Methods for Evaluating AI Tools in Journalism

CNTI's Global AI & Journalism Research Group Briefing #5 | October 2026

A note from CNTI Executive Director, Amy Mitchell

AI in News. In even just the past two years, what that means has expanded in myriad ways. So too have the ways news providers use AI tools. Early uses of AI by news providers tended to focus on direct content outputs, with the goal and evaluation metric of 100% accuracy. Now, however, AI is used to aid staff in far more complex ways. So how can a provider decide what tool or set of tools works best when? And how can research and guidelines help?

As with most outside tools, no one set of criteria or measures works across all use cases. Rather, determining the best AI tool and customization techniques depends on several factors such as the problem one is seeking to address, the goals in addressing the problem, internal capacity and sustainability. Even between seemingly similar tasks, the optimal AI tool will differ based on how the output will be used, the structure of the dataset and more. Consider the use of an AI tool to cull through a complex dataset. What is the desired output, and how will it be used by staff? How much internal coding will be applied? How constant or dynamic is the dataset itself? The answer to each of these questions will help determine the optimal tool or mix of tools. The journalism industry needs to develop the right set of questions about the use case and then, and only then, the ways to evaluate the fit.

Given this complexity, how can research be most helpful? Perhaps by 1) examining a wide range of successful AI uses, both among teams within the journalism industry and teams outside the journalism industry, and synthesizing the kinds of questions the teams asked, the systems and workflows they set up and the measures they used to evaluate strength and success, 2) learning from other industries that have established techniques for evaluating AI tools — an approach rarely used in journalism, 3) designing new, forward-looking research on questions such as how well coding agents understand the responses in vibe coding, the effectiveness of the interface, the effectiveness of human reviews and correction techniques to name just a few that have surfaced. And there is also the question of how to adjust evaluation steps as models update. These are not easy projects and may require a level of access into providers' internal workings, but this deeper and forward-looking approach will bring longer-term value to the industry and the public it serves. Shining a light on areas where we need to build stronger research is a central goal of our Global AI and Journalism Research Working Group and warrants attention in this brief.

Methods for Evaluating AI Tools in Journalism

This briefing is a bit different from others developed by CNTI's Global AI and Journalism Research Working Group. Rather than exploring a particular use of AI, our goal is to understand what the current research says — and misses — about how news providers evaluate AI tools across different uses.

Despite the breadth of research on the uses of AI in journalism, little attention has been given to methods for evaluating AI tools. Further, the research that does exist does not address the business (and revenue) aspects of journalism. This briefing starts with research from fields outside of journalism, and identifies practices that are transferable to journalism, highlights journalism-specific research, and concludes with recommendations for further research.

*

This is the fifth in a series of reports from the AI and Journalism Research Working Group convened by the Center for News, Technology & Innovation (CNTI). The working group currently consists of 23 cross-industry members from around the world, bringing research, journalism and technology expertise to the discussions.

The goal of the working group is to offer succinct summaries of global research on specific topics at the intersection of journalism and AI. Each quarter, the working group will synthesize the state of research across two to three topics for journalism practitioners, researchers and industry leaders around the world, focusing on actionable recommendations for journalism — not other fields that are concerned with AI.

In each report, we lay out the general findings of the research to date, suggested considerations and/or actions for practitioners and areas where more or new research is needed. This report was prepared by the research and professional staff of CNTI in partnership with several external contributors who collectively authored this briefing. If you have ideas or research findings that are important for CNTI and the working group to include, please email them to info@cnti.org.

What do we mean by "AI"?

This report uses the OECD definition: *"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."*

Wherever possible, we try to use specific terms rather than "AI" to avoid conflation or confusion. Journalism has been adopting forms of automation for more than 50 years,[1] but widespread use of the term "AI" is more recent — and may include both newer technologies and those that have been in use for quite some time.

Research we reviewed outside journalism suggests:

- **Evaluating AI tools is more complex than feature comparison** across fields.[2] Even technologists and engineers widely acknowledge the difficulties.
  - Most practices for evaluating AI tools start by looking at outputs because it is functionally impossible to examine the inner workings of the AI tools or the underlying models.
  - By design, AI systems are nondeterministic: the same input can lead to different outputs,[3] which means evaluations often cannot be replicated.
  - Superficial differences between prompts can have outsize effects on outputs.[4]
  - Many AI tools were developed by and for a relatively small user base, and do not work equally well across the full spectrum of human variation.[5]
- **Standardized evaluations (benchmarks and scorecards) aren't as helpful as one might hope.** While these evaluations exist across fields, they tend to be abstract and infrequently align well with specific use cases.[6]
  - Metrics like accuracy and speed vary between different specific tasks such as developing headlines, captioning images, summarizing meetings or categorizing datasets.
  - Computer scientists have attempted to create journalism benchmarks. Unfortunately, several of these efforts suffer from limited understanding of day-to-day journalistic work.[7]
- Decision-makers benefit from **starting with goals, not with tools.**
  - Methods for evaluating an AI tool are stronger when they align with the goals for using the tool.[8] The evaluation measures must also tie to specific institutional needs. In the journalism context, that would mean starting from journalistic values.[9]
  - It's easier for people to suggest solutions than to reflect on how they do things, but both are important sources of data. Instead of relying solely on the words people use to describe the tool they want, observing actual user behaviors will enrich evaluation.[10] That's just as true for staff as for audiences. Including user researchers in the process from the beginning can help.
- Evaluating AI tools needs to be **an ongoing practice, not a one-time decision**.[11] Small changes to underlying models can impact outputs in unpredictable ways.

Journalism-specific research we reviewed suggests:

- Several recent studies are **converging on a set of dimensions** that are relevant across many different journalism uses of AI tools.[12] How to weigh them will depend on the intended use of the AI tool, and none of these dimensions are binary. This remains a baseline, not an exhaustive list.
  - **Accuracy**: How frequently is the tool accurate? How much work is required to verify its outputs?
  - **Neutrality & Objectivity**: Are there underlying biases in results? How often does this occur, and how can users ensure that it doesn't?
  - **Transparency:** Does the tool explain the logic behind its outputs, and are users aware of the limitations?
  - **Autonomy & Independence**: How will third-party updates impact journalists' work? Would losing access to particular models or APIs impact their process? How can they ensure they are accountable for their work product?
  - **Efficiency**: Does it save time and energy, increase scale or both? Do those gains offset any extra work required?

Lessons from Technical Fields

What is AI evaluation, and why is it difficult?

Evaluation is the practice of determining whether a tool is good enough for a specific purpose in a specific context.

We know more about developers than users.

Developers building their own AI tools evaluate them using a range of methods, and there is research on those methods. At the same time, anyone considering adopting new AI tools also needs to evaluate them to decide if they will work in their specific context. To date, there has been much less research on the evaluation practices of end users and institutional decision-makers, compared with developers.[13]

It's difficult to evaluate the tools directly.

Across fields, people evaluate AI tools by examining outputs because it is almost never possible to examine the inner workings of the tools directly. Evaluating AI tools is especially hard because they are nondeterministic by design.[14] That means they won't give the same results every time, and errors may not follow a predictable pattern. Large language models (LLMs) in particular are sold as "general-purpose" technologies that can be adapted for a wide range of use cases that may not have a single correct answer.[15] That makes context-specific evaluation all the more important. In practical contexts, evaluation should mean identifying the gap between what technology can do and what people actually need.[16]

Even major technology companies haven't reached a consensus.

One might expect large, well-resourced technology companies to set the evaluation standards for everyone else. After all, these companies have been working with AI tools longer than anyone else. But even they struggle to evaluate whether the quality of LLM tools is "good enough," especially for outward-facing work. A mixed-methods study of multiple product teams at Microsoft[17] corroborates major challenges to setting evaluation standards identified in other research, many of which are relevant to journalism:

- "Correct" or "good" answers are subjective and underspecified.
- Non-deterministic technology means results cannot be replicated exactly.
- Small model tweaks can create big changes to outputs.

The same study found several emerging solutions to these challenges that may be relevant to journalism. In response to changing evaluation metrics, Microsoft employees defined custom metrics through collaboration with domain experts, combined qualitative and quantitative metrics, and laid out clear rubrics. To counter nondeterminism, they rolled out small-scale releases and ran A/B tests so they could identify unanticipated issues before full release. They responded to changing models by automating scheduled evaluation at regular frequencies and establishing internal standards.

The study also cautions that while these evaluation practices are becoming standardized, there is little data about whether they are actually effective.

There are two approaches with complementary strengths and weaknesses.

Across fields, there are two primary evaluation methods in use:

- Benchmarks are, essentially, standardized tests. A benchmark consists of a large dataset, a series of tasks and a scoring procedure. Typically, benchmarks are simple so that outputs can be scored automatically, often by other machine learning models.
- Real-world testing, or human annotation, requires experts to review outputs manually and qualitatively. This form of evaluation measures the AI tool's judgments against human ones, but there are limits to the applicable scale.

The problems with benchmarks and scorecards

Computer science and engineering were the first fields to evaluate AI tools, and they offer lessons for other fields. However, the way they frame the problem can seem opaque and/or irrelevant to other fields.

Benchmarks are inexpensive and easy to compare.

These fields rely heavily on benchmarks. This approach has some clear advantages. Most obviously, benchmarks make it easy to compare different models.[18] Many benchmarks can be scored automatically, which is relatively inexpensive.[19]

Takeaways for journalism

- General-purpose benchmarks and scorecards are unlikely to be informative for journalism purposes.
  - They don't always measure what they're supposed to measure.
  - The results don't always translate to real-world performance.
  - Developers can optimize for benchmark performance, leading to unrealistically positive results.
  - Few benchmarks attempt to measure bias or safety issues.
  - They tend to assume people are much less diverse than we actually are.
  - They don't provide much insight into why a tool succeeds or fails.
- Journalism-specific benchmarks could be valuable, but they may run into the same pitfalls as other benchmarks.

Benchmarks don't translate well to the real world.

But benchmarks have many weaknesses. First, they don't always measure what they're supposed to measure (construct validity).[20] Second, the results don't always translate to real-world performance (ecological validity).[21] For example, benchmarks can be sensitive to minor differences in prompts, such as formatting.[22] This is especially important in real-world settings. Third, benchmarks can be overly positive: it is possible for developers to optimize for benchmark performance rather than real-world performance. Many benchmarks remain static over time and rely on multiple-choice items with a single correct

answer. That means models can be trained directly on test items, and they often are.[23] Fourth, few benchmarks attempt to measure bias or safety issues.[24]

Current benchmarks assume people are culturally homogeneous. We're not.

Fifth, benchmarks tend to favor particular cultural and linguistic assumptions that are not directly related to task performance. A recent audit of 51 publicly released multilingual benchmarks found that nearly all of them fall short in terms of language coverage, representativeness, rigor and transparency.[25] Specifically, they overrepresent Indo-European languages and Latin script. That's important because language family and script type impact how AI models break up text, which in turn affects computation costs and speed. In other words, efficiency gains may vary based on the language being used. The same study also found that more than half the non-English content in the benchmark datasets was translated from English, not written in the language being tested.[26] That means benchmarks are less likely to be culturally representative. Moreover, only a few of the benchmarks are transparent about the linguistic composition of the training set.

Benchmarks don't explain why AI tools fail or how to improve them.

Sixth, typical benchmarks don't provide much insight into the reasons for success or failure.[27] They can't explain why models perform poorly or what changes could help them perform better. That also means benchmarks may have more value for academic publication than applied value. In fact, several studies have suggested that researchers are more likely to rely exclusively on benchmarks, compared with people making product or policy decisions.[28]

Decision-makers tend to see human evaluation as a necessary complement to benchmarks, given the shortcomings of benchmarks on their own.

The value of real-world testing

Takeaways for journalism

- Start from specific goals, not tools. If you know what you want to achieve, it's easier to develop custom metrics or rubrics.
- The strongest rubrics are developed by a mix of domain experts and technical experts.
- The best rubrics consider the specific cultural and linguistic context and the institution.
- Effective evaluation is ongoing, not one-time.

Real-world testing can be combined with benchmarks for more robust evaluation.

In contrast to benchmarking, real-world testing can help identify the maximum capabilities of new AI tools.[29] However, it cannot provide a sense of average performance or the range of typical behavior. That makes this approach on its own insufficient for decision-making.

A few additional studies have explored institutional decision-making about AI tools outside technical industries, largely through interviews.[30] These studies identify similar considerations and pitfalls. Domain expertise and context-

sensitive-tests complement benchmarks.[31] Across sectors, the following considerations came up repeatedly:[32]

- **Sociocultural concerns**. There are major disparities in linguistic performance, even within single models.[33] In some cases, customizing AI tools for particular languages and cultural contexts requires incorporating multiple models or providing extensive customization.[34] Domain expertise is critical to achieving this.
- **Institutional concerns**. Policies and compliance structures vary between organizations and sectors. Decision-makers should consider the impact on the institution and on other internal and external stakeholders.[35] If there is insufficient capacity to incorporate and verify domain expertise on an ongoing basis, deployment may be unsustainable.[36]
- **Technological concerns**. Flexibility and ongoing evaluation are essential so that workflows can accommodate third-party technical changes. Small-scale rollouts can serve as an additional form of evaluation to identify real-world issues.[37]

When decision-makers see AI use as the goal itself, it creates risk.

Two additional interview studies of organizational decision-makers suggest that they may not always be aware of their own limitations. The first study, which included participants from journalism alongside law, medicine and the public sector, found a need for increased AI literacy to ensure that decision-makers have more accurate perceptions of model capabilities and weaknesses, as well as tools to ensure that adoption does not create overreliance.[38] The second study included 25 interviewees. Only one was from the media, but the findings were consistent across sectors.[39] These researchers found that leaders across sectors overestimate the impact of adopting AI tools while underestimating the technical effort and risks involved. They also identified a lack of awareness of methods for identifying user needs and a failure to iterate and ideate before moving forward with AI tools. These challenges are related: spending more time ideating would help organizations identify costs and risks before building out a solution that isn't viable. Leaders may also be deceived by the ease of building prototypes, assuming this speed translates to full implementation, which is often untrue.

Both studies suggest that additional ideation and literacy about risks could improve AI evaluation and adoption practices across industries.

Core Values: Journalism-specific Decision-making

Three bodies of literature attempt to address the difficulties of evaluating LLMs and other AI tools for journalism use cases. Unfortunately, they are largely siloed and do not benefit from one another's insights.

Computer science researchers have tried to do this work without journalism domain experts, which has been flawed.

First, several computer science researchers have tried to develop LLM benchmarks specifically for journalism tasks. One study, which measured writing proficiency and adherence to journalistic ethics across five editorial tasks in written Chinese, found considerable ethical shortfalls.[40] For example, all the models they tested failed to identify headlines with potential social harm. Another study measured factuality, logical consistency, importance, readability, objectivity and journalistic style for English-language journalism.[41] But these two studies include only limited participation from domain experts, so they may not fully reflect real-world use cases. For example, one of these studies describes journalists' work as *"complex real-world web mining workflows."*[42] It is far from clear whether journalists would agree.

Researchers have shared practices from large news outlets with in-house data science teams, but it may be difficult for smaller organizations to apply their lessons.

At the other extreme, a small number of large news outlets have in-house data science teams that build custom AI tools. For example, Bloomberg News is applying its guidelines for rules-based automation to new tools, including an internal LLM trained to perform specific tasks, including headline suggestions and identifying sentiment in financial stories.[43] But even organizations of this size and complexity need to establish specific thresholds for "good enough" performance that balance a range of business needs. An ethnographic study of the BBC[44] found that new AI tools that are built or adapted in-house are first tested offline against current products, then evaluated by editors and finally A/B tested with end users. There is no single threshold or metric that determines "good enough" in those cases; instead, evaluation is always negotiated between data science, editorial and business stakeholders.

The most valuable studies to date are interdisciplinary and participatory, with a focus on eliciting values to reflect on, rather than specific metrics or thresholds.

A promising approach to evaluation includes journalists in the design process without assuming they have in-house technical expertise.[45] In these cases, technologists collaborate closely with domain experts to understand their needs and concerns. Two separate groups of researchers identified similar core values through this method, as well as similar questions to ask during evaluation:[46]

- **Accuracy**: AI tools can introduce falsehoods. How frequently is the AI tool accurate, and how much work is required to verify its outputs? The goal is not necessarily perfect accuracy, but "accurate enough" to be useful.

- **Neutrality & Objectivity:** AI tools may have underlying biases or may output language that breaks with journalistic norms of avoiding value judgments. How frequently does this occur, and how can users ensure that it doesn't?
- **Transparency:** Most AI tools do not provide fully explainable or replicable results. Does the AI tool explain the logic behind its outputs, and are users aware of the limitations? If so, is it relatively easy for users to edit the final product to meet their overall standards?
- **Autonomy & Independence:** Most use cases depend on access to third-party models. If models change or journalists lose access, will their work or processes be at risk? Moreover, the lack of transparency may mean that AI tools are subtly steering users' decisions. How can journalists ensure they are fully accountable for their work product? And how much control do journalists have over their data?
- **Efficiency:** Organizations adopt AI tools to save time and energy, increase scale or both. But using AI tools also requires additional time to prompt, verify and edit their outputs. Do the efficiency gains offset the additional labor?

These core values overlap and reinforce one another. For example, a tool that saves time in other ways may still be valuable at a relatively lower accuracy threshold. We have spoken with news organizations that avoid creating a seamless user experience because friction in the process encourages journalists to critically evaluate key decisions and produce higher-quality work.

Nascent efforts to develop journalism benchmarks may eventually make it easier to compare different tools' or processes' adherence to these core values.[47]

Evaluation is a process, not an answer.

These same researchers also identified the following best practices:

- Evaluations should be grounded in real-world practice and specific tasks.
- Practitioners should be involved at multiple stages.
- Evaluation procedures that build end-user literacy would be especially valuable. In particular, users may struggle to understand how LLMs work conceptually and why they provide false information.[48] Making it easier to understand the uncertainty of outputs could pay off.
- Customizable templates or "cookbooks" can help create replicable practices.

A third interdisciplinary research team[49] offers twelve questions that can be used to evaluate both external and in-house AI tools. We reproduce them here verbatim.

Generaliz-ability	Is the AI tool applicable to multiple contexts?	**Journalistic values**	Does the AI tool reflect journalistic values such as reporting accuracy, relevance, and diversity? Were journalists involved and consulted in the development process?
Costs	Are there open-source datasets, AI models, or software provided by, for instance, journalistic organizations that can be used to build or enhance the AI tool? Can a crowdsourced approach be applied to enhance the quality and quantity of the data used for training the AI tool?	**Independence**	To what extent do journalists and newsrooms own the AI tool they use (data, code, infrastructure)? How are journalists affected by changes to components that they do not own?
User-friendliness	Is the AI tool designed in a user-friendly way, making it useful for users with different levels of technical skills?	**Transparency**	Is the AI tool's source code publicly available and well-documented, making it easy to understand and open to scrutiny?
Accuracy	Does the AI tool consistently produce reliable and accurate results in line with journalistic standards? Are there technical or human safeguards in place for the AI tool that minimize erroneous results, and, to the extent possible, mitigate their impact?	**Data privacy**	Have precautions been taken to prevent or mitigate the risk of the model sharing sensitive information from users with actors external to the journalistic organization?

Perspectives

Working group member Gary Mundy writes:

Emerging evidence from newsrooms around the world suggests that adoption decisions are often shaped by quite different realities. A recent Thomson Media study of 20 news outlets found that interest in AI was driven primarily by staffing shortages, financial pressures, and the need to improve efficiency, rather than by formal innovation and evaluation strategies. We see this in our work with small newsrooms globally, where adoption typically occurs through informal experimentation rather than structured evaluation approaches. In these instances, adopters typically give greater weight to immediately visible ***efficiency*** *benefits such as cost, time saved and ease of use. They may not fully consider issues of* ***neutrality*** *and bias,* ***accuracy****,* ***transparency****, data and information security, or longer-term dependence on technology providers (that is,* ***independence****). These dimensions are currently difficult to systematically assess within resource-constrained newsrooms, even when these newsrooms recognize their importance.*

These experiences point to the need for evaluation approaches that reflect the conditions in which journalists actually use AI. Accuracy and efficiency remain important, but they are only part of the picture. Newsrooms also need to understand how tools perform in local languages, where editorial responsibility sits when AI is used, and whether staff can identify and manage errors when they occur. Dependence on external technology providers is another consideration, particularly for smaller organizations with limited influence over how products are developed or changed.

The challenge is to embed evaluation into the everyday practice of smaller newsrooms rather than treating it as a separate technical exercise. Approaches must be proportionate to the realities in which these newsrooms operate, with limited staff, time, resources and technical capacity. while still providing enough structure to support informed and accountable decisions about when, where and how AI can improve performance.

Where More Research would be Helpful

Evaluation remains an emerging area of research and practice. Some of the most important gaps we identified for the field of journalism are:

- **Operational uses**: There is very little research exploring uses of AI tools in journalism organizations for revenue, staffing and other operational uses. In some cases, research from other fields may be transferable. For example, using AI tools to target advertisements may have fewer field-specific constraints than newsgathering uses. However, more research in these areas would provide a more complete understanding of AI use in journalism.
- **Evaluating vibe coding tools and other support uses**: To date, there is a near-total absence of research on vibe coding in journalism. Vibe coding tools may be more straightforward to evaluate than other kinds of tools: they need to consistently output executable code that passes standard quality assurance checks. But in the absence of research, it remains difficult to understand how vibe coding may be changing the types of outputs that journalists can produce.
- **Additional journalistic values and concerns**: The emerging consensus lays out a baseline, but it is not yet comprehensive.
- **Specific methods to evaluate journalistic values**: While there is a growing consensus about which journalistic values need to be evaluated, there is still a great deal of work to do on how to best evaluate them.
- **Scalable evaluation methods**: There remains a need to develop scalable methods that can measure best-case performance, average performance and variation along a range of context-specific metrics (e.g., accuracy, transparency, etc.).
- **Understanding smaller organizations**: Much of the research on AI decision-making in journalism has looked at larger organizations with dedicated data science or AI teams. But the majority of news organizations are small enough that they need to rely primarily on off-the-shelf tools. Additional research could shine light on how they decide which AI tools to use and when.
- **Multilingual evaluation and multilingual use cases**: Our 2025 briefing on translation and transcription identified downstream effects of gaps between languages which remain relevant here.
- **Efficacy of decision-making practices**: Several of the included papers describe current decision-making practices, but do not address whether these practices are actually effective.
- **More collaboration**: Research on AI evaluation remains largely siloed. Collaboration between researchers from different disciplines and industries could greatly benefit practitioners.

References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A Primer for Evaluating Large Language Models in Social-Science Research. *Advances in Methods and Practices in Psychological Science*, *8*(2), 25152459251325174. https://doi.org/10.1177/25152459251325174

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which Humans? https://doi.org/10.31234/osf.io/5b26t

Baack, S., Bohacek, M., & Buschek, C. (2026). Unsteady Metrics and Benchmarking Cultures of AI Model Builders. *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency, FAccT '26,* 840–872. https://doi.org/10.1145/3805689.3812240

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 39:1-39:45. https://doi.org/10.1145/3641289

Chien, Y.-C., Wang, K.-D., Wang, W.-Y., & Peng, W.-C. (2025). *NEWSAGENT: Benchmarking Multimodal Agents as Journalists with Real-World Newswriting Tasks* (arXiv:2509.00446). arXiv. https://doi.org/10.48550/arXiv.2509.00446

Cifliku, B., & Heuer, H. (2026). They Think AI Can Do More Than It Actually Can: Practices, Challenges, & Opportunities of AI-Supported Reporting In Local Journalism. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*, 1–20. https://doi.org/10.1145/3772318.3791130

Cordero, J. M., Henn, T., Holtel, F., Sánchez Gómez, J. Á., Arenas, D., Šipka, A., & Vollmer, S. (2025). Developing Effective and Value-Aligned AI Tools for Journalists: 12 Critical Questions to Reflect upon. *Journalism Practice*, *0*(0), 1–20. https://doi.org/10.1080/17512786.2025.2465894

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., & Fernandez-Llorca, D. (2025). Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(1), 850–864. https://doi.org/10.1609/aies.v8i1.36595

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Supryadi, Yu, L., Liu, Y., Li, J., Xiong, B., & Xiong, D. (2023). *Evaluating Large Language Models: A Comprehensive Survey* (arXiv:2310.19736). arXiv. https://doi.org/10.48550/arXiv.2310.19736

Hall, A., & Goldfarb, R. (2026, May 21). Introducing NewsBench [Substack newsletter]. *Forum AI*. https://forumai.substack.com/p/introducing-newsbench

Hardy, A., Reuel, A., Jafari Meimandi, K., Soder, L., Griffith, A., Asmar, D. M., Koyejo, S., Bernstein, M. S., & Kochenderfer, M. J. (2025). More than Marketing? On the Information Value of AI Benchmarks for Practitioners. *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, 1032–1047. https://doi.org/10.1145/3708359.3712152

Jaźwińska, K., & Chandrasekar, A. (2025, September 18). Journalists need their own benchmark tests for AI tools. *Columbia Journalism Review*. https://www.cjr.org/analysis/journalists-need-their-own-benchmark-tests-for-ai-tools.php

Kapoor, S., Kirgis, P., Schwartz, A., Rabanser, S., Allaire, J. J., Bommasani, R., Coppock, H., Dubois, M., Hadfield, G. K., Hall, A. B., Hooker, S., Lazar, S., Newman, S., Papailiopoulos, D., Tekofsky, S., Toner, H., Ududec, C., & Narayanan, A. (2026). *Open-World Evaluations for Measuring Frontier AI Capabilities* (arXiv:2605.20520). arXiv. https://doi.org/10.48550/arXiv.2605.20520

Kohavi, N., & Weichselbraun, A. (2026). Human tests for machine models: What lies "Beyond the Imitation Game"? *Journal of Linguistic Anthropology*, *36*(1), e70035. https://doi.org/10.1111/jola.70035

Laire, K. (2025, July 8). How to build an AI toolkit for your newsroom. *RJI Innovation in Focus*. https://rjionline.org/news/how-to-build-an-ai-toolkit-for-your-newsroom/

Li, C. (2026, April 29). *Capturing Journalists' Needs in LLM Uncertainty Communication*. Medium. https://generative-ai-newsroom.com/capturing-journalists-needs-in-llm-uncertainty-communication-8e3f84e5b06f

Li, C., Hagar, N., Nishal, S., Gilbert, J., & Diakopoulos, N. (2026). *Towards Real-World Validity in Generative AI Benchmarks: Understanding and Designing Domain-Centered Evaluations for Journalism Practitioners* (arXiv:2511.05501). arXiv. https://doi.org/10.48550/arXiv.2511.05501

Li, M., Chen, M.-B., Tang, B., ShengbinHou, S., Wang, P., Deng, H., Li, Z., Xiong, F., Mao, K., Peng, C., & Luo, Y. (2024). NewsBench: A Systematic Evaluation Framework for Assessing Editorial Capabilities of Large Language Models in Chinese Journalism. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9993–10014. https://doi.org/10.18653/v1/2024.acl-long.538

Liao, Q. V., & Xiao, Z. (2025). *Rethinking Model Evaluation as Narrowing the Socio-Technical Gap* (arXiv:2306.03100). arXiv. https://doi.org/10.48550/arXiv.2306.03100

Lucas, L. (2026, January 28). We graded how five free AI chatbots performed journalism tasks. Here's what we learned. *News Media Help Desk*. https://www.newsmediahelpdesk.org/benchmarking-ai-performance-journalism-tasks/

Mari, W. (2024). The Pre-History of News-Industry Discourse Around Artificial Intelligence. Emerging Media, 2(3), 499–522. https://doi.org/10.1177/27523543241279577

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2026). Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. *IEEE Transactions on Artificial Intelligence*, *7*(1), 22–39. https://doi.org/10.1109/TAI.2025.3569516

Mollick, E. (2026, February 18). *Giving your AI a Job Interview*. https://www.oneusefulthing.org/p/giving-your-ai-a-job-interview

Nahar, N., Kästner, C., Butler, J., Parnin, C., Zimmermann., T., & Bird, C. (2025). Beyond the Comfort Zone: Emerging Solutions to Overcome Challenges in Integrating LLMs into Software Products. *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 516–527. https://doi.org/10.1109/ICSE-SEIP66354.2025.00051

Orr, W., & Kang, E. B. (2024). AI as a Sport: On the Competitive Epistemologies of Benchmarking. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 1875–1884. https://doi.org/10.1145/3630106.3659012

Quinonez, C., & Meij, E. (2024). A new era of AI-assisted journalism at Bloomberg. *AI Magazine*, *45*(2), 187–199. https://doi.org/10.1002/aaai.12181

Rønning Pedersen, T., Øvlisen, K., Connelly, L., Assent, I., & Graves Petersen, M. (2026). Breaking News or Breaking Trust? Exploring Challenges and a Design Space for Trustworthy LLM Integration in Journalism. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*, 1–19. https://doi.org/10.1145/3772318.3791457

Rønning Pedersen, T., Øvlisen, K., & Graves Petersen, M. (2026). Inside the Newsroom: How Media Company AI Strategies Fall Short in Practice. *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems, CHI EA '26*, 1–5. https://doi.org/10.1145/3772363.3799296

Schellmann, H. (2025, August 19). I Tested How Well AI Tools Work for Journalism. *Columbia Journalism Review*. https://www.cjr.org/analysis/i-tested-how-well-ai-tools-work-for-journalism.php

Schjøtt, A., & Blanke, T. (2026). Making machine learning good enough – studying the political endeavour of finding 'right' metrics and thresholds. *Big Data & Society*, *13*(2), 20539517261429198. https://doi.org/10.1177/20539517261429198

Shukla, P., Bui, P., Levy, S. S., Kowalski, M., Baigelenov, A., & Parsons, P. (2025). De-skilling, Cognitive Offloading, and Misplaced Responsibilities: Potential Ironies of AI-Assisted Design. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25*, 1–7. https://doi.org/10.1145/3706599.3719931

Sibinga, E. (2026, August 12). Critically Evaluating LLMs: What Data Visualization Can Teach Us. Nightingale. https://nightingaledvs.com/critically-evaluating-llms/

Trusov, M., Hwang, M., Jamal, Z., & Chandra, S. (2025). *Strategic Decision Framework for Enterprise LLM Adoption* (arXiv:2511.18589). arXiv. https://doi.org/10.48550/arXiv.2511.18589

Uppadhyay, M., Beniwal, H., Kodali, P., & Sitaram, S. (In review). *The State and Fate of Multilingual, Contextual Evaluation in the NLP World*.

van der Maden, W., Sadek, M., Xiao, Z., Mottelson, A., Liao, Q. V., & Zhu, J. (2026). Results-Actionability Gap: Understanding How Practitioners Evaluate LLM Products in the Wild. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*, 1–17. https://doi.org/10.1145/3772318.3791069

Varuvel Dennison, D., Jain, M., Ganu, T., & Vashistha, A. (2026). Designing Culturally Aligned AI Systems For Social Good in Non-Western Contexts. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3772318.3791513

Wilczek, B., Haim, M., & Thurman, N. (2024). Transforming the value chain of local journalism with artificial intelligence. *AI Magazine*, *45*(2), 200–211. https://doi.org/10.1002/aaai.12174

Wirth, A., Ludec, C. L., Girard-Chanudet, C., & Cornet, M. (2025). Who Decides When Artificial Intelligence Works? Algorithmic Production as a Continuous Evaluation Process. *Digital Society*, *4*(3), 65. https://doi.org/10.1007/s44206-025-00224-4

Xie, S., Xiao, S., Peng, C., Mani, G., Zimmerman, J., & Eslami, M. (2026). Investigating How Leaders Decide on AI Innovations: Opportunities for HCI. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*, 1–12. https://doi.org/10.1145/3772318.3790484

Yu, R., Chen, V., Talwalkar, A., & Heidari, H. (2025). Why Do Decision Makers (Not) Use AI? A Cross-Domain Analysis of Factors Impacting AI Adoption. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(3), 2772–2784. https://doi.org/10.1609/aies.v8i3.36758

Endnotes

1. See Mari, 2024
2. Nahar et al., 2025; van der Maden et al., 2026; Wirth et al., 2025; Xie et al., 2026. See Schjøtt & Blanke (2026) and C. Li et al. (2026) for journalism in particular.
3. Both Abdurahman et al., 2025 and Nahar et al., 2025 consider strategies to address non-determinism in different fields.
4. Abdurahaman et al., 2025; Hardy et al., 2025; McIntosh et al., 2026
5. Atari et al., 2023; Cifliku & Heuer, 2026; C. Li et al., 2026; McIntosh et al., 2026; Uppadhyay et al., in review; Varuvel Dennison et al., 2026
6. Hardy et al., 2025; van der Maden et al., 2026
7. M. Li et al., 2024; Chien et al., 2025
8. Abdurahaman et al., 2025; C. Li et al., 2026; McIntosh et al., 2026; Varuvel Dennison et al., 2026
9. Cordero et al., 2025; C. Li et al., 2026; Schjøtt & Blanke, 2026; Wirth et al., 2025
10. Xie et al., 2026
11. Abdurahaman et al., 2025; Laire, 2025; Lucas, 2026; Nahar et al., 2025; Wirth et al., 2025
12. C. Li et al., 2026; Rønning Pedersen et al., 2026a; see also Cordero et al., 2025 and Sibinga, 2026.
13. See Jaźwińska & Chandrasekar, 2025 on the gap in journalism.
14. Both Abdurahman et al., 2025 and Nahar et al., 2025 consider strategies to address non-determinism in different fields.
15. Liao & Xiao, 2025; Orr & Kang, 2024
16. See Liao & Xiao, 2025
17. Nahar et al., 2025
18. Chang et al., 2024; Guo et al., 2023. In practice, however, most benchmarks are used by a single developer, making comparison difficult (Baack et al., 2026).
19. Chang et al., 2024; Guo et al., 2023
20. McIntosh et al., 2026; Orr & Kang, 2024; C. Li et al., 2026
21. Hardy et al., 2025; Liao & Xiao, 2025; C. Li et al., 2026
22. Abdurahman et al., 2025; Chang et al., 2024; Guo et al., 2023; Hardy et al., 2025
23. Chang et al., 2024; Guo et al., 2023; Orr & Kang, 2024; Uppadhyay et al., in review
24. Baack et al., 2026; Eriksson et al., 2025; McIntosh et al., 2026
25. Uppadhyay et al., in review
26. Uppadhyay et al., in review
27. Eriksson et al., 2025; Guo et al., 2023
28. Eriksson et al., 2025; Hardy et al., 2025; Nahar et al., 2025; Wirth et al., 2025
29. Kapoor et al., 2026
30. Hardy et al., 2025; Varuvel Dennison et al., 2026; Xie et al., 2026; Yu et al., 2025
31. Hardy et al., 2025; Varuvel Dennison et al., 2026
32. Varuvel Dennison et al., 2026 provide this framework.
33. CNTI, 2025; Uppadhyay et al., in review
34. Varuvel Dennison et al., 2026
35. Yu et al., 2025
36. Varuvel Dennison et al., 2026
37. Varuvel Dennison et al., 2026
38. Yu et al., 2025
39. Xie et al., 2026
40. M. Li et al., 2024
41. Chien et al., 2025
42. Chien et al., 2025
43. Quinonez & Meij, 2024
44. Schjøtt & Blanke, 2026
45. Cordero et al., 2025; C. Li, 2026; C. Li et al., 2026; Rønning Pedersen, et al., 2026a
46. C. Li et al., 2026; Rønning Pedersen, et al., 2026a
47. Jaźwińska & Chandrasekar, 2025
48. C. Li, 2026; Rønning Pedersen, et al., 2026b
49. Cordero et al., 2025, p. 16

Current Research Working Group Members

Akintunde Babatunde	Executive Director, Centre for Journalism Innovation and Development
Claudia Báez	Associate Consultant, Fathm
Jay Barchas-Lichtenstein	Senior Research Manager, Center for News, Technology & Innovation
Madhav Chinnappa	Independent Media Consultant
Mark Chonofsky	AI Fellow, Chicago Public Media; Fellow, Knight Lab, Medill School of Journalism at Northwestern
Utsav Gandhi	PhD Student, University of Illinois Chicago
Gregory Gondwe	Assistant Professor, Journalism & Creative Media, University of Alabama
Johannes Klingbiel	Research Manager, Media Lab Bayern
K.V. Kurmanath	Senior Journalist and Academic
Amy Mitchell	Executive Director, Center for News, Technology & Innovation
Chris Moran	Head of Editorial Innovation, Guardian News & Media
Sophie Morosoli	Postdoctoral Researcher, AI, Media & Democracy Lab, University of Amsterdam
Afrooz Mosallaei	Research Associate, Center for News, Technology & Innovation
Gary Mundy	Director of Research, Policy and Impact, Thomson Foundation
Oluwapelumi Oginni	Project Manager, AI Initiatives, Centre for Journalism Innovation and Development
Joshua Olufemi	Executive Director, Dataphyte Foundation
Oluseyi Olufemi	Nigeria Country Director, Dataphyte
Esteban Ponce de León	Resident Fellow, Digital Forensic Research Lab (DFRLab) at the Atlantic Council
Amy Ross Arguedas	Research Fellow, Reuters Institute for the Study of Journalism
Zara Schroeder	Researcher, Research ICT Africa
Felix M. Simon	Research Fellow in AI and News, Reuters Institute for the Study of Journalism Research Associate, Oxford Internet Institute, University of Oxford
Sabina Tomkins	Assistant Professor of Information, School of Information and Faculty Associate, Center for Political Studies, Institute for Social Research, University of Michigan
Jaemark Tordecilla	Independent Media Advisor, Philippines