Launch and Aftermath: Contrasting Social Media Responses to Chatbot Releases. The Cases of Meta's Galactica and OpenAI's ChatGPT

Maximilian Weber[1], **Johannes B. Gruber**[2]

[1]University of Mainz, Germany, [2]GESIS, Germany

Correspondence: maximilian.weber@uni-mainz.de

Abstract

In November 2022, Meta's Galactica and OpenAI's ChatGPT were released within fifteen days of each other, two transformer-based language models that were architecturally similar and built on comparable underlying technology, yet experienced starkly different outcomes. Where they diverged was not in technical kind but in domain positioning and epistemic framing: Galactica was explicitly marketed as a reliable scientific assistant, while ChatGPT was presented as a general-purpose conversational tool. Using Twitter data collected via the Twitter Research API, we conduct a comparative analysis of early social media discourse surrounding both models. Through sentiment classification, zero-shot harm and risk annotation, and LLM-based topic modeling, we find that negative sentiment escalated rapidly for Galactica while remaining comparatively stable for ChatGPT in the release period. Galactica experienced a marked escalation in criticism during its first week, eventually structuring much of the conversation. In contrast, ChatGPT's early discourse remained more evenly distributed across hype, experimentation, practical engagement, and criticism. We argue that domain positioning and epistemic expectations, rather than any meaningful technological difference, played a central role in shaping public perception, with Galactica's scientific presentation making its well-documented hallucinations appear far more damaging in public opinion.

1 Introduction

In November 2022, two large language models (LLMs) were released to the public with strikingly different trajectories. Meta introduced Galactica on November 15, 2022, a specialized LLM designed explicitly for scientific applications. Introduced as capable of summarizing academic literature, solving mathematical problems, generating Wikipedia articles, writing scientific code, and annotating molecules and proteins, the model was positioned as a tool for the research community (Taylor et al., 2022). However, within just two days of its public demo launch, Meta withdrew the service following criticism. According to Meta's Chief AI Scientist Yann LeCun, the model was effectively driven offline by public backlash: "Galactica, the LLM for scientists from Meta [...]. It was murdered by a ravenous Twitter mob. The mob claimed that what we now call LLM hallucinations was going to destroy the scientific publication system" (Yann LeCun [@ylecun], 2023).

In contrast, OpenAI's ChatGPT, released just fifteen days later on November 30, 2022, experienced rapid and widespread adoption, becoming a mainstream phenomenon that reportedly reached 700 million weekly active users within three years of its launch.[1]

This contrast provides a quasi-experimental opportunity to examine the two model launches: Why did one model face shutdown within days, while the other achieved remarkable success? Were the concerns raised about Galactica equally applicable to ChatGPT, but less central to the overall public discourse?

This study addresses three research questions through comparative analysis of social media discourse. First, how did public discourse differ between the two launches? Second, did harm- and risk-related discourse escalate within Galactica tweets over time? Third, did the target domain (scientific knowledge versus general-purpose use) affect public reception?

2 Previous research

While a growing body of research has examined public discourse around LLM releases on social media, particularly around ChatGPT (Koonchanok et al., 2024; Weber, 2024; Rauchfleisch et al.,

[1] https://openai.com/index/how-people-are-using-chatgpt/ (OpenAI usage report, as of mid-2025).

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 95–102

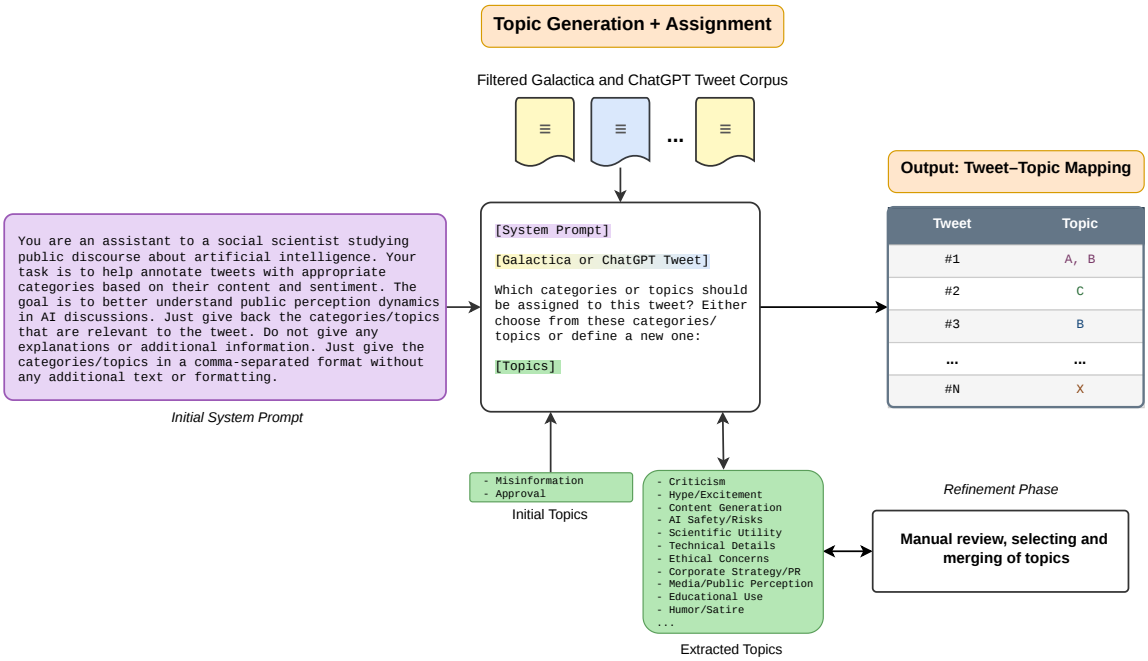

Figure 1: Overview of the Topic Generation and Assignment Pipeline. Given a filtered corpus of Galactica and ChatGPT tweets and a set of manually-curated initial topics, an LLM iteratively assigns topics to each tweet. After every 100 documents, the framework enters a refinement phase in which topics are manually reviewed, merged, and selected, yielding a final curated topic list and a topic assignment for all sampled documents.

2025), comparative analyses of social media reactions to Galactica versus ChatGPT remain absent from the literature. Chartier-Edwards et al. (2024) provide a critical account of the Galactica release, arguing that the controversy reflects tensions between AI for science promissory marketing claims, and epistemic expectations in scientific domains. They further highlight concerns about scientific misinformation and the framing of LLMs as epistemic oracles. However, they do not conduct a large-scale analysis of social media discourse, which we address in this study.

3 Data and Methods

3.1 Data

We conduct an analysis of English-language Twitter discourse during the initial release periods of both models. Using the Twitter Research API, we collected 109,694 initial tweets: 2,077 tweets mentioning "Galactica" from November 15 through December 15, 2022, and 107,726 tweets mentioning "ChatGPT" from November 30 through December 7, 2022, with 283 tweets mentioning both systems. To ensure comparability, our main analyses focus on the first eight days following each release.

Since the term *Galactica* is also associated with unrelated content on social media, such as the television series *Battlestar Galactica*, we first filtered out tweets that did not refer to Meta's Galactica model. To do so, we randomly sampled 250 tweets mentioning Galactica and manually annotated them as relevant or irrelevant. We then embedded all tweet texts using the snowflake-arctic-embed2 model (via rollama: Gruber and Weber, 2024) and trained a logistic regression classifier with LASSO regularization. The classifier achieved an accuracy of 0.908 and a macro F1 score of 0.873, after which it was applied to classify all remaining tweets mentioning Galactica.

3.2 Sentiment Annotation

To assess the emotional tone of tweets, we employed a RoBERTa-based classifier fine-tuned on Twitter data (Loureiro et al., 2022) and trained to annotate tweets for negative, neutral, and positive sentiment[2]. Each tweet was preprocessed following the conventions (replacing @mentions with `@user` and URLs with `http`) before classification. The model assigns a probability score to each of the three sentiment classes; the class with the highest

[2] `cardiffnlp/twitter-roberta-base-sentiment-latest`

score is taken as the predicted sentiment.

3.3 Harms and Risks Annotation

To identify Galactica tweets discussing potential harms or risks, we employed zero-shot annotation using `meta-llama/Llama-3.3-70B-Instruct` (Grattafiori et al., 2024), loaded in 4-bit NF4 quantization from Hugging Face. Each tweet was passed to the model with the following prompt: *Is this tweet about potential harm and risk of the AI model (LLM) Galactica? Answer with just yes or no.* Classification was based on the normalized probabilities of yes and no tokens. The system prompt instructed the model to act as an assistant to a social scientist studying public discourse about AI, providing context that tweets mentioning Galactica refer to Meta's large language model designed to assist scientific research. To evaluate annotation quality, two human annotators independently labeled a random sample of 100 tweets, achieving an inter-annotator agreement of $\kappa = 0.67$ (84% agreement). Evaluated against each annotator separately, the model achieved a κ of 0.688 and a macro-average F1 of 0.843 against annotator 1, and a κ of 0.766 and a macro-average F1 of 0.883 against annotator 2, indicating substantial agreement between automated annotations and human judgment.

3.4 Topic Classification

Recent work has explored the use of generative large language models for topic modeling and theme extraction through prompt-based frameworks (Pham et al., 2024; Sharma, 2025; Liu et al., 2025; van Wanrooij and Manhar, 2024). These approaches typically leverage generative LLMs to generate candidate topics and refine them through iterative prompting. We adopt a similar pipeline to Pham et al. (2024), as can be seen in Figure 1, but omit their second annotation phase in which a final fixed topic list is used to label a held-out set of documents. The resulting LLM-based zero-shot topic modeling and annotation pipeline incorporates a human-in-the-loop refinement stage, in which topics are reviewed and consolidated by the researchers. This design gives us direct control over the granularity and coherence of the final topic set, ensuring that the extracted themes are both meaningful and well-suited to the domain of AI discourse on social media.

Topic generation and assignment were performed using the open-weight Llama 3.3 70B (`llama3.3:70b-instruct-q4_K_M`), an instruction-tuned variant with 4-bit quantization, run via Ollama. The system prompt framed the task as assisting a social scientist in annotating public discourse about artificial intelligence, and restricted output to a comma-separated list of topic labels without additional explanation. Topic classification was applied to all tweets in both corpora: a random subset of 1,640 tweets for ChatGPT due to computational constraints, and the full set of Galactica tweets.

Day	Galactica (%)	ChatGPT (%)	Sig.
1	5.1 [2.0, 12.5]	12.0 [9.5, 15.2]	
2	12.1 [8.9, 16.1]	17.5 [16.6, 18.5]	*
3	35.6 [30.3, 41.3]	18.0 [17.4, 18.7]	***
4	33.6 [26.4, 41.6]	18.8 [18.1, 19.6]	***
5	49.1 [39.9, 58.3]	18.0 [17.4, 18.6]	***
6	42.4 [32.8, 52.6]	20.3 [19.8, 20.8]	***
7	31.2 [24.5, 38.8]	20.2 [19.7, 20.8]	**
8	35.8 [28.2, 44.1]	21.2 [20.6, 21.8]	***

Table 1: Negative sentiment rate (%) for the first 8 days after each model launch. Wilson 95% CIs in brackets. Significance based on Fisher's exact test (BH-adjusted): * $p < .05$, ** $p < .01$, *** $p < .001$.

4 Results

Table 1 and Figure 2 illustrate the temporal dynamics of sentiment during the first eight days following each model's release. For Galactica, negative sentiment accounts for only 5.1% of tweets on Day 1. However, this share rises rapidly over the subsequent days, reaching 35.6% by Day 3 and peaking at 49.1% on Day 5. Following the announcement that the demo was taken offline (November 17), overall tweet volume declines substantially.

ChatGPT exhibits a markedly different pattern. Although tweet volume increases steadily after release, reaching substantially higher levels than Galactica, negative sentiment remains comparatively stable throughout. On Day 1, 12.0% of ChatGPT tweets are classified as negative, and this share fluctuates between 17% and 21% across the first week. Unlike Galactica, ChatGPT does not experience a comparable escalation in negative sentiment.

As shown in Table 1, differences between the two models are statistically significant on all days except on the respective release date (Fisher's exact test), particularly from Day 3 onward. Overall, negative sentiment intensified markedly for Galactica while remaining stable for ChatGPT.

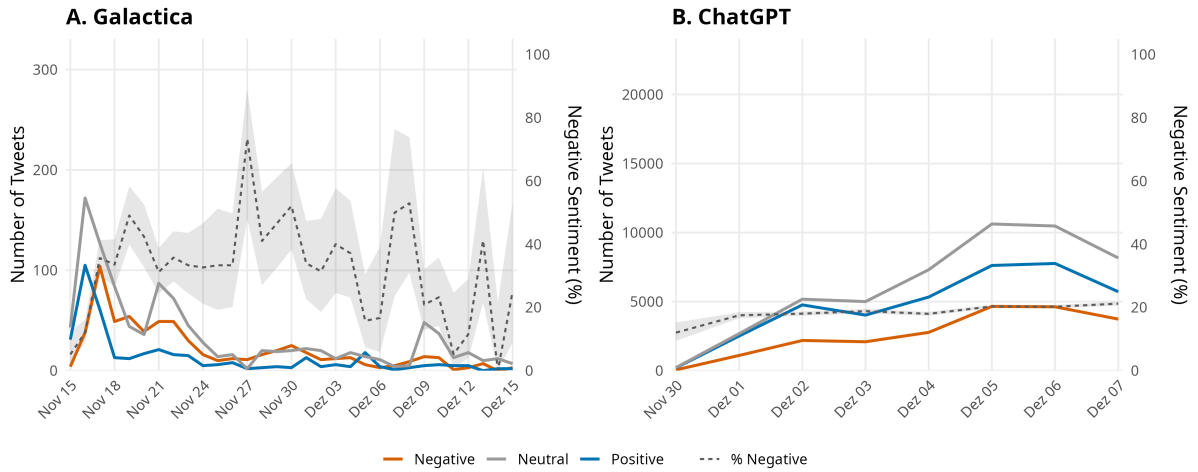

Figure 2: Temporal dynamics of Galactica (Figure A) and ChatGPT (Figure B) tweets. Lines show daily sentiment counts; the dashed line indicates the percentage of negative-sentiment tweets.

Focusing on Galactica alone, potential harm and risk annotation follows a similar pattern (Table 3). Harm-related tweets constitute only 8.3% on Day 1, but this share rises rapidly, reaching 43.9% by Day 3 and 57% by Day 5. Thus, while criticism was not dominant at launch, it came to structure a majority of Galactica discourse within the first week. Notably, following the model being taken offline on November 17, discourse shifted further toward harm- and risk-related concerns, suggesting that the withdrawal itself amplified rather than resolved the debate.

Figures 3 and 4 show the distribution of topics identified via zero-shot topic classification. ChatGPT discourse is dominated by *Innovation*, followed by *Technical Details*, *Hype/Excitement*, and *Criticism*. Although *Criticism* and *Misinformation* are present, they do not structure the majority of the conversation.

In contrast, Galactica-related discourse is led by *Criticism*, followed by *Scientific Utility* and *AI Safety/Risks*. Hype-related categories are comparatively less dominant. This suggests that Galactica's reception became increasingly structured around concerns about epistemic risk and factual reliability, particularly after the first three days following its release. During the initial three days (marked in red), however, tweets more frequently focused on *Scientific Utility* and technical details, with *Criticism* becoming more prominent thereafter.

This suggests that while concerns were present in the discourse around both models, they were relatively more prominent in the case of Galactica, particularly regarding misinformation, truthfulness, credibility, and scientific legitimacy. ChatGPT discourse, while not free of criticism, was more broadly characterized by exploratory and practical engagement.

Model	Example Tweet
Galactica	"Is this really what AI has come to, automatically mixing reality with nonsense so finely we can no longer recognize the difference?"
Galactica	"Shocked that it only took a handful of questions before Meta's new Galactica model produced racist garbage when asked about linguistic prejudice."
Galactica	"A whole new level of AI-generated academic misconduct to deal with now."
ChatGPT	"It boggles my mind how the world keeps spinning like nothing happened despite #ChatGPT. People don't understand the danger."
ChatGPT	"#ChatGPT could make it easy to cheat on written tests and homework. You can no longer give take-home exams."
ChatGPT	"ChatGPT is down and I'm having an existential crisis because I can't paste my code in. Please come back."

Table 2: Illustrative tweets from the early release period of Galactica and ChatGPT. Tweets are lightly edited for clarity and anonymized.

5 Discussion

Returning to our research questions, the findings reveal differences in the public trajectories of the two model launches. First, public discourse differed not only in tone but in temporal dynamics: while both models were accompanied by early criticism, Galactica's reception shifted rapidly toward potential harm- and risk-centered discourse, which came

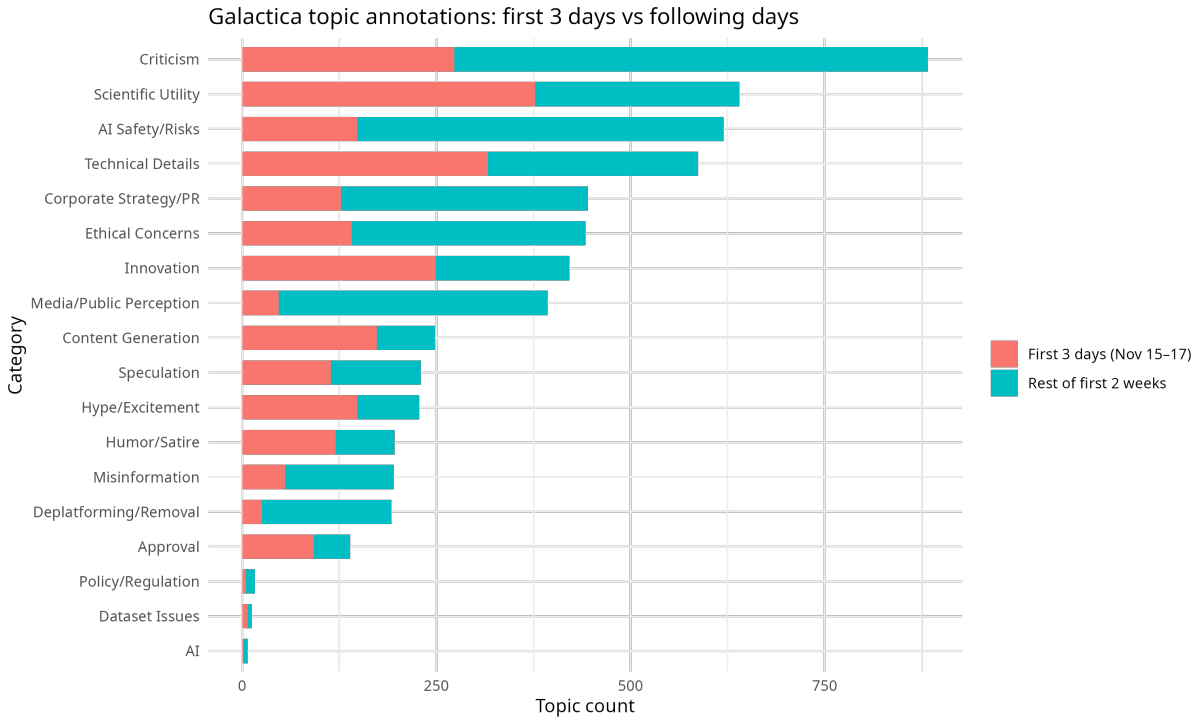

Figure 3: Topic distribution in Galactica-related tweets during the first 2 weeks following release, derived from zero-shot topic classification using Llama 3.3 (70B).

to structure a majority of tweets within days. ChatGPT discourse, by contrast, expanded in volume while maintaining a comparatively stable mix of enthusiasm, experimentation, practical engagement, and critique. Second, these patterns are consistent with the interpretation that domain positioning shaped public reception (Chartier-Edwards et al., 2024). Arguably, it is possible that ChatGPT performed better for users, leaving them with a more positive outlook on the technology. However, Taylor et al. (2022) themselves benchmarked Galactica against `text-davinci-002`, a model in the GPT-3 family (Brown et al., 2020) that underpinned the initial release of ChatGPT, and found that Galactica outperformed it on scientific knowledge probes and several bias and toxicity benchmarks. The main technical difference was the reinforcement learning from human feedback (Ouyang et al., 2022) applied in `text-davinci-003`, the model ChatGPT used at launch, which was intended in part to make the model less prone to assert falsehoods with confidence. Yet even this technical difference reflects a broader strategic divergence that led to the outcome we observed: Galactica's framing as a reliable scientific assistant likely heightened epistemic expectations, making hallucinations and factual errors normatively consequential within a domain where credibility is central. ChatGPT's general-purpose positioning, in contrast, appears to have allowed criticism to coexist with hype rather than dominate the discourse. While our design does not permit strong causal claims, the comparative evidence suggests that epistemic framing and expectation management play a critical role in shaping the early public legitimacy of AI systems.

5.1 Limitations

This study has several limitations. Our dataset comprises only original tweets, excluding retweets, and is restricted to the initial release period of both models, leaving unexamined longer-term shifts in discourse. Additionally, topic classification for ChatGPT was conducted on a random subsample of 1,640 tweets, which may not capture the full breadth of discussion around the model.

Furthermore, since the tweets analyzed predate the release of the Llama 3.3 models, it is possible that some of this content was included in the models' pretraining or fine-tuning data. This could introduce bias, as the model may have been exposed to content about Galactica and ChatGPT during training.

Following Galactica's withdrawal on November 17, discourse necessarily became retrospective, as users could no longer interact with the model.

This asymmetry with ChatGPT, which remained live throughout, should be considered when interpreting the results.

6 Conclusion

Our analysis reveals differences in public reception between Galactica and ChatGPT. Galactica faced substantially more criticism, particularly centered on concerns about misinformation and the generation of false or misleading scientific information. ChatGPT, by contrast, generated more positive discourse characterized by hype, experimentation, and demonstrations of utility, despite being subject to many of the same technical limitations. This comparative case study highlights the role of domain positioning, epistemic expectations, and expectation management in shaping public acceptance of AI technologies.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Nicolas Chartier-Edwards, Etienne Grenier, and Valentin Goujon. 2024. Galactica's dis-assemblage: Meta's beta and the omega of post-human science. *AI & SOCIETY*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Johannes B. Gruber and Maximilian Weber. 2024. rollama: An R package for using generative large language models through Ollama. *arXiv preprint*. ArXiv:2404.07654 [cs].

Ratanond Koonchanok, Yanling Pan, and Hyeju Jang. 2024. Public attitudes toward chatgpt on twitter: sentiments, topics, and occupations. *Social Network Analysis and Mining*, 14(1):106.

Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, and Yining Li. 2025. LLM-guided semantic-aware clustering for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18420–18435, Vienna, Austria. Association for Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 27730–27744, Red Hook, NY, USA. Curran Associates Inc.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Adrian Rauchfleisch, Joshua Philip Suarez, Nikka Marie Sales, and Andreas Jungherr. 2025. Winning and losing with Artificial Intelligence: What public discourse about ChatGPT tells us about how societies make sense of technological change. *Telematics and Informatics*, 103:102344.

Yash Sharma. 2025. MALTopic: Multi-Agent LLM Topic Modeling Framework. In *2025 IEEE World AI IoT Congress (AIIoT)*, pages 0707–0712, Seattle, WA, USA. IEEE.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

Cascha van Wanrooij and Omendra Kumar Manhar. 2024. Topic Modeling for Small Data using Generative LLMs. In *Proceedings of the 36th Benelux Conference on Artificial Intelligence (BNAIC) and the 33rd Belgian-Dutch Conference on Machine Learning (BeNeLearn) 2024*, Utrecht, Netherlands.

Maximilian Weber. 2024. Social Group Differences in the Social Media Discussion about ChatGPT and Bing Chat. In *ACM Web Science Conference*, pages 114–118, Stuttgart Germany. ACM.

Yann LeCun [@ylecun]. 2023. Galactica, the LLM for scientists from Meta, was released a couple of weeks before ChatGPT but was taken down after... [Post].

A Appendix

Day	% Harms/Risks	95% CI
1	8.3	[3.9, 17]
2	21.5	[17.1, 26.7]
3	43.9	[38.2, 49.8]
4	45.0	[36.4, 53.9]
5	57.3	[47.6, 66.4]
6	57.6	[47, 67.6]
7	55.7	[47.2, 63.9]
8	54.5	[45.2, 63.4]

Table 3: Potential harms/risks prediction rate (%) for the first 8 days after Galactica's launch. 95% CIs based on Wilson score interval.

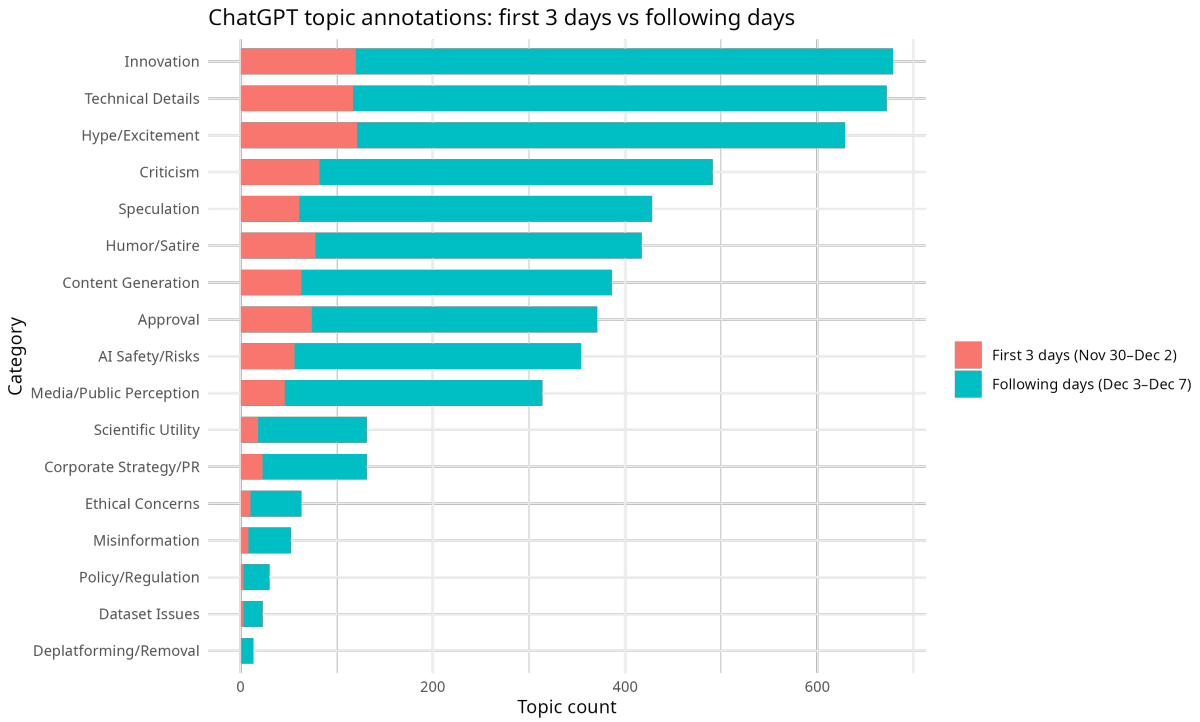

Figure 4: Topic distribution for ChatGPT tweets during the first eight days following release. Topics are derived from zero-shot topic classification using Llama 3.3 (70B) applied to a random subsample of 1,640 tweets.

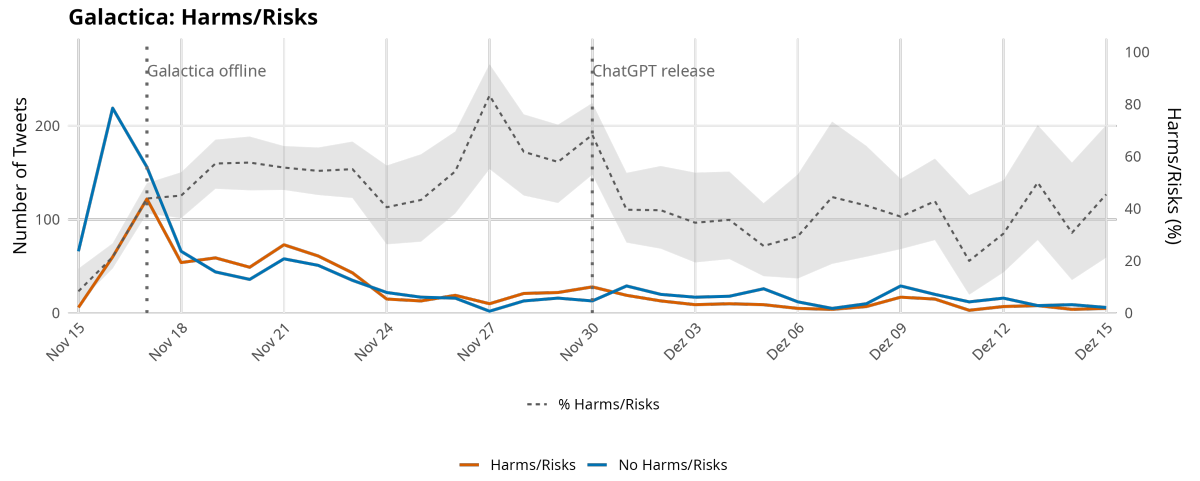

Figure 5: Temporal dynamics of potential harm/risk predictions for Galactica tweets. Lines show daily counts of predicted harms/risks and no harms/risks; the dashed line indicates the percentage of tweets classified as harms/risks, with shaded band representing the 95% Wilson confidence interval.