

Multimodal AI Analysis of Politically Sensitive Material: A Case Study on Pepe the Frog



Stephanie Nguyen, Trisha Agrawal, Yuanxi Li, Noor Fatima Hasan
University of Washington

Background

- As LLMs are increasingly competent at categorizing multimodal material, we explore how their assessments shape the qualitative dimensions of the datasets they process
- Building on research on the use of AI for image captioning and analysis¹, this project compares how LLMs versus human raters assess social media posts

Methodology & Research Design

- Analyzed 3,407 tweets (image + caption) pertaining to the Pepe the Frog meme, sourced from Twitter in the months surrounding the Jan 6th 2021 Capitol riot
- Developed a codebook with four categories: contextual references, political positionality, hate content, and sexually explicit content
- Queried four models (Claude Opus 4 and Sonnet 4, Mistral Pixtral 12B, and LLaVa-NeXT) with zero-shot prompting to generate categorizations for each tweet
- Compared LLM categorizations with three human raters' categorizations, with minimally acceptable agreement at a benchmark of 0.55

Results

► Finding 1: LLMs identified political positioning less confidently than human raters, especially in right-wing content.

Figure 2. Distribution of Political Position

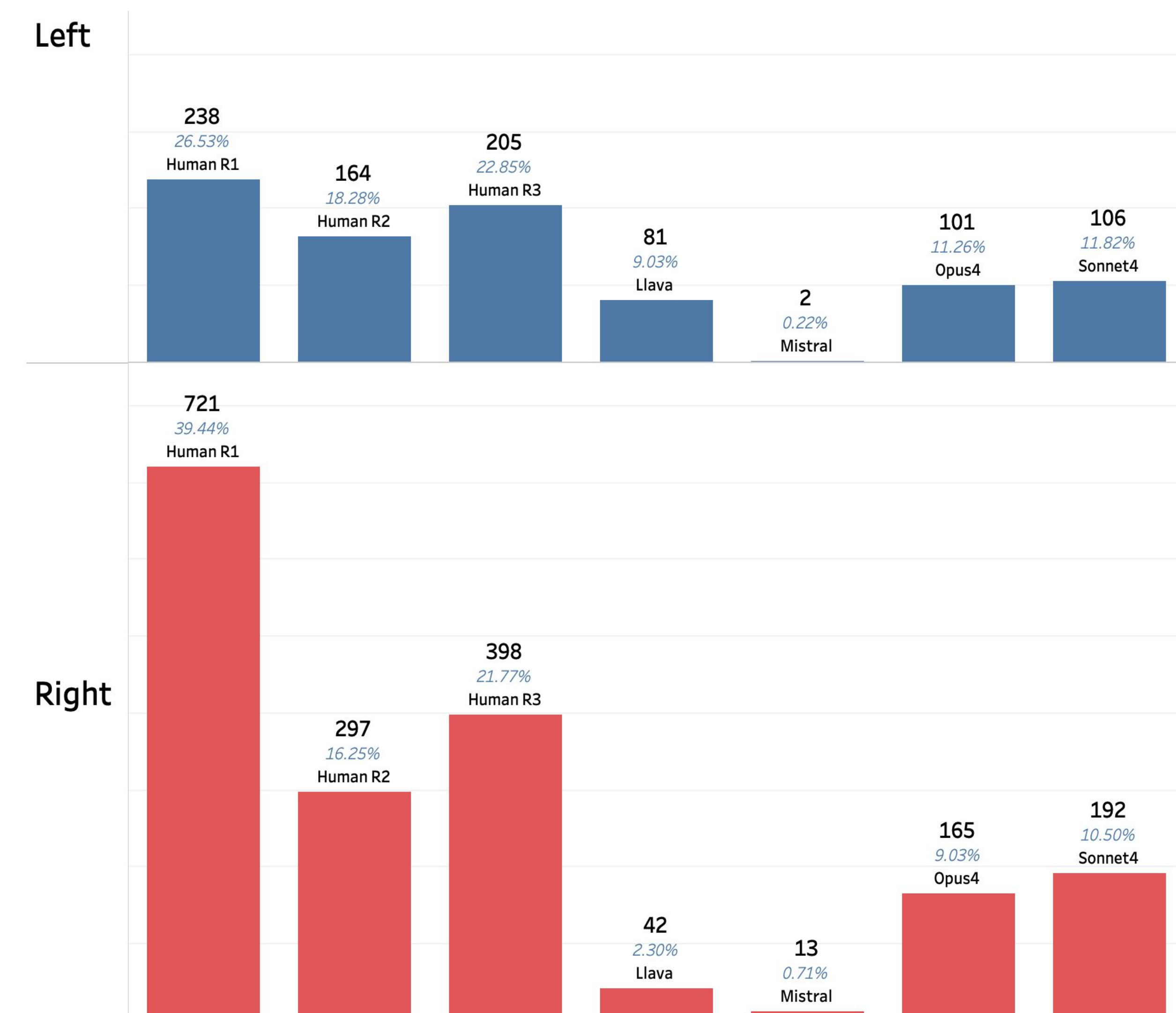


Figure 1. Poster for *Feels Good Man* (2020). Source: Wikipedia



Implications & Future Directions

- Downplaying of individual + social responsibility
- Majority of our data consisted of 4chan screenshots reposted on Twitter
- Potential functionality of LLMs for classification tasks in humanities research

Acknowledgement

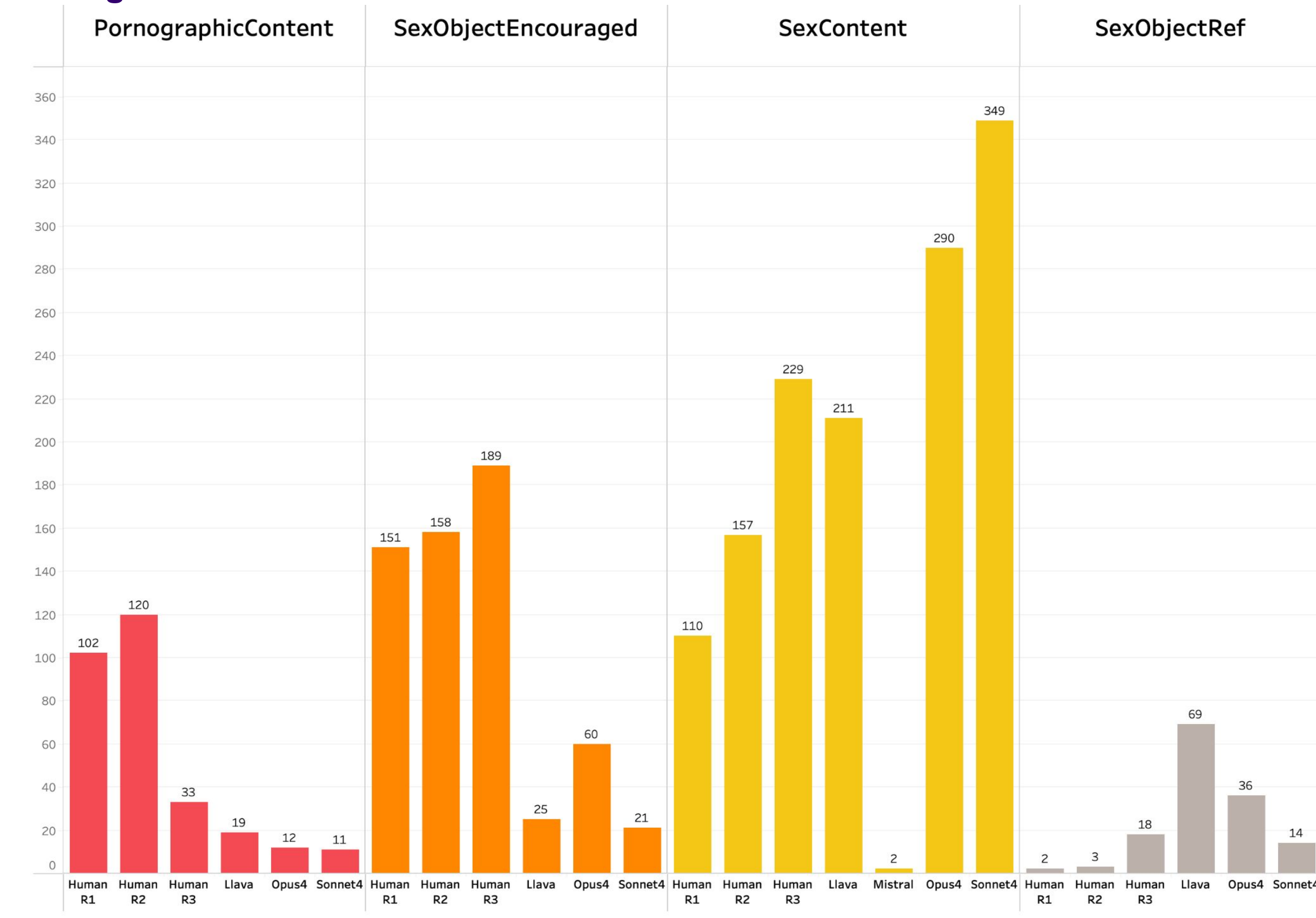
We thank our PI Adair Rounthwaite for her leadership and RA Nikoloz Nadirashvili for his contributions throughout this project. Many thanks to the Humanities Data Science Summer Institute and eScience Institute for hosting this research program and providing funding. This project was also supported through funding from the UW Research Computing Club.

References

- Shamshiri, Alireza, Kyeong Rok Ryu, and June Young Park. 2024. "In-Context Learning for Long-Context Sentiment Analysis on Infrastructure Project Opinions." *arXiv*. <https://doi.org/10.48550/arxiv.2410.11265>.

► Finding 2: LLMs tended to use less severe labels for sensitive content, including hate, abuse, or explicit sexual material.

Figure 3. Distribution of Sexual Content



► Finding 3: Political positioning had the lowest agreement, showing it was the most subjective category.

Figure 4. Heatmap of Cohen's Kappa Scores Across Categories

Rater pair	H1-H2	H2-H3	H3-Sonnet	H3-Opus	H2-Opus	H2-Sonnet	H1-H3	H1-Sonnet	H1-Opus	Opus-Llava	Sonnet-Llava	H3-Llava	H2-Llava	H1-Llava	Opus-Mistral	Sonnet-Mistral	H3-Mistral	H1-Mistral	H2-Mistral	Mistral-Llava
	0.9050	0.8870	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130	0.8130
	0.9360	0.6960	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850	0.5850
	0.8030	0.6380	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080	0.4080
	0.8590	0.5010	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330	0.4330
	0.8720	0.5040	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480	0.4480
	0.7920	0.4420	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290	0.4290
	0.7950	0.4440	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070	0.4070
	0.7290	0.5960	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330	0.3330
	0.7100	0.3860	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640
	0.7090	0.3860	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740	0.3740
	0.0310	0.0800	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520	0.0520
	0.0300	0.0800	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490	0.0490
	0.0290	0.0490	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450	0.0450
	0.0230	0.0550	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430	0.0430
	0.0220	0.0450	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410	0.0410

Contextual Sexual Hate & Abuse Political Position

