Breaking News: Case Studies of Generative AI's Use in Journalism

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno
Franziska Roesner, Niloofar Mireshghallah
University of Washington
{nbrigham, chongjiu, 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

LLMs present the opportunity to increase productivity in newsrooms and 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 a threat to journalistic integrity (Wihbey, 2024).

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 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.

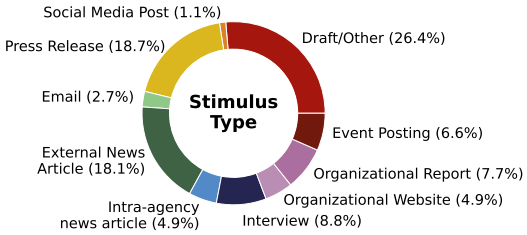


Figure 1: The distribution of different input stimulus material types provided by journalists to LLMs, over the WildChat conversations that are matched to online articles from the two identified news agencies.

To achieve this, we probe the 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 1. Inputs include articles from other agencies or private conversations, real case examples of which are depicted in Figures 2 and 3 (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 and Cui, 2023), a state-of-the-art commercial LLM-generation detector.

Our findings raise concerns about the level of diligence and journalistic integrity, as we find sub-

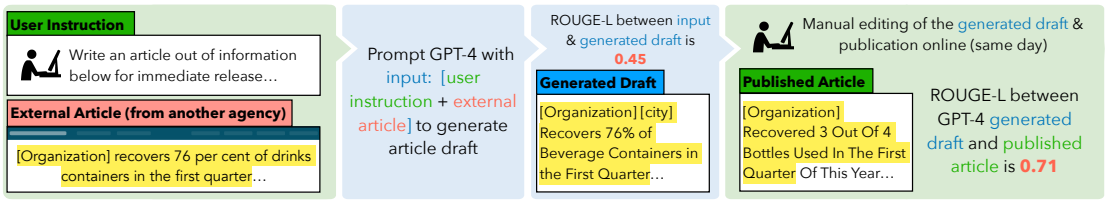


Figure 2: Case study of a single-turn journalist-LLM interaction for article generation where an external article by another agency is used as the input stimulus. The generated draft is published by the journalist with little modification as the ROUGE-L between the published article we identified and the generation is 0.71. The '[' and ']' symbols denote portions of the text that have been replaced to minimize identifiability.

stantial 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. Alarmingly, 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 guidelines surrounding LLM use by journalists, and also 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.

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.

We acknowledge that this process may result in false negatives, missing cases of journalistic activity in WildChat that we did not identify. For reproducibility of results we have uploaded the list of verified WildChat conversation IDs along with our submission. (See ethical considerations 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 (Whit-

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

[2] `lakshyakh93/deberta_finetuned_pii`

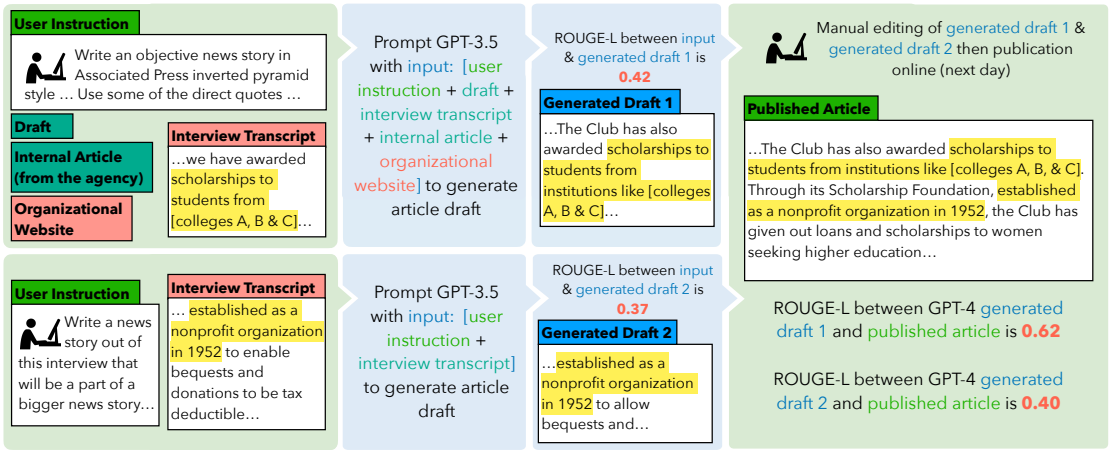


Figure 3: Case study of a muli-turn article generation using multiple stimuli, including an internal article from the same agency and an interview transcript. The '[' and ']' symbols denote portions of the text that have been replaced to minimize identifiability.

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

temore 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 used.

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 and 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 2, a case of a single-turn article generation (from Agency B) and Figure 3, a multi-turn process of article generation (from Agency A). Below we discuss our observations from the two agencies. Appendices D and E provide more fine grained results for Sections 3.1 and 3.2, respectively.

3.1 What do journalists prompt LLMs with?

Figure 1 shows the distribution of input stimulus material. 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. Other types, like interview transcripts, are generated internally but may contain information and material from external sources, including potentially sensitive information from or about those sources.

In Figure 2, 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*

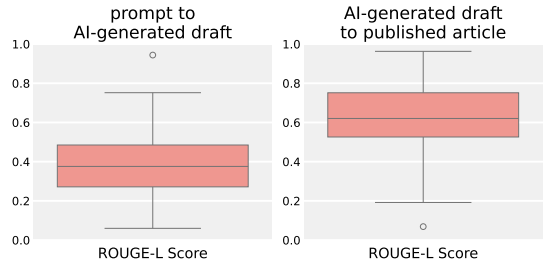


Figure 4: Box plots of ROUGE-L scores between outputs from users' article generations prompts and LLM response (left) as well as the LLM-generated article and the corresponding published article (right).

taking information from them but in very sensitive information give them credit." Conversely, Figure 3 shows Agency A employing a mix of internal and external stimuli.

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 provide more thorough results for the identified task types in Appendix B.

3.2 How much do journalists modify model outputs before publication?

The median ROUGE-L score between the machine-generated drafts and published article text is 0.62 (see Figure 4) 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 2 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 3, the ROUGE-L scores between the drafts and published articles are slightly lower but still relatively high. We also found a separate query (see Appendix C) to generate the headline for this article, and the user publishes the machine-generated headline with no editing.

3.3 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 2 and 3 are consistent with these findings, as illustrated in Figure 5.

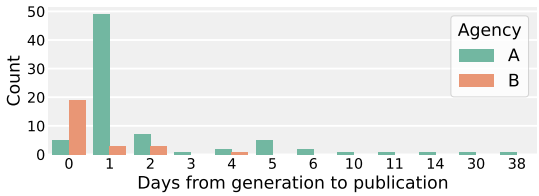
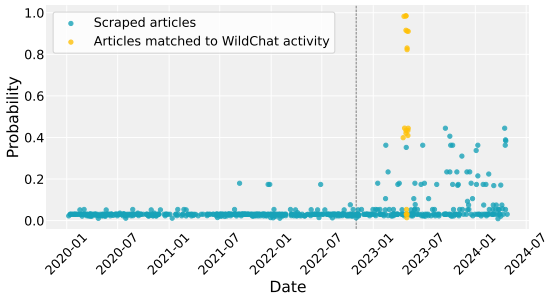
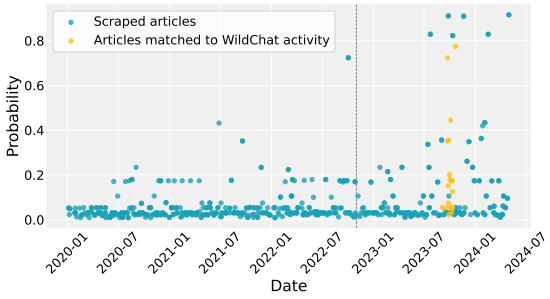


Figure 5: Distribution of days from generation (based on date in WildChat) to publication (based on published article's date) for published articles matched to turns of article generation in WildChat.



(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.

3.4 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 necessitate guidelines for responsible AI in journalism.

Acknowledgements

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

Limitations

We acknowledge that we are constrained by the WildChat dataset itself. We can only identify and study journalist-AI interactions included in the dataset, which limits our ability to generalize our results beyond the two agencies identified. Furthermore, given the nature of WildChat, we can only examine interactions with GPT-4 and GPT-3.5.

There may also be interactions that we were unable to identify within the dataset. To mitigate this, we performed additional searches for other cases of article generation, headline generation, and article editing, using relevant keywords and prompting GPT to annotate instances of non-fictional journalistic activity. Upon manually reviewing these results, we did not identify any additional conversations.

Additionally, there are concerns and questions regarding the effectiveness and feasibility of detecting machine-generated text (Solaiman et al., 2019; Chakraborty et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023). We recognize that GPTZero, the tool used for detection in this study, does not provide ground truth predictions.

Ethical Considerations

This study is motivated by a recognition that generative AI capabilities are rapidly evolving and their potential impacts on journalism call for thoughtful, thorough research. Our goal is to provide insights that can help shape guidelines and policies governing the responsible use of AI in this domain.

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 the WildChat dataset. 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 as well as recognize that, even without it, others could independently identify the news agencies through WildChat. 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.

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. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Comput. Surv.*, 50(3).

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

Associated Press. AP'S Local News AI Intiative. Accessed: 2024-05-13.

Suhib Y. Bdoor and Mohammad Habes. 2024. *"Use Chat GPT in Media Content Production Digital Newsrooms Perspective"*, pages 545–561. Springer Nature Switzerland, Cham.

Souradip Chakraborty, A. S. Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the Possibilities of AI-Generated Text Detection. *ArXiv*, abs/2304.04736.

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

Gregory Gondwe. 2023. 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.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. 2023. "Privacy Implications of Retrieval-Based Language Models". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14887–14902, Singapore. Association for Computational Linguistics.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Sogaard. 2023. Copyright Violations and Large Language Models. *ArXiv*, abs/2310.13771.

Tharindu Kumarage, Amrita Bhattacharjee, Djordje Padejski, Kristy Roschke, Dan Gillmor, Scott Ruston, Huan Liu, and Joshua Garland. 2023. J-Guard: Journalism Guided Adversarially Robust Detection of AI-generated News. *Preprint*, arXiv:2309.03164.

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. 2024a. Monitoring AI-modified content at scale: A case study on the impact of chatgpt on AI conference peer reviews. *ArXiv*, abs/2403.07183.

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. 2024b. Mapping the increasing use of LLMs in scientific papers. *ArXiv*, abs/2404.01268.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ansel MacLaughlin, John Wihbey, Aleszu Bajak, and David A. Smith. 2020. 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.

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

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. "On the Risk of Misinformation Pollution with Large Language Models". In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Partnership on AI. AI Adoption for Newsrooms: A 10-Step Guide. Accessed: 2024-05-14.

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

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *ArXiv*, abs/2303.11156.

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

Edward Tian and Alexander Cui. 2023. GPTZero: Towards detection of AI-generated text using zero-shot and supervised methods. *GPTZero*.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1):26.

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

John Wihbey. 2024. AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge? Available at SSRN.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 4(2):100211.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*.

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.

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 and 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

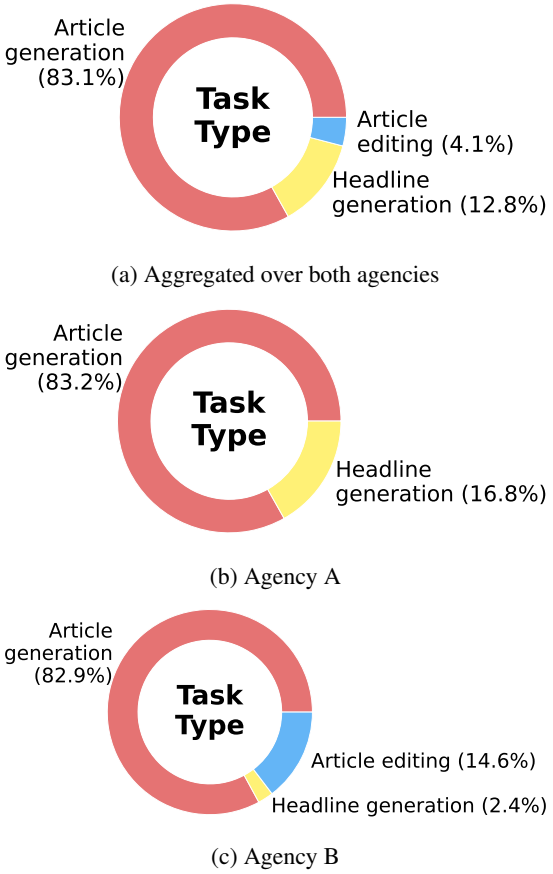
Table 2 and Figure 7 provide breakdown of task types.

C Additional Case Study

Figure 8 is an extension of Figure 3 where the user generates a headline for the machine-drafted article.

D Stimuli

Table 3 provides stimuli types, descriptions, and frequencies. Table 4 includes the top five most frequent combinations of stimuli used in a turn



(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.

of article generation. Figure 9 visualize the distributions of stimuli types for Agencies A and B, respectively.

E ROUGE-L

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

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.

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.

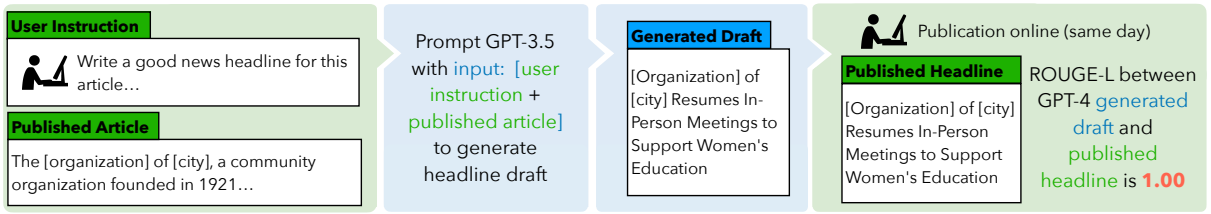
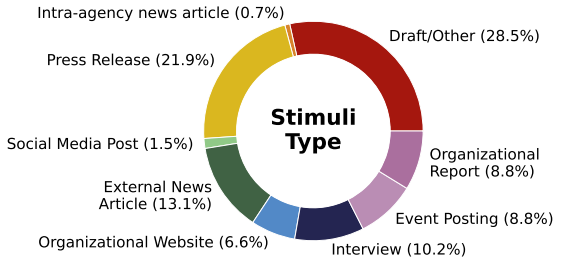
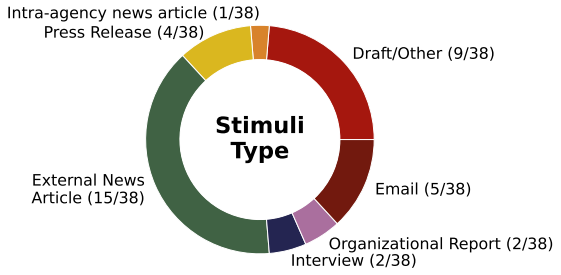


Figure 8: A case study of headline generation, extending from Figure 3. The '[' and ']' symbols denote portions of the text that have been replaced to minimize identifiability.

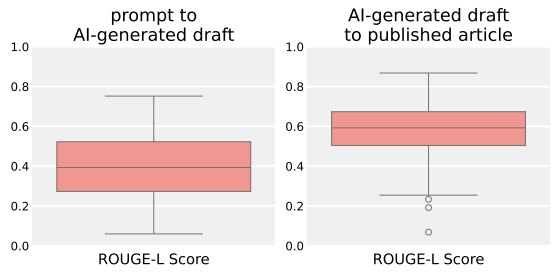


(a) Agency A

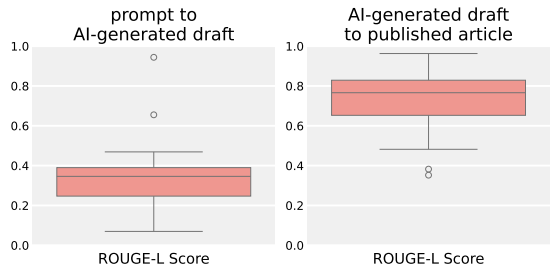


(b) Agency B

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



(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.