LLMs for Extracting Agenda Items from City Council Meetings and Generating Newsworthy Headlines

David Xia*
Mentors: Professor Jeffrey Bigham†, Professor Chris Maury†,‡
*University of Illinois Urbana-Champaign
†Carnegie Mellon University
‡InformUp

Introduction and Motivation

Social media increasingly competes with traditional news sources for public attention, often diverting focus from local news to national headlines. It can undermine trust through the spread of misinformation and unverified commentary [1]. As reliable local journalism declines, so too does civic engagement [2]. This raises a critical question:

- **Can large language models (LLMs) make local public meeting coverage more affordable and accessible?**

To promote local engagement each week, the nonprofit **InformUp** reviews city council meetings, selects the three most important topics, and publishes an article featuring headlines and summaries for each. To automate this task, an LLM would need to **segment** meeting documents, **identify** newsworthy topics, and **generate** engaging, accurate headlines and summaries.

While prior research has explored the use of LLMs for segmentation and accurate summarization [3], there is limited work on their ability to assess newsworthiness and produce compelling headlines. This project focuses on filling that gap. We specifically investigate two key questions:

- How well do LLMs identify engaging topics compared to expert human writers?
- How well do LLMs generate compelling headlines compared to expert human writers?

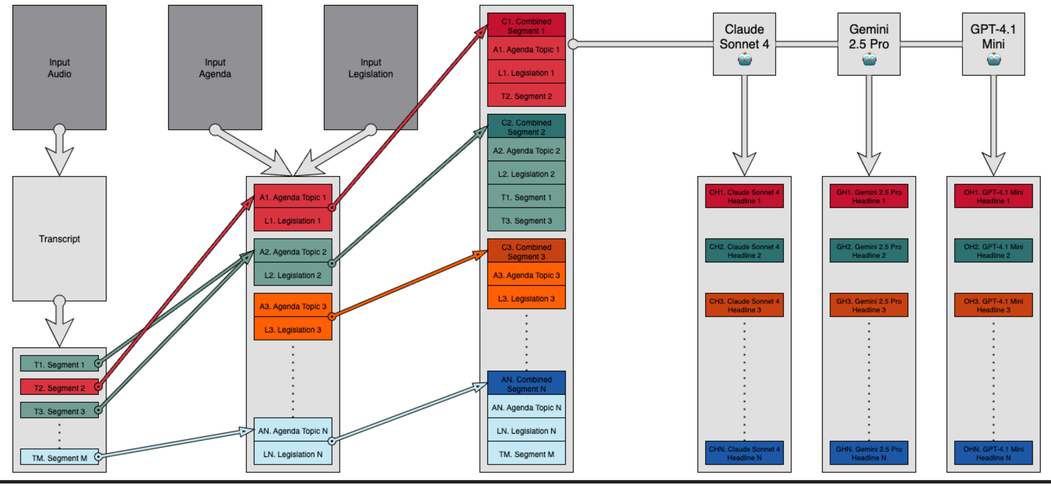
Description of Data

We focus on city council meetings in Pittsburgh due to the availability of meeting recordings, documents, and a pre-existing manual dataset of headlines and summaries. For each Pittsburgh City Council meeting:

- The official City Channel Pittsburgh YouTube Channel uploads the video recording of the meeting. We transcribe the audio to produce a transcript.
- The official meeting agenda is uploaded to the city council website.
- For each agenda item, the corresponding legislation is also uploaded to the city council website.

We manually identify and segment each agenda item, then manually pair it with the corresponding legislation and relevant portions of the meeting transcript. This results in a "combined segment" that brings together the agenda item, its associated legislation, and related discussion from the meeting.

For each combined segment, we prompt Claude 4 Sonnet, Gemini 2.5 Pro, and GPT-4.1 Mini to generate a headline. These models were selected because they offer the best performance within their respective model families while still being able to fit the full combined segment (>100k tokens) within their context windows. This ensures that each model has complete access to all material. We focus on headlines as readers usually decide to read and interact with an article based solely on the headline [4].



Establishing a Human-Evaluated Ground-Truth

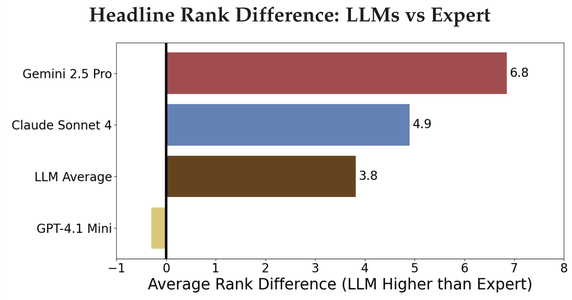
The primary objective of InformUp is to promote civic engagement. Research shows that civic engagement is often rooted in strong emotional or opinionated responses from readers [5]. To evaluate whether LLMs can effectively identify such topics, we develop a metric based on real reader reactions.

We analyzed an average of 30 relevant combined segments per week over a four-week period in April 2025. Of these 30 combined segments, five include pre-existing expert-written headlines, in addition to headlines generated by the three different LLMs.

To collect reader response data, we use the crowdsourcing platform Prolific. On average, 100 participants a week were each presented with 20 randomized pairs of headlines. In each pair, they were asked to select the headline that evoked a stronger opinion or emotional reaction—an approach aligned with our goal of fostering civic engagement. We evaluated all possible pairwise comparisons among the headlines.

We then applied the TrueSkill [6] ranking model to compute a full ranking of headline quality for each of the three LLMs.

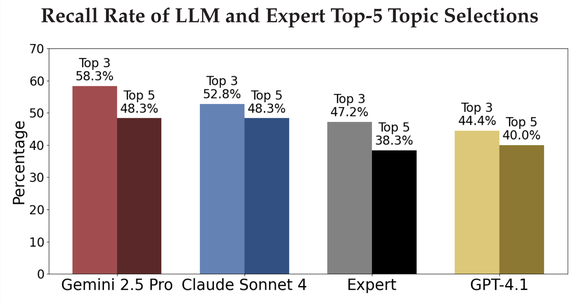
Evaluating Headline Quality



The wording and style of a headline can significantly influence how it is perceived by readers. To evaluate the impact of author style on human perception, we focus on agenda items that have both an LLM-generated headline and a corresponding expert-written version.

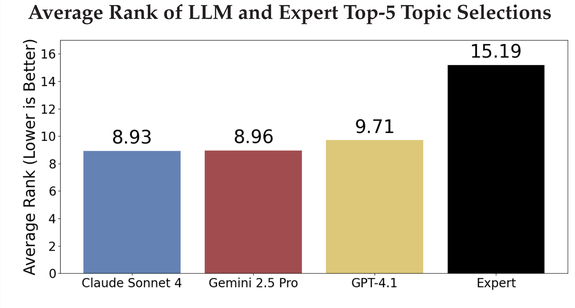
The graph shows the average human-evaluated rank distance between LLM-generated and expert-written headlines for the same topic, using the expert version as a baseline. Higher distances indicate LLM-generated headlines draw stronger opinions than the expert-written headlines.

Selecting Important Topics



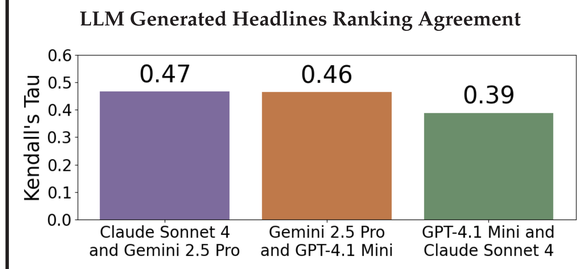
Beyond generating readable and engaging headlines, it is equally crucial that the topics selected by an LLM resonates with readers. To assess topic selection, we prompted Gemini 2.5 Pro, Claude Sonnet 4, and GPT-4.1 to rank topics through pairwise comparisons, using the TrueSkill model [6].

We take the top-5 ranked topics from each LLM as their respective selections. The graph above shows the average proportion of the human-evaluated top-3 and top-5 topics that were included in each author's selected five.



The graph above shows the average true rank of the five topics chosen by each LLM and the expert. A lower rank corresponds to a stronger selection.

LLM Agreement



A strong ranking agreement between LLMs indicates that wording and writing style matter less in evoking human opinion and engagement. The above graph shows the Kendall's Tau [7] score between each LLM's headlines. A score of 1 is perfect match, while a score of -1 is a complete reverse match.

Key Takeaways

- **LLMs outperform expert writers in headline generation.** Gemini 2.5 Pro and Claude Sonnet 4 consistently produced headlines that evoked stronger reader opinion than both the expert and GPT-4.1 Mini, as reflected in the significant positive rank differences.
- **LLMs more effectively identify engaging topics.** Gemini 2.5 Pro and Claude Sonnet 4 selected a greater share of top-3 and top-5 topics, though the advantage here is more modest compared to headline performance. The lower (better) average rank of LLM chosen topics also indicates that LLMs are capable of identifying engaging topics better.
- **In addition to the well-studied tasks of segmentation and summarization, our results suggest that LLMs can effectively automate the affordable and accessible coverage of local public meetings.**
- Given the moderate Kendall's Tau agreement scores between the three LLM generated headline rankings, it appears that the specific wording used by each model **does** influence human perception and opinion.

Next Steps

- **Full pipeline automation.** The transcript, agenda, and legislation were manually segmented and paired, so we could focus our experiments on the quality of the topics selected and headlines generated. We are interested in automating this process by utilizing and fine-tuning LLMs.

Acknowledgements

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

References

[1] M. Caro, "Is Local News Failing To Hold Public Officials Accountable?," *Local News Initiative*, Jun. 06, 2023. https://localnewsinitiative.northwestern.edu/posts/2023/06/06/medill-local-news-poll/

[2] D. Hayes and J. L. Lawless, "The Decline of Local News and Its Effects: New Evidence from Longitudinal Data," *The Journal of Politics*, vol. 80, no. 1, pp. 332–336, Jan. 2018, doi: https://doi.org/10.1086/694105.

[3] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking Large Language Models for News Summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 39–57, Jan. 2024.

[4] N. Siller, "Online headlines shift from concise to click-worthy," *Phys.org*, May 19, 2025. https://phys.org/news/2025-05-online-headlines-shift-concise-click.html

[5] S. Rathje, J. J. Van Bavel, and S. van der Linden, "Out-group animosity drives engagement on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 26, Jun. 2021, doi: https://doi.org/10.1073/pnas.2024292118.

[6] R. Herbrich, T. Minka, and T. Graepel, "TrueSkill(TM): A Bayesian Skill Rating System," *Advances in Neural Information Processing Systems*, vol. 20, pp. 569–576, Jan. 2007.

[7] K. M. G, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938, doi: https://doi.org/10.2307/2332226.