Revitalizing Local Democracy: A Human-Centered Audit of LLMs in City Council Journalism

David Xia
The University of Illinois Urbana-Champaign
Champaign, USA
davidx3@illinois.edu

Chris Maury
Carnegie Mellon University
Pittsburgh, USA
cmaury@andrew.cmu.edu

Abstract

The collapse of traditional local journalism has created a widespread civic information deficit, disconnecting citizens from their elected officials and from one another. Large Language Models (LLMs) present a potential solution to augment resource-constrained news production. While text segmentation and summarization are widely established benchmark capabilities of modern LLMs, their utility in higher-order journalistic tasks, such as identifying newsworthy topics from raw proceedings and concisely articulating this information for readers, remains to be fully investigated. Our work addresses this gap through a human-centered evaluation of LLMs in the domain of city council reporting. We conducted a crowdsourced study where respondents compared the newsworthiness of LLM-generated headlines against those from an expert journalist. Our findings demonstrate that LLMs can perform the core tasks of producing and prioritizing headlines with quality meeting or exceeding that of a professionally written standard. This highlights the potential for LLMs to automate routine reporting tasks, allowing human journalists to focus their limited resources on high-value work. The code and dataset for this work are available at: `https://github.com/davidxia3/llm-civic-audit`.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

Human-Centered Evaluation, LLM Auditing, Local Journalism

ACM Reference Format:
David Xia and Chris Maury. 2026. Revitalizing Local Democracy: A Human-Centered Audit of LLMs in City Council Journalism. In *Proceedings of Conference on Human Factors in Computing Systems (CHI '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The United States is facing a critical local news crisis, driven primarily by economic forces that have raised journalism costs and decimated newsroom employment [15]. This decline has created "news deserts," leaving communities with inadequate reporting on their own local governance, which diminishes civic engagement [1, 6]. This resource-constrained environment presents a clear opportunity for augmentation using Large Language Models (LLMs). By automating routine reporting tasks, LLM systems could empower under-resourced newsrooms to scale their coverage and help bridge the civic information gap.

However, effective local reporting requires more than simple text segmentation and summarization, tasks where LLMs are already well-benchmarked [8, 16]. The core function of a journalist involves complex, higher-order cognitive skills, including the ability to sift through hours of proceedings to identify the most consequential topics and the skill of ranking those topics and crafting headlines that are accurate, concise, and engaging. The central question of our research is whether current LLMs can perform these nuanced journalistic tasks at an expert level in the specialized domain of municipal government.

To investigate this, we designed a human-centered evaluation pipeline that benchmarks LLM performance against professional journalistic standards on a weekly basis. Each week, an expert journalist reviewed city council records to author a small set of headlines summarizing the most significant local government developments. In parallel, LLMs were given access to the same records and tasked with generating and prioritizing their own candidate headlines based on perceived newsworthiness. To ensure a fair comparison, we isolated the top-ranked LLM headlines to match the exact volume of the expert's output for that week. These competing headlines were then presented to human evaluators on the Prolific platform [13], who ranked the journalist's and the LLMs' work together based on clarity, impact, and local relevance. This procedure allows us to measure not only the quality of the headlines the LLMs produce but also their editorial judgment in identifying which stories truly matter to a community.

Our findings show that LLMs can perform this core reporting work at a level comparable and even exceeding that of an expert journalist. However, we also recognize the limitations of current LLMs. These models can only report on data that exists in a digital format (e.g., a recording of a meeting) and cannot perform tasks that require new information gathering, such as reporting on meetings that are not recorded, conducting interviews, asking follow-up questions, or building source relationships. Therefore, we conclude that LLMs show significant potential to augment human reporting by automating the core coverage of public meetings. This automation can, in turn, free human journalists to focus their limited and valuable resources on the high-impact, investigative work where they provide the most value[1].

[1]Project repository: https://github.com/davidxia3/llm-civic-audit

CHI '26, Barcelona, Spain

ACM ISBN 978-X-XXXX-XXXX-X/XXXX/XX
https://doi.org/XXXXXXX.XXXXXXX

2 Methodology

We designed a pipeline to replicate the complete journalistic workflow, encompassing data collection, headline generation, and topic prioritization. To isolate LLM capacity for complex editorial judgment, we focus our preliminary analysis on two primary objectives: 1) the generation of compelling headlines, and 2) the prioritization of newsworthy topics. This exploratory study proceeds in four main stages: data collection and corpus creation, LLM-based headline generation, human evaluation of headline generation, and human evaluation of topic prioritization.

2.1 Data Collection and Corpus Creation

The study focuses on the publicly available records of the Pittsburgh City Council. This context was chosen due to the comprehensive availability of meeting agendas, detailed legislation, and video recordings. As a benchmark for human expert performance, we utilize an expert journalist's weekly reports summarizing four to six key topics from these meetings, complete with headlines for each topic [10].

We implemented a data processing pipeline to create a unified data corpus. This workflow, illustrated in Figure 1, was executed for nine city council meetings across four consecutive weeks in April 2025 and consisted of the following steps:

(1) **Agenda Collection and Segmentation:** For each meeting, the agenda was sourced from a publicly available PDF file. We utilized the `pdfplumber` library[2] to extract its textual content. The agenda was then segmented into discrete topics using Claude Haiku 3.5 [2] (see Appendix 7.1). This model was selected because the meeting agendas follow a highly consistent, structured format. We observed that the task of segmenting these documents is sufficiently straightforward that modern LLMs perform it with near-zero error. Using a lightweight model like Haiku allowed for efficient processing without compromising accuracy.
(2) **Legislation Extraction:** Hyperlinks embedded within the agenda text pointed to the full legislative proposals hosted on the Legistar agenda management system[3]. We followed these hyperlinks to retrieve the full text, which was then associated with the agenda segment containing the source hyperlink.
(3) **Transcript Generation and Alignment:** Video recordings of each meeting were obtained from the Pittsburgh City Council official YouTube channel[4]. We employed OpenAI's Whisper model [14] to generate a transcript of the proceedings. This transcript was then manually segmented and manually aligned with the relevant agenda topics.
(4) **Corpus Assembly:** For each agenda topic, the extracted agenda text, its corresponding legislative text, and the aligned transcript segment were aggregated. This aggregated unit is hereafter referred to as a *Combined Segment*.

Over the four-week period, this process yielded 31, 34, 27, and 30 Combined Segments, respectively (totaling 122). During this same period, the expert journalist identified and wrote headlines for 4, 5, 5, and 6 of the agenda topics for the respective weeks.

[2]https://github.com/jsvine/pdfplumber
[3]https://granicus.com/solution/govmeetings/legistar/
[4]https://www.youtube.com/@CityChannelPittsburgh

2.2 LLM-Based Headline Generation

For the first task of generating compelling headlines, we prompted three different LLMs to generate a one-sentence headline focusing on the "most newsworthy action or decision" for each combined segment (see Appendix 7.2). The models used were:

- Claude Sonnet 4 [3]
- Gemini 2.5 Pro [5]
- GPT-4.1 mini [12]

The "mini" variant of the GPT-4.1 family was selected specifically to accommodate the extensive length of some combined segments, which exceeded the context window limitations of more powerful models. This step resulted in three distinct sets of LLM-generated headlines for the entire corpus.

2.3 Human Evaluation of Headline Generation

To assess the quality of the generated headlines, we established a human-evaluated ground truth using a crowd-sourcing framework and derived specific rankings for analysis.

2.3.1 Crowd-sourced Evaluation. To establish a human-evaluated ground truth, we conducted a large-scale human-centered study on the crowd-sourcing platform Prolific. The study employed a pairwise comparison design to rank headlines based on their ability to engage a reader.

To assess headline characteristics such as newsworthiness and intrigue, which are known to drive reader engagement [9], U.S.-based respondents were shown two headlines at a time and asked: "Which headline do you have a stronger opinion about?". We selected this metric based on prior findings that content evoking high-arousal emotions, such as strong positive or negative opinions, are primary drivers of online engagement [4]. Over the four weeks, we collected responses from 89, 121, 73, and 93 respondents, who completed 1761, 2188, 1453, and 1842 pairwise comparisons, respectively.

2.3.2 Headline Rankings. The crowd-sourced pairwise comparison data was used to rank headlines via the TrueSkill ranking algorithm [7]. To isolate the performance of each LLM against the expert benchmark, we generated a distinct **Headline Ranking** ($\mathcal{R}_{head}^{g,w}$) for each Generating LLM $g \in G$ and each week $w \in W$. These rankings serve as our ground truth for headline quality:

- $\mathcal{R}_{head}^{\text{Claude},w}$: Human preference ranking of Claude Sonnet 4-generated headlines for week w **and** expert-written headlines for week w.
- $\mathcal{R}_{head}^{\text{Gemini},w}$: Human preference ranking of Gemini 2.5 Pro-generated headlines for week w **and** expert-written headlines for week w.
- $\mathcal{R}_{head}^{\text{GPT},w}$: Human preference ranking of GPT-4.1 mini-generated headlines for week w **and** expert-written headlines for week w.

2.4 Human Evaluation of Topic Prioritization

To investigate whether LLMs can replicate professional editorial judgment, we first established a human-validated ground truth for

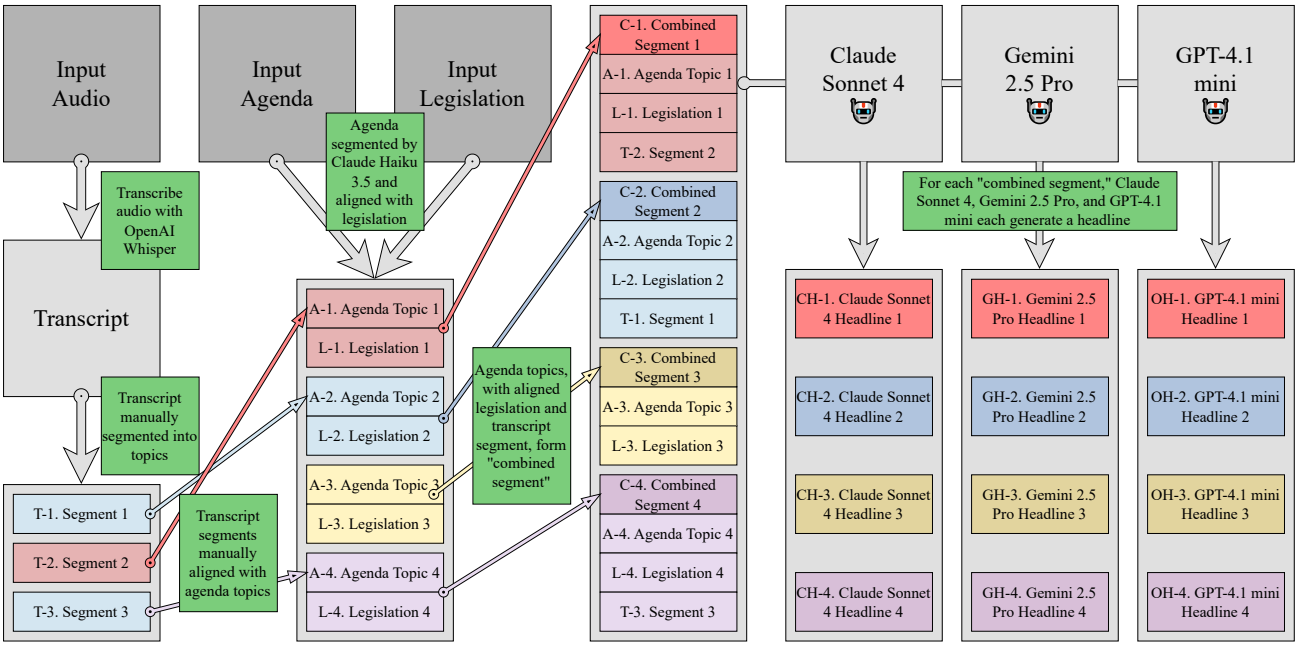

Figure 1: Overview of the data collection and corpus creation pipeline for an illustrative week.

topic importance and then evaluated the models' ability to select topics that align with that standard.

2.4.1 Topic Rankings. We utilized the Headline Rankings ($\mathcal{R}^{g,w}_{head}$) as a proxy for topic newsworthiness. To isolate the intrinsic value of a topic from the quality of its presentation, we controlled for variations in writing style by analyzing each generating LLM independently. By holding the "author" constant (e.g., only analyzing headlines written by Claude), we eliminate confounding variables such as tone or verbosity. In this controlled environment, the relative ranking of two headlines acts as a clean proxy for the human interest in their underlying topics: if a reader prefers a Claude-written headline about Topic A over a Claude-written headline about Topic B, we infer that Topic A is intrinsically more newsworthy.

Based on this rationale, we constructed our ground truth for topic importance using an exclusion process on the Headline Rankings. Consider an illustrative week w containing 30 agenda topics, where the expert journalist selected 5 topics to cover.

(1) We took the Headline Ranking for a given model ($\mathcal{R}^{g,w}_{head}$), which contains 35 items (30 LLM + 5 Expert).
(2) We removed the 5 expert-written headlines from this list, retaining the relative order of the remaining 30 LLM-generated headlines.

By removing the expert-written headlines from the Headline Rankings while maintaining the relative order of the remaining items, we derived the **Topic Rankings** ($\mathcal{R}^{g,w}_{topic}$) for each Generating LLM $g \in G$ and each week $w \in W$. These rankings serve as our ground truth for topic importance:

- $\mathcal{R}^{\text{Claude},w}_{topic}$: Human preference ranking of Claude Sonnet 4-generated headlines **only**.
- $\mathcal{R}^{\text{Gemini},w}_{topic}$: Human preference ranking of Gemini 2.5 Pro-generated headlines **only**.
- $\mathcal{R}^{\text{GPT},w}_{topic}$: Human preference ranking of GPT-4.1 mini-generated headlines **only**.

2.4.2 Modeling LLM Editorial Selection. We then evaluated LLM capacity for editorial judgment using a cross-evaluation framework. We formally distinguish between two roles: the *Generating LLM* (g), responsible for generating the headlines, and the *Judging LLM* (j), tasked with prioritizing those headlines and by proxy, the topics, through pairwise comparisons. To ensure consistency, all Judging LLMs were provided with a uniform definition of "importance" that emphasized changes to the status quo, broad population impact, and relevance to marginalized groups (see Appendix 7.3). The Judging LLMs used were:

- Claude Sonnet 4
- Gemini 2.5 Pro
- GPT-4.1 [11]

The standard variant of the GPT-4.1 family was selected for this task instead of the mini variant. Since pairwise comparison is a concise task, the model does not require the expanded context windows.

The pairwise preferences were aggregated using the TrueSkill algorithm to produce a final ranked list. This process created a full 3×3 evaluation matrix where every Judging LLM ranked the outputs of every Generating LLM, resulting in nine distinct ranking lists per week.

2.4.3 Selection Sets Definition. To compare the LLM selections against the human expert's selections, we applied an adaptive threshold. We define the **Selection Sets** (S) as follows:

- **Expert Selection ($S_{e,w}$):** The set of topics actually published by the expert journalist in week w. We define $N_w := |S_{e,w}|$.
- **Judging LLM Selection ($S_{j,g,w}$):** The set of the top-N_w topics selected by Judge j from the headlines generated by Generator g in week w.

By forcing the LLM to select the same number of topics as the expert (N_w), we ensure a fair comparison of recall and rank metrics.

3 Results

To formalize our analysis, we define a rank retrieval function $rank(\mathcal{R}, a, t)$ which returns the ordinal rank of the headline for topic t written by author a within the ranking list $\mathcal{R}$, where author $a \in A := G \cup \{e\}$ (G denotes the set of Generating LLMs and e denotes the expert).

3.1 Evaluating Headline Quality

To quantify the quality of LLM-produced headlines, we utilize the **Headline Rankings** ($\mathcal{R}^{g,w}_{head}$). We calculate the *Rank Difference* (Δr) for each topic t in the Expert's Selection ($S_{e,w}$). This is defined as the rank of the expert's headline minus the rank of the LLM's headline for the same topic:

$$\Delta r_{g,w,t} = rank(\mathcal{R}^{g,w}_{head}, e, t) - rank(\mathcal{R}^{g,w}_{head}, g, t) \tag{1}$$

Since lower numerical ranks indicate stronger preference (e.g., 1st is preferred over 5th), a positive $\Delta r_{g,w,t}$ indicates that the LLM-generated headline was preferred over the expert-written headline for topic t. We aggregate these comparisons to calculate the *Average Rank Difference* ($\overline{\Delta r}(g)$) for a given Generating LLM g across all weeks $w \in W$ and all topics selected by the expert:

$$\overline{\Delta r}(g) = \frac{\sum_{w \in W} \sum_{t \in S_{e,w}} \Delta r_{g,w,t}}{\sum_{w \in W} N_w} \tag{2}$$

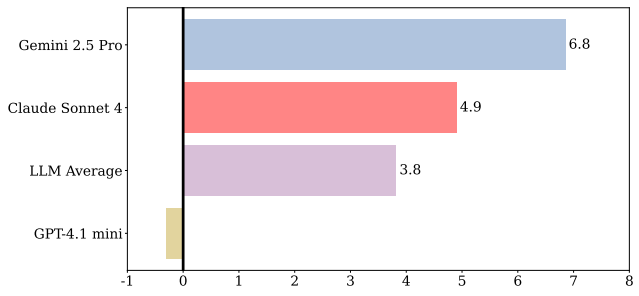

Figure 2: Average rank difference between LLM-generated and expert-written headlines ($\overline{\Delta r}$).

Figure 2 displays the $\overline{\Delta r}(g)$ for each Generating LLM g across the four-week study. In aggregate, LLM-generated headlines outperformed the expert writer, achieving a positive difference of +3.8. Individually, Gemini 2.5 Pro and Claude Sonnet 4 showed strong performances, ranking 6.8 and 4.9 positions higher than the expert, respectively. GPT-4.1 mini performed comparably to the expert but slightly worse, with headlines ranked 0.3 positions lower on average.

3.2 Evaluating Topic Prioritization with Recall Rate

To quantify the efficacy of identifying high-value topics, we calculate *Prioritization Recall* (ρ_k) for evaluation depths $k \in \{3, 5\}$. Intuitively, this metric answers the question: *What percentage of the human-verified top-k stories did each author successfully identify?*

We rely on the **Topic Rankings** ($\mathcal{R}^{g,w}_{topic}$) as our ground truth. We first define the *Ground Truth Set* $T_{g,w}(k)$ as the set of top-k topics in a given list:

$$T_{g,w}(k) = \{t : rank(\mathcal{R}^{g,w}_{topic}, g, t) \leq k\}$$

We then calculate the recall for a Judging LLM j by averaging the intersection of their selection ($S_{j,g,w}$) with this ground truth set across all Generating LLMs $g \in G$ and weeks $w \in W$:

$$\rho_k(j) = \frac{1}{k \cdot |G| \cdot |W|} \sum_{g \in G} \sum_{w \in W} |S_{j,g,w} \cap T_{g,w}(k)| \tag{3}$$

Similarly, the *Expert's Prioritization Recall* ($\rho_k(e)$) is defined by substituting the Expert Selection ($S_{e,w}$) as the Selection Set:

$$\rho_k(e) = \frac{1}{k \cdot |G| \cdot |W|} \sum_{g \in G} \sum_{w \in W} |S_{e,w} \cap T_{g,w}(k)| \tag{4}$$

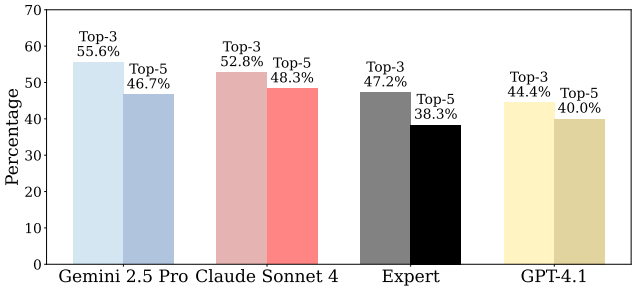

Figure 3: Top-3/5 prioritization recall (ρ_k).

Figure 3 displays the $\rho_k(a)$ for each author $a \in A$ and evaluation depth $k \in \{3, 5\}$. Gemini 2.5 Pro demonstrated the highest efficacy in identifying the most critical stories, achieving a top-3 recall of 55.6% and a top-5 recall of 46.7%. Claude Sonnet 4 yielded comparable performance, securing the highest recall for the top-5 category (48.3%) and a strong recall for the top-3 (52.8%). In contrast, both the expert writer and GPT-4.1 exhibited lower recall metrics. The expert identified 47.2% of the top-3 and 38.3% of the top-5 topics. GPT-4.1 trailed in the top-3 category with 44.4%, though it performed slightly better than the expert in the top-5 category with 40.0% recalled. Across all judges, recall rates for the top-3 category were higher than those for the top-5 category, suggesting higher consensus on the most obvious, high-impact news items compared to secondary stories.

3.3 Evaluating Topic Prioritization with Average Ranks

To assess the overall quality of the editorial choices beyond strict top-k alignment, we calculate the *Average Selection Rank* ($\overline{r}$). This computes the average ground-truth rank of all topics selected by an author. Using the **Topic Rankings** ($\mathcal{R}^{g,w}_{topic}$) as the ground truth reference, the metric for a Judging LLM j is:

$$\overline{r}(j) = \frac{\sum_{g\in G}\sum_{w\in W}\sum_{t\in S_{j,g,w}} rank(\mathcal{R}^{g,w}_{topic}, g, t)}{\sum_{g\in G}\sum_{w\in W} N_w} \quad (5)$$

The *Expert's Selection Rank* $\overline{r}(e)$ is calculated by substituting the Expert Selection ($S_{e,w}$) as the Selection Set. Note that the author argument is fixed to g as the Topic Rankings exclusively contain LLM-generated headlines:

$$\overline{r}(e) = \frac{\sum_{g\in G}\sum_{w\in W}\sum_{t\in S_{e,w}} rank(\mathcal{R}^{g,w}_{topic}, g, t)}{\sum_{g\in G}\sum_{w\in W} N_w} \quad (6)$$

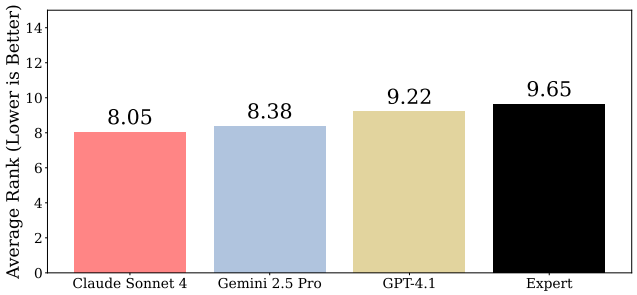

Figure 4: Average rank of topics selected by LLMs and expert ($\overline{r}$).

Figure 4 displays the $\overline{r}(a)$ for each author $a \in A$. Lower average ranks indicate closer alignment with human newsworthiness preferences. Claude Sonnet 4 and Gemini 2.5 Pro demonstrated the highest precision in identifying high-value stories, achieving average selection ranks of 8.05 and 8.38, respectively. Most notably, every LLM in our study outperformed the human expert in this metric. Even the lower-performing GPT-4.1 achieved an average rank of 9.22, showing a more consistent alignment with the human majority than the expert journalist, whose selections resulted in an average rank of 9.65.

4 Discussion

The preliminary findings of this exploratory study suggest that LLMs demonstrate significant potential for augmenting routine journalistic workflows. By deconstructing our evaluation into two distinct competencies (headline generation and topic prioritization), we observe promising trends in machine-augmented journalism.

4.1 Headline Quality

The results in Figure 2 demonstrate that existing LLMs, specifically Gemini 2.5 Pro and Claude Sonnet 4, are capable of significantly outperforming the expert in headline generation tasks. However, the performance gap between these models and GPT-4.1 mini suggests that model capacity plays a critical role in this domain. While GPT-4.1 mini was able to produce headlines of comparable quality to an expert, it failed to match the stronger ranking capabilities of the larger models.

This disparity is likely attributable to the reduced parameter size of the "mini" architecture. It is worth noting that larger models from the OpenAI GPT family were excluded from this study due to context window limitations, which prevented them from processing the full length of the source documents. This finding implies that while smaller, efficient models can achieve human parity, maximizing headline quality currently requires the extensive context reasoning capabilities found in larger foundation models.

4.2 Topic Prioritization Recall Rates

The superior recall rates achieved by Gemini 2.5 Pro and Claude Sonnet 4 suggest that existing LLMs are capable of outperforming the expert in topic prioritization tasks. The data indicates that these models possess a refined ability to align with aggregate human judgment regarding story importance, particularly when identifying the most salient "lead" stories.

It is notable that recall performance degraded across all models when expanding the scope from the top-3 to the top-5. This trend implies that while LLMs are more effective at identifying the undisputed most important news items (top-3), distinguishing between mid-tier stories (rank 4 and 5) remains a more subjective and challenging task. Nevertheless, the consistent performance gap between leading LLMs, like Gemini 2.5 Pro and Claude Sonnet 4, and the expert underscores the potential for LLM-driven systems to serve as reliable editorial assistants for content curation.

4.3 Topic Prioritization Average Ranks

Analysis of the average rank metric reveals that all three Judging LLMs selected, on average, higher-priority topics than the expert. This reinforces the viability of using LLMs for editorial prioritization. However, the margin of improvement is relatively narrow. With an average total pool of 30.5 topics per week, the separation between the best-performing model (Claude Sonnet 4) and the expert is only 1.6 ranks.

A critical insight emerges when synthesizing the recall rates with the average rank data. As noted previously, the models successfully recalled roughly 50% of the top-3 and top-5 stories. Given that the average selection size was $\overline{N_w} = 5$, a "perfect" model would achieve an average rank close to 3.0. The observed average ranks of 8.05 (Claude Sonnet 4) and 8.38 (Gemini 2.5 Pro) indicate a high degree of variance in the models' non-top-tier selections.

Specifically, for the mathematical average to exceed 8.0 despite the inclusion of highly-ranked (top-1–3) stories, the models must be populating the remainder of their selections with low-priority topics, likely including those ranked in the bottom quartile (ranks 20–30) by human evaluators. This suggests that while LLMs are excellent at identifying the undisputed "lead" stories, they occasionally misidentify importance in niche or irrelevant topics, diverging significantly from human consensus on the lower end of the selection list.

4.4 Implications for Automated Reporting

Collectively, the analysis of headline quality and topic prioritization provides preliminary evidence of the potential for LLMs to mitigate resource constraints in local newsrooms. The best models (Gemini 2.5 Pro and Claude Sonnet 4) demonstrated an ability to identify critical civic events and draft headlines that consistently outperformed our expert baseline in human-validated trials. This confirms that the routine labor of monitoring municipal proceedings, a task often abandoned in "news deserts" due to staffing shortages, is technically feasible for LLM automation.

However, the divergence observed in the average rank metric serves as a critical guardrail against fully autonomous deployment. While the models reliably captured the "lead" stories, their tendency to intersperse high-priority topics with low-relevance outliers necessitates a human-in-the-loop workflow. The LLMs demonstrates the capacity to act as a highly effective "junior reporter" that is able to sift through massive transcripts to extract the signal from the noise. However, it lacks the consistent editorial judgment required to act as a publisher. Consequently, these findings support the feasibility of an augmentation framework where LLMs drastically reduce the time burden of information gathering and drafting, allowing human journalists to exercise final editorial judgments.

5 Limitations

Our findings should be interpreted within the context of several constraints inherent to this exploratory study. Specifically, we acknowledge limitations regarding the temporal scope of our dataset, the reliance on a single expert baseline, and the potential disconnects between our computational objectives and the subjective proxies used for human evaluation.

5.1 Temporal Scope and Sample Size

The dataset employed for this analysis was limited to city council meetings occurring over a four-week period. While this window provided a snapshot of model performance, it may not capture the full variance of legislative discourse, which can fluctuate significantly based on seasonal budget cycles, election periods, or specific controversies. A larger corpus spanning multiple months or years would be necessary to validate the consistency of these findings.

5.2 Expert Baseline Representativeness

The expert baseline for this study relied on the output of a single writer. While this individual possesses professional experience in journalism, relying on a solitary data point introduces the risk of bias. Consequently, the "expert" performance recorded here should be viewed as a singular benchmark rather than a definitive upper bound of human capability.

5.3 Evaluator Bias and Ranking Confidence

The subjective evaluation of headlines and topic importance was conducted using the Prolific platform. While crowd-sourcing provides scalability, the respondent pool was not demographically representative of the specific locality where the city council meetings took place. Evaluators lacking local context may prioritize topics differently than actual residents. Furthermore, while pairwise comparisons offer a robust method for ranking, the number of comparisons collected per headline was limited. Increasing the volume of pairwise comparisons, including repeated match-ups, would yield a more concrete ranking.

5.4 Prompt Sensitivity and Construct Definitions

A potential misalignment exists between the objective definition of importance provided to the LLMs and the subjective proxy used for human evaluation. While the LLMs were explicitly instructed to prioritize civic relevance and impact on vulnerable populations, the human evaluators on Prolific were asked to select the headline they held a "stronger opinion" about. Although we utilized opinion as a proxy for engagement based on prior literature, "opinion" and "importance" are not synonymous. A high-impact but bureaucratic topic (e.g. a unanimous budget adjustment) may satisfy the LLM's definition while failing to elicit a strong opinion from a human-valuator. Conversely, a polarizing but low-impact topic might dominate the human evaluation. This discrepancy implies that the LLMs may have been optimizing for impact, while the human evaluated "ground truth" was measuring emotional engagement.

6 Conclusion

This study sought to explore whether LLMs could transcend simple text summarization to perform the higher-order cognitive tasks essential to local journalism, such as newsworthy prioritization and engaging headline generation. In the context of a widening local news crisis characterized by "news deserts" and resource-constrained newsrooms, our findings offer evidence for the potential of using LLMs to help bridge this civic information gap.

Our results demonstrate that models with sufficient context reasoning capabilities, specifically Gemini 2.5 Pro and Claude Sonnet 4, can perform these nuanced reporting tasks at a level comparable to, and in some metrics exceeding, an expert journalist. These models successfully parsed hours of municipal proceedings to identify the most consequential topics, achieving superior recall rates for top-tier stories and generating headlines that human evaluators consistently ranked as more newsworthy than the expert baseline.

Crucially, these findings validate the potential for an "augmentation" model of local journalism rather than full displacement. While LLMs demonstrate proficiency in processing existing meeting data, they remain bounded by their input. They cannot conduct the interviews, cultivate the sources, or gather the external context that defines investigative reporting. Therefore, we conclude the possibility that deploying LLMs to automate the coverage of public meetings can effectively free journalists from the constraints of routine monitoring, allowing them to redirect their limited resources toward the high-impact, relationship-driven work where they provide unique value to their communities.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. 2349558.

References

[1] Penelope Muse Abernathy. 2020. *News Deserts and Ghost Newspapers: Will Local News Survive?* University of North Carolina Press.

[2] Anthropic. 2024. Claude Haiku 3.5. https://www.anthropic.com/news/claude-3-5-haiku Large Language Model.
[3] Anthropic. 2025. Claude Sonnet 4. https://www.anthropic.com/news/claude-4 Large Language Model.
[4] Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research* 49, 2 (2012), 192–205.
[5] Gemini. 2025. Gemini 2.5 Pro. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking Large Language Model.
[6] Danny Hayes and Jennifer L Lawless. 2015. As local news goes, so goes citizen engagement: Media, knowledge, and participation in US House elections. *The Journal of Politics* 77, 2 (2015), 447–462.
[7] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill: A Bayesian skill rating system. In *Advances in neural information processing systems*. 569–576.
[8] Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 469–473.
[9] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.
[10] Chris Maury. 2024. InformUp. https://informup.org/ Accessed: 2025-05-26 – 2025-08-15.
[11] OpenAI. 2025. GPT-4.1. https://openai.com/index/gpt-4-1/ Large Language Model.
[12] OpenAI. 2025. GPT-4.1 mini. https://openai.com/index/gpt-4-1/ Large Language Model.
[13] Prolific. 2014. Prolific. https://www.prolific.com/ Accessed: 2025-05-26 – 2025-08-15.
[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.
[15] Mason Walker. 2021. U.S. newsroom employment has fallen 26% since 2008. *Pew Research Center* (July 2021). https://www.pewresearch.org/fact-tank/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/
[16] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics* 12 (2024), 39–57.

7 Appendix: Prompt Templates

7.1 Agenda Segmentation Prompt

```
You are a professional city council meeting
    assistant.

Given the full raw text of a meeting agenda,
     segment it into distinct agenda items.
    For each item:

1. Include the full agenda item title (e.g.,
     "Bill 2023-114: Amending the zoning
    regulations").
2. Each **bill, paper, resolution, or
    ordinance** (e.g., "ORD. 2023-114", "RES
    . 2023-R016", "PAPER #412") counts as a
    **separate agenda item**, even if
    multiple items fall under the same
    section.
3. If a bill number or ordinance number
    appears, include it as part of the
    agenda item title.
4. Under each agenda item, include **all the
     text** that falls under it until the
    next agenda item begins.
5. Keep the original wording and formatting.
     Do not summarize or shorten the text.
6. Do not skip or omit any part of the
    agenda. This includes routine items such
     as 'Roll Call,' 'Public Comment,' and
    other procedural sections.

Return the segmented agenda as a JSON array
    of strings. Each string is one agenda
    item with its full text.

**Important:** Return **only** the JSON
    array of strings with no additional text
    , explanation, or commentary. The output
     must be a valid JSON array.

[
  "[Agenda item title]\\n   [Full text under
      the item]",
  "[Next agenda item]\\n   [Full text under
     the item]",
  ...
]

Agenda:
\"\"\"[AGENDA TEXT INSERTED HERE]\"\"\"
```

7.2 Headline Generation Prompt

```
You are a local government reporter covering
     city council meetings.

You will receive:
- A section from a **city council meeting
    agenda** (note: this may be vague or
    generic)
- A section from the related **official
    legislation**
- A section from the **meeting transcript**

Your task is to write a **clear, one-
    sentence headline** that:
- Focuses on the **most newsworthy action or
     decision**
- Summarizes what the **council actually did
    **, proposed, debated, or approved
- Highlights **specific outcomes**, impacts,
     or controversial statements
- Is written at an **eighth-grade reading
    level**
- Contains **no commentary** or extra
    background

Do *not* copy or paraphrase the agenda title
    . Use the transcript and legislation
    instead.

---
```

```
[COMBINED SEGMENT INSERTED HERE]
Headline:
```

7.3 Headline Ranking Prompt

```
You will be shown two headlines from city
    council meetings.

### Your Task
Select the headline that is more important,
    using the definition below.

### What Does 'Important' Mean?
A headline is important if:
- It reflects a major change to the status
    quo,
- OR it has a large impact on a large number
     of people,
- OR it has a large impact on a marginalized
     group (e.g., people facing poverty,
    discrimination, or limited access to
    resources),
- OR it covers an issue that is especially
    newsworthy due to its civic relevance,
    urgency, or long-term consequences.

### Consider These Factors
- **Scope**: How many people in the city are
     affected?
- **Depth**: How significant or lasting is
    the impact?
- **Equity**: Does it affect vulnerable or
    underserved communities?

---

### Compare the Headlines Below

Headline 1: [HEADLINE 1 INSERTED HERE]
Headline 2: [HEADLINE 2 INSERTED HERE]

Your output should be a single line: either
    'Headline 1' or 'Headline 2' - no
    explanation.

---

### Examples

**Example 1**
Headline 1: City Council Approves $20
    Million Affordable Housing Project
Headline 2: Council Discusses Adding Public
    Art
**More Important**: Headline 1

**Example 2**
Headline 1: City Declares 'Local History
    Month'
Headline 2: Council Votes to Close Health
    Clinic Despite Protests
**More Important**: Headline 2
```