URL https://api.semanticscholar.org/CorpusID:201666234.

Edward Tian and Alexander Cui. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods. *GPTZero*, 2023. URL https://gptzero.me/.

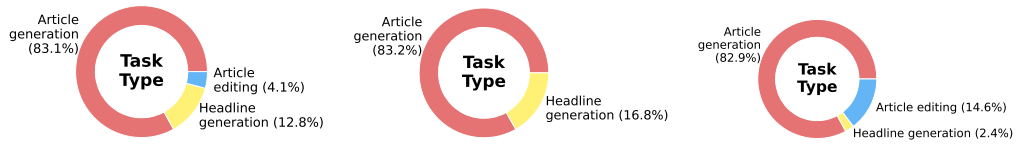
Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26, 2023. ISSN 1833-2595. doi: 10.1007/s40979-023-00146-z. URL https://doi.org/10.1007/s40979-023-00146-z.

Robin Whittemore, Susan K. Chase, and Carol Lynn Mandle. Validity in Qualitative Research. *Qualitative Health Research*, 11(4):522–537, 2001. doi: 10.1177/104973201129119299. URL https://doi.org/10.1177/104973201129119299. PMID: 11521609.

John Wihbey. AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?, April 20 2024. URL https://ssrn.com/abstract=4805026. Available at SSRN.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 4(2):100211, 2024. ISSN 2667-2952. doi: https://doi.org/10.1016/j.hcc.2024.100211. URL https://www.sciencedirect.com/science/article/pii/S266729522400014X.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.



(a) Aggregated over both agencies (b) Agency A (c) Agency B

Figure 7: The distribution of different task types across turns from verified Wildchat journalist-LLM conversations.

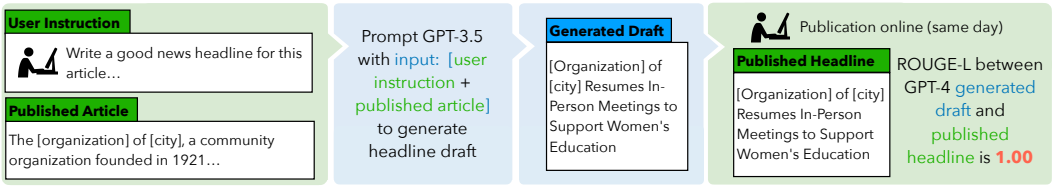


Figure 8: A case study of headline generation, extending from Figure 2. The '[' and ']' symbols denote portions of the text that have been replaced to minimize identifiability.

A Extended Related Work

Research at the intersection of journalism and generative AI is still in its early stages. However, work has begun to investigate its potential and actual uses in the field. As part of the Associated Press's Local News AI Initiative, Diakopoulos et al. surveyed 292 individuals in the news industry about their use and opinions of generative AI in newsrooms. The survey revealed that generative AI is already being used for tasks, such as content production, and changing workflows and role definitions in the newsroom. Gondwe (Gondwe, 2023) investigated the use of ChatGPT by journalists in sub-Saharan Africa, finding that the system's training on a non-representative African corpus limits its utility for the studied population.

Bdoor and Habes (Bdoor & Habes, 2024) experimented with using GPT to generate news content and discussed the trade-offs around its adoption in the newsroom. Pavlik (Pavlik, 2023) 'co-authored' an essay with ChatGPT to demonstrate both the capacity and limitations of generative AI in journalism and media education.

To address concerns about AI-generated news media, Kumarage et al. (Kumarage et al., 2023) developed J-Guard, a framework for directing supervised AI-generated text detectors to identify AI-generated news articles.

B Task Type

Figure 7 provide breakdown of task types.

C Additional Case Study

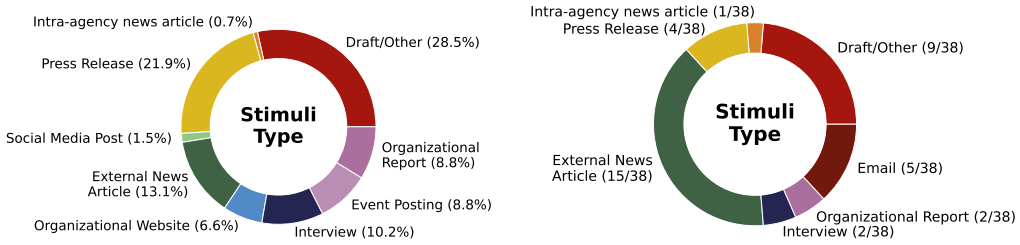
Figure 8 is an extension of Figure 2 where the user generates a headline for the machine-drafted article.

D Stimuli

Table 4 includes the top five most frequent combinations of stimuli used in a turn of article generation. Figure 9 visualize the distributions of stimuli types for Agencies A and B, respectively.

Stimuli type(s)	Agency A frequency	Stimuli type(s)	Agency B frequency
Draft or other	16	External news article	13
Press release	16	Draft or other	5
Organizational report	11	Press release	3
External news article	6	Email	3
Draft or other, external news article	5	Organizational report	2

Table 4: Top five most frequent combinations of stimuli types used in an individual turn of article generation.

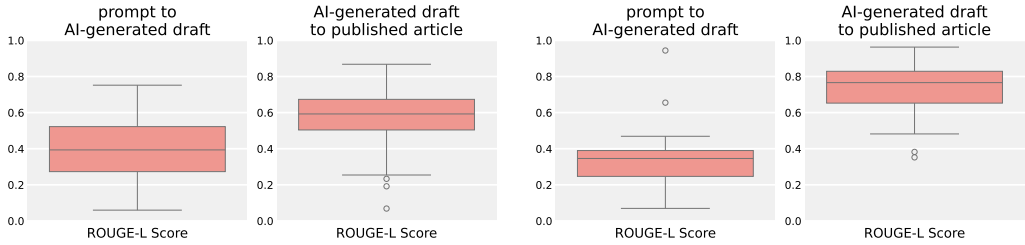


(a) Agency A (b) Agency B

Figure 9: The distribution of input stimuli types over the verified Wildchat journalist-LLM interactions for both agencies.

E ROUGE-L

Figure 10 depicts ROUGE-L distributions for prompts to GPT outputs and those outputs to published articles for each agency.



(a) Agency A (b) Agency B

Figure 10: ROUGE-L scores for prompt to machine-generated output (left) and that output to published article text (right) for both agencies.