On-Premise AI for the Newsroom: Evaluating Small Language Models for Investigative Document Search

Nick Hagar
nicholas.hagar@northwestern.edu
Northwestern University
Evanston, IL, USA

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University
Evanston, IL, USA

Jeremy Gilbert
jeremy.gilbert@northwestern.edu
Northwestern University
Evanston, IL, USA

Abstract

Investigative journalists routinely confront large document collections. Large language models (LLMs) with retrieval-augmented generation (RAG) capabilities promise to accelerate the process of document discovery, but newsroom adoption remains limited due to hallucination risks, verification burden, and data privacy concerns. We present a journalist-centered approach to LLM-powered document search that prioritizes transparency and editorial control through a five-stage pipeline—corpus summarization, search planning, parallel thread execution, quality evaluation, and synthesis—using small, locally-deployable language models that preserve data security and maintain complete auditability through explicit citation chains. Evaluating three quantized models (Gemma 3 12B, Qwen 3 14B, and GPT-OSS 20B) on two corpora, we find substantial variation in reliability. All models achieved high citation validity and ran effectively on standard desktop hardware (e.g., 24 GB of memory), demonstrating feasibility for resource-constrained newsrooms. However, systematic challenges emerged, including error propagation through multi-stage synthesis and dramatic performance variation based on training data overlap with corpus content. These findings suggest that effective newsroom AI deployment requires careful model selection and system design, alongside human oversight for maintaining standards of accuracy and accountability.

Keywords

investigative journalism, semantic search, retrieval-augmented generation, large language models, hallucination detection

1 Introduction

Investigative journalism often involves analyzing vast document collections [6]. While semantic search has emerged as a powerful tool for navigating these collections, the sheer volume of information often remains overwhelming.

Large language models (LLMs) offer a promising solution to this analytical bottleneck. Their ability to rapidly synthesize information from multiple sources, identify patterns, and generate coherent summaries could theoretically transform weeks of document review into hours of targeted analysis [8]. Retrieval-augmented generation (RAG) systems, which combine semantic search with LLM synthesis, have already demonstrated success in knowledge-intensive domains, producing fluent answers grounded in retrieved documents while maintaining citations to source material [15].

Yet newsroom adoption of such AI assistants remains limited [10], constrained by a fundamental tension between the efficiency gains these systems promise and the verification standards journalism demands. When every published claim carries potential legal and ethical implications [9], the well-documented tendency of LLMs to hallucinate becomes a serious threat to journalistic credibility [19, 20]. Moreover, the proprietary and opaque nature of leading LLM services raises concerns about data privacy [26] and editorial independence [22]. In this context, evaluating new LLM-driven approaches to reporting is a key concern for practice [27].

This paper presents a journalist-centered approach to LLM-powered document search that addresses these concerns through transparent, modular design. Rather than treating AI as an autonomous agent that produces final answers, we position it as a sophisticated intermediary that executes systematic search strategies while maintaining complete auditability. Our system leverages small, locally-deployable language models to preserve data security and editorial independence [4], and it maintains explicit citation chains from every claim to its source documents. Through this design, we aim to capture the analytical power of AI while preserving the transparency that investigative journalism requires.

We evaluate our approach using two document collections that represent common journalistic challenges. The first mirrors the task of getting up to speed on a complex topic, requiring the system to synthesize environmental impact reports and academic papers into a coherent briefing. The second simulates an investigation, where the system acts as a research assistant to find potential story leads in a collection of legal and business documents. Our results demonstrate that small models can achieve high citation accuracy and meaningful document synthesis on standard newsroom hardware [25], though with important variations in reliability that underscore the need for continued human oversight. We find that while technical capabilities like plan adherence and citation management prove robust across models, the tendency to hallucinate varies dramatically based on both model architecture and corpus characteristics, suggesting that deployment decisions must consider the investigative context.

2 Background

Investigative journalists often face a *needle-in-the-haystack* challenge of sifting through massive troves of documents (e.g. FOIA records, leaks) on tight deadlines to find pertinent information [6]. In such large collections, keyword searches alone often fail to surface all relevant documents because journalists may not guess the exact terms hidden in the data. This limitation has motivated the adoption of *semantic search*, which matches queries to results by meaning rather than literal keywords. Using now-standard techniques where documents are chunked, embedded, and indexed for fast retrieval, this capability enables journalists to discover themes, connections, and patterns that traditional search would miss [7, 8].

An emerging frontier in this area is the integration of generative language models with semantic search pipelines. Rather than simply returning a list of relevant documents for the journalist to read, generative systems direct an LLM to first read the retrieved documents, then produce a synthesized answer. This approach, called retrieval-augmented generation (RAG), represents an attempt to handle knowledge-intensive queries by combining the strengths of search (accuracy and grounding in up-to-date, specific information) with the strengths of LLMs (fluid natural language answers) [15]. In theory, this yields the "best of both worlds": The model can cite specific source documents as evidence, which allows the user to verify claims, while retrieving relevant context reduces the model's reliance on potentially outdated parametric memory [17].

Despite the potential of RAG, newsroom adoption of such AI assistants has been cautious [10]. Because of their tendency to hallucinate, many newsrooms do not deem LLM outputs themselves trustworthy, requiring humans to verify information against source documents [3, 13]. And from a technical perspective, while LLM capabilities have increased dramatically over the past five years, hallucinations remain an issue, in some cases even increasing as models become more capable [19, 20].

For newsrooms, these concerns motivate a focus on *automating routine news production* and *enabling faster research* while preserving transparency [8, 13]. Rather than tools that act as autonomous black boxes, journalists see the largest benefit from AI systems that augment their capabilities while keeping them in the driver's seat.

The technology industry's answer to this challenge is a new class of agentic "deep research" tools capable of autonomous synthesis [16]. While powerful, their operation as opaque, cloud-based services makes them fundamentally unsuitable for sensitive investigative work that demands confidentiality and source protection. This points to the need for a parallel approach designed for the newsroom: a journalistic deep research system that combines sophisticated reasoning with the transparency, security, and auditable citation chains that the profession requires.

This motivates our central research question: **How can we design an LLM-powered document search assistant that amplifies journalists' investigative capabilities while preserving transparency and control?**

3 System design

Our system design draws on key considerations in supporting investigative reporting for newsrooms, and on established best practices and design patterns for LLM-based agents.

3.1 Newsroom Considerations

Privacy and Security. The document collections that investigative journalists need to search are often highly sensitive, containing material that carries legal risks [9] and privacy concerns [26]. As such, journalists require a high level of data security for any technical systems that will interact with these materials, ranging from restricted access to airgapped computers [4]. Commercial cloud services are often not sufficient for these use cases; rather, on-premise tools allow the level of control and legal protection required.

Resource Constraints. Many newsrooms do not operate with extensive budgets for technical infrastructure or software engineering expertise [25]. As such, technical solutions must be *lightweight* (i.e., not requiring specialized machines) and *easily maintainable*.

Transparency. Finally, the high-stakes nature of investigative journalism means published claims must be verifiable against primary sources. In the context of an agentic system, this means that every step requires transparency into the LLM's operation, as well as direct citations to source documents—no part of the workflow can rely solely on the LLM's output without verification.

3.2 Agent Best Practices

Rather than diving straight into a task or series of tasks, effective agents often first carry out a planning step—a pattern that allows the LLM to enumerate a checklist or sequence of steps, ensuring a comprehensive, high-level understanding of the goal [1]. After planning, effective agents often execute searches in isolated threads, before combining them into a final answer [2]. Finally, the increasing capabilities of small language models—both generally and for tool use—make them increasingly viable for agentic use cases [5].

3.3 System Implementation

We deploy a five-stage, on-premise pipeline designed for small, locally run LMs and transparent tool use. Each stage emits plaintext artifacts for auditability, and every claim carries an explicit citation key—generated via document chunk hashing—that resolves to the exact source passage. Full prompts, code, hyperparameters, and system outputs are available in the project repository.[1]

Stage 1–Corpus synopsis: Summarize document headers/snippets to produce an outline of topics, recurring themes, and gaps.

- *Example: For the Trump Scotland corpus, the synopsis might highlight key categories of documents like "Scottish planning and environmental assessments" and "U.S. legal and court documents."*

Stage 2–Search planning: Given the synopsis and a prompt specifying the required output format for search plans, propose 5–7 independent threads with objectives, sub-objectives, and candidate queries.

- *Example: For the AI & environment corpus, a search thread might direct the model to "investigate AI applications in e-waste management."*

Stage 3–Execution & reporting: Run each thread in isolation over a local semantic index; draft a thread report with citations plus open questions/next-step queries.

- *Example: Reports contain claims ("For 2015, the company incurred a loss of £1.650k, with total equity at a deficit of £13,682k") and citation keys ("[7db3cb:0]").*

Stage 4–Evaluation: A lightweight judge scores relevance and coverage; only passing reports proceed to synthesis.

Stage 5–Synthesis: Merge surviving reports into an executive brief, de-duplicating findings and aggregating next steps while preserving end-to-end citation chains.

- *Example: An executive summary for the AI & environment corpus might contain high-level information on electricity and water usage, e-waste concerns, and any gaps in the corpus.*

[1] https://github.com/NHagar/semantic-search-assistant

Practical notes. The system runs on a single workstation; all plans, reports, metrics, and model versions are logged for reproducibility. Depending on the model and corpus, a typical end-to-end run of this system took 2–3 hours. Implementation details (indexing, chunking, embedding model, vector store) are documented in the repository.

4 Evaluation

We evaluate our system design on two tasks: A reference task (i.e., getting up to speed on a high-level collection of documents) and an investigation task (i.e., highlighting potential leads in a document corpus).

For the reference task, we collect 64 news articles, academic articles, and technical reports cited in the Wikipedia article on the environmental impact of AI[2]. For the investigation task, we leverage the "Trump Scotland" corpus on Aleph[3], a database maintained by the Organized Crime and Corruption Reporting Project. Compiled by Adam Davidson, this corpus contains documents—legal filings, building designs, permits, and correspondence—pertaining to Donald Trump's golf course in Aberdeen, Scotland. We extract 292 documents from this corpus.

We test our system with three small language models, chosen to provide a balance of model capability and efficiency: Gemma 3 12B[4] [23], Qwen 3 14B [24], and GPT-OSS 20B [18]. All models are quantized to 4-bit precision. All evaluation runs are executed on a MacBook Air with an M3 chipset and 24 GB of unified memory, using LM Studio.

4.1 Protocol

We run the system with identical model configurations for each dataset. For each run, we collect all thread reports (6–7 reports per run) and the final synthesized report. We then conduct claim-level annotation on each thread report. We segment outputs into atomic claims and, for each claim, record support status, citation validity, and (if applicable) hallucination severity.

4.2 Metrics

We report four metrics, averaged across all thread reports, to capture factuality, hallucination risk, citation validity, and process reliability (exact calculation details in the repository).

Claim support rate. Share of model claims that are directly supported by at least one cited passage from the corpus.

Hallucination severity index (HSI).. Counts hallucinated claims and weights them by severity—minor (1), major (2), critical (3)—then averages across reports; lower is better [20].

Invalid citation rate. Fraction of citations that do not correspond to actual document chunks.

Plan adherence. Fraction of planned sub-objectives that were either satisfied with supported evidence or explicitly concluded as unsupported after documented search attempts.

[2]https://en.wikipedia.org/wiki/Environmental_impact_of_artificial_intelligence
[3]https://aleph.occrp.org/datasets/2705
[4]Due to memory constraints on our hardware, the Gemma 3 12B model could not process the high token count of the "Trump Scotland" corpus required for the initial synopsis stage. This condition is therefore omitted from our results.

4.3 Final Reports

Because the final reports for each search are generated via a synthesis of the thread reports, we do not reexamine errors that were already present in the thread reports. Instead, we report instances where the final report introduced additional unsupported claims, invalid citations, or hallucinations into the synthesized material.

5 Results

Table 1 summarizes the performance of our three tested models across both corpora. The results reveal substantial variation in model reliability: Qwen 3 14B demonstrated the most consistent performance across all metrics, while GPT-OSS 20B showed more variable outcomes depending on the corpus and task complexity.

Table 1: Model performance (search thread-level averages) across corpora. Qwen 3 14B reliably produces valid citations and the fewest hallucinations.

Model	Corpus	Claim Support	Halluc. Index	Invalid Citation	Plan Adherence
Qwen	AI & env.	0.95	1.83	0.00	1.00
Gemma	AI & env.	1.00	8.57	0.10	0.79
GPT-OSS	AI & env.	0.92	22.50	0.00	1.00
Qwen	Trump Scot.	0.98	1.00	0.00	1.00
Gemma	Trump Scot.	–	–	–	–
GPT-OSS	Trump Scot.	1.00	1.50	0.00	1.00

5.1 Instruction Following and System Reliability

With the exception of Gemma 3 12B, the tested models demonstrated strong capabilities in following multi-stage pipeline instructions. Both Qwen 3 14B and GPT-OSS 20B achieved perfect plan adherence scores, addressing all specified sub-objectives within their assigned search threads. This high adherence rate indicates that these models can reliably execute planned search strategies. The models also demonstrated robust citation management, accurately replicating citation keys that corresponded to actual document chunks retrieved from the vector index. This technical reliability provides a crucial foundation for journalistic use cases.

5.2 Validity of Generated Reports and Hallucination Patterns

The validity of generated reports varied considerably across models and corpora, revealing important patterns in how different architectures handle document-grounded reasoning tasks. Across all tested conditions, the vast majority of claims aligned correctly with their provided citations, indicating that when models cite sources, they generally do so accurately. However, variation emerged in the frequency and severity of hallucinations.

Gemma 3 12B proved the most problematic, generating substantial hallucinations with a severity index of 8.57 on the AI & environment corpus. The model's poor performance extended beyond hallucinations to plan adherence, managing only 79% completion of its planned sub-objectives.

Qwen 3 14B emerged as the most reliable performer, producing the equivalent of 1-2 minor hallucinations per report across both test corpora. Its consistency across different document types and investigation contexts suggests robust grounding capabilities that align well with journalistic verification requirements.

GPT-OSS 20B showed marked corpus sensitivity, performing well on the Trump Scotland documents (HSI: 1.50) but poorly on the AI & environment collection (HSI: 22.50). This difference may be linked to the model's tendency to incorporate information from its training data. On the AI & environment corpus—a topic likely well-represented in the model's training—GPT-OSS frequently referenced documents and facts that it had not actually encountered during the search process. Notably, this pattern did not emerge with the Trump Scotland corpus, possibly because this specialized legal and business documentation represents a more niche domain with less representation in model training data.

5.3 Final Report Synthesis Challenges

Our analysis of the final synthesized reports revealed systematic issues with the multi-stage aggregation process. The system lacks any correction mechanism, meaning hallucinated or incorrect information from individual thread reports propagates unchanged into the final output. This error persistence represents a fundamental challenge for multi-stage agentic systems in high-stakes domains like journalism. We also observed one concerning instance where Qwen 3 14B, despite its generally strong performance at the thread level, introduced two severe hallucinations and one invalid citation during the synthesis stage of the Trump Scotland corpus analysis. This finding suggests that even reliable models can introduce errors during the reasoning required to merge information threads.

5.4 Qualitative Observations and System Limitations

Several qualitative patterns emerged that illuminate both capabilities and constraints of the current system design. Rather than producing entirely fabricated reports when encountering sparse search results, both Qwen 3 14B and GPT-OSS 20B demonstrated appropriate uncertainty handling. They reliably flagged instances where no relevant documents were retrieved for specific search queries. GPT-OSS proved especially transparent in this regard, producing detailed logs of attempted searches when encountering dead ends on investigative threads.

5.4.1 Corpus Composition and Breadth Effects. The relationship between corpus composition and search performance proved complex and context-dependent. The Trump Scotland corpus presented particular challenges due to its heterogeneous document types, ranging from architectural renderings to legal filings. In our current configuration, search agents struggled with this diversity, focusing predominantly on the corpus's primary themes (building plans and permitting processes) while overlooking potentially valuable auxiliary material.

The AI & environment corpus maintained much tighter topical focus, yet produced less reliable results from GPT-OSS, suggesting that corpus breadth alone does not determine efficacy. The interaction between model architecture, training data overlap, and document diversity requires further investigation to optimize system performance across different investigative contexts.

5.4.2 Document Processing and Technical Infrastructure. Document preprocessing quality emerged as a critical factor in search effectiveness. OCR errors and malformed text occasionally confounded the models' ability to locate relevant chunks, underscoring the importance of high-quality document ingestion pipelines. However, the models also demonstrated surprising robustness, successfully extracting precise figures from malformed tabular data and correcting minor OCR errors during report generation. This suggests that while perfect document processing is ideal, the underlying system can generate useful results even with imperfect input data.

6 Discussion

Our evaluation demonstrates that small language models can serve as capable research assistants for investigative journalism, though with important caveats that illuminate broader questions about human–AI collaboration in high-stakes domains. While models like Qwen 3 14B achieved near-perfect citation accuracy and plan adherence, the substantial variation in hallucination rates and the systematic error propagation during synthesis underscore that effective deployment requires careful model selection, transparent system design, and sustained human oversight.

These findings suggest the path forward is not a choice between human expertise and AI, but a need for systems that amplify journalistic capabilities while preserving verification standards. [11]. Rather than attempting to make model reasoning interpretable—a potentially impossible task [21]—our approach maintains a clear chain of evidence from query to citation to source document. Our multi-stage design also demonstrates that pursuing transparency does not require abandoning sophisticated analytical capabilities. Through careful prompt engineering and modular architecture, the system provides systematic document exploration that would be prohibitively time-consuming for human researchers, while ensuring that every claim remains traceable to source material.

As with many newsroom technologies, deploying LLMs is an editorial, not just technical, decision [12]. Our results show that model selection reflects fundamental choices about verification standards. Editors must consider how a model's training might bias its performance on certain topics. For example, a model that excels on niche legal documents might prove unreliable on topics well-represented in its training data, where it may confuse retrieved evidence with its own parametric knowledge.

6.1 Practical Implications

Our findings address two critical barriers to newsroom AI adoption: cost and data security. The ability to run effective semantic search and synthesis pipelines on a single workstation with 24 GB of memory means that even resource-constrained newsrooms can experiment with AI assistance [14]. This finding is particularly important for investigative teams working with sensitive documents that cannot be uploaded to commercial AI services.

However, successful implementation requires recognizing these systems as extensions of journalistic capability rather than replacements for editorial judgment. Our results suggest that the most productive deployment approach treats AI as a sophisticated research

assistant that can surface relevant material and salient patterns, while leaving editorial decisions about significance, newsworthiness, and story direction firmly in human hands.

6.2 Limitations

Our evaluation prioritized functional accuracy and citation validity over editorial judgment. Future work should incorporate newsworthiness assessments. Our test corpora, while drawn from real investigative contexts, may differ from typical document dumps in ways that affect system performance. In particular, this system design focuses solely on text-based research. Future work should explore evaluation approaches for multimodal corpora. Finally, the probabilistic nature of language model inference means that our quantitative results may not reflect consistent real-world performance. Small models, in particular, can show significant run-to-run variation that our limited evaluation runs may not have captured. Newsrooms considering deployment need extensive testing on their specific document types and use cases.

6.3 Conclusion

The integration of AI assistance into investigative journalism requires balancing sophisticated analytical capabilities with the transparency and accountability that credible reporting demands. Our findings demonstrate this balance is achievable through careful system design that prioritizes auditability, treats model selection as an editorial decision, and positions AI as an extension of rather than replacement for journalistic expertise. While significant technical and methodological challenges remain, the demonstrated capability of local, resource-efficient models suggests that newsrooms can experiment with AI-assisted investigation without compromising their core values of verification, independence, and editorial control.

Acknowledgments

This work was conducted with support from the Knight Foundation.

References

[1] Anthropic. 2024. Building Effective AI Agents. https://www.anthropic.com/engineering/building-effective-agents

[2] Anthropic. 2025. How we built our multi-agent research system. https://www.anthropic.com/engineering/multi-agent-research-system

[3] Kim Björn Becker, Felix M. Simon, and Christopher Crum. 2025. Policies in Parallel? A Comparative Study of Journalistic AI Policies in 52 Global News Organisations. *Digital Journalism* (Jan. 2025), 1–21.

[4] Riley Beggin. 2016. Security for journalists: How to keep your sources and your information safe. https://www.ire.org/security-for-journalists-how-to-keep-your-sources-and-your-information-safe/

[5] Peter Belcak et al. 2025. Small Language Models are the Future of Agentic AI. doi:10.48550/arXiv.2506.02153 arXiv:2506.02153 [cs].

[6] Matthew Brehmer et al. 2014. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 2271–2280.

[7] Sorup Chakraborty et al. 2025. A scalable framework for evaluating multiple language models through cross-domain generation and hallucination detection. *Scientific Reports* 15, 1 (Aug. 2025), 29981. doi:10.1038/s41598-025-15203-5

[8] Besjon Cifliku and Hendrik Heuer. 2025. "This could save us months of work" - Use Cases of AI and Automation Support in Investigative Journalism. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–8. doi:10.1145/3706599.3719856

[9] Grayson Clary et al. 2025. A Journalist's Guide to Reporting on Information Illegally Obtained by a Third Party. https://www.rcfp.org/resources/reporting-on-information-illegally-obtained-by-third-party/

[10] Hannes Cools and Nicholas Diakopoulos. 2024. Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities. *Journalism Practice* (2024).

[11] Nicholas Diakopoulos and Michael Koliska. 2017. Algorithmic Transparency in the News Media. *Digital Journalism* 5, 7 (Aug. 2017), 809–828.

[12] Nick Hagar and Nicholas Diakopoulos. 2019. Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. *Media and Communication* 7, 1 (Feb. 2019), 117. doi:10.17645/mac.v7i1.1801

[13] Lucinda Jordaan. 2025. Divining data: How AI invigorates The New York Times' approach to investigative reporting. https://wan-ifra.org/2025/05/divining-data-how-ai-invigorates-the-new-york-times-approach-to-investigative-reporting/

[14] Martin Kleppmann, Adam Wiggins, Peter Van Hardenberg, and Mark McGranaghan. 2019. Local-first software: you own your data, in spite of the cloud. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*. ACM, Athens Greece, 154–178. doi:10.1145/3359591.3359737

[15] Patrick Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vol. 793. Curran Associates Inc., Vancouver, BC, Canada, 16.

[16] Minghao Li et al. 2025. ReportBench: Evaluating Deep Research Agents via Academic Survey Tasks. doi:10.48550/arXiv.2508.15804 arXiv:2508.15804 [cs].

[17] Rick Merritt. 2025. What Is Retrieval-Augmented Generation aka RAG? https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

[18] OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. doi:10.48550/arXiv.2508.10925 arXiv:2508.10925 [cs].

[19] OpenAI. 2025. *OpenAI o3 and o4-mini System Card*. Technical Report. OpenAI. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

[20] Vipula Rawte et al. 2023. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 2541–2573. doi:10.18653/v1/2023.emnlp-main.155

[21] Parshin Shojaee et al. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.06941 [cs.AI] https://arxiv.org/abs/2506.06941

[22] Felix M. Simon. 2022. Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism* 10, 10 (Nov. 2022), 1832–1854. doi:10.1080/21670811.2022.2063150

[23] Gemma Team. 2025. Gemma 3 Technical Report. doi:10.48550/arXiv.2503.19786 arXiv:2503.19786 [cs].

[24] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[25] Nicol Turner Lee and Courtney C. Radsch. 2024. Journalism needs better representation to counter AI. https://www.brookings.edu/articles/journalism-needs-better-representation-to-counter-ai/

[26] Maartje Van Der Woude, Tomás Dodds, and Guillén Torres. 2024. The ethics of open source investigations: Navigating privacy challenges in a gray zone information landscape. *Journalism* (Aug. 2024), 19. doi:10.1177/14648849241274104

[27] Joris Veerbeek and Nicholas Diakopoulos. 2024. Using Generative Agents to Create Tip Sheets for Investigative Data Reporting. In *Computation + Journalism Symposium*. doi:10.48550/arxiv.2409.07286