Can LLMs Be Efficient Predictors of Conversational Derailment?

Kaustubh Olpadkar and Vikram Sunil Bajaj and Leslie Barrett
Bloomberg
{kolpadkar, vbajaj11, lbarrett4}@bloomberg.net

Abstract

Conversational derailment — when online discussions stray from their intended topics due to toxic or inappropriate remarks, is a common issue on online platforms. These derailments can have negative impacts on users and the online community. While previous work has focused on post hoc identification of toxic content, recent efforts emphasize proactive prediction of derailments before they occur, enabling early moderation. However, forecasting derailment is difficult due to the context-dependent emergence of toxicity and the need for timely alerts. We prompt pretrained large language models (LLMs) to predict conversational derailment without task-specific fine-tuning. We compare a range of prompting strategies, including chain-of-thought reasoning (CoT) and few-shot exemplars, across small and large-scale models, and evaluate their performance and inference-cost trade-offs on derailment benchmarks. Our experiments show that the best prompting configuration attains state-of-the-art performance, and forecasts derailments earlier than existing approaches. These results demonstrate that LLMs, even without fine-tuning, can serve as effective tools for proactive conversational moderation.

1 Introduction

The task of forecasting conversation derailment, defined as the proactive prediction of future antisocial behavior in otherwise civil online discussions, has received increasing attention in recent years. Earlier research primarily focused on post hoc classification of toxic behavior (Zhang et al., 2018); however, recent advancements aim to detect early warning signs of derailment to enable timely interventions. This task is unusual and challenging because the predictive model must take into account previous and current conversational turns, not simply the dialog as a whole. Also, conversations can "derail" in various ways, and different datasets reflect particular properties of this phenomenon. For example in the Conversations Gone Awry (CGA) dataset (Zhang et al., 2018) conversations are task-oriented dialogues between Wikipedia editors; in contrast, the Reddit ChangeMyView (CMV) corpus (Tan et al., 2016) covers a broader range of topics. In both datasets, a derailment indicates that the conversation devolved into personal attacks (i.e., became "toxic"). These represent the two main datasets commonly used for this task.

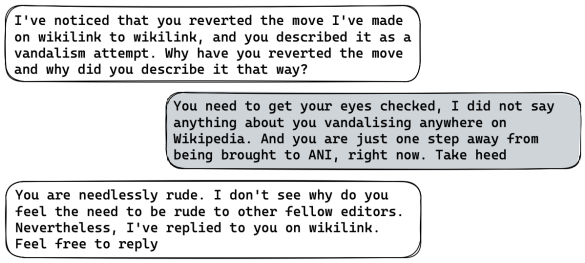

Figure 1: Sample conversation containing a personal attack.

In this paper, we conduct experiments using LLMs to predict the existence of a derailment in a conversation as early as possible, and address the following research questions:

RQ1: Are LLMs able to predict conversational derailments?
RQ2: Can LLMs predict derailments as early as current models?
RQ3: How do LLMs compare on performance and cost parameters against current models?

Our results on two datasets with four LLMs show that these models excel at the prediction task itself and are able to predict the derailment quite early. We use only the data present in these datasets, without additional features or fine-tuning. In addition, we analyze the costs and benefits of deploying such models in real-world scenarios where toxic conversations may need to be detected dynamically in real time.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15104–15112
November 4-9, 2025

2 Previous Approaches

A clear methodological trajectory emerges across previous studies. Zhang et al. (2018) establish the foundational task and dataset, relying on static linguistic features and crowd-annotated personal attacks. Chang and Danescu-Niculescu-Mizil (2019) move to dynamic prediction via hierarchical RNNs (CRAFT), explicitly modeling comment-to-comment interactions over time. Kementchedjhieva and Søgaard (2021) further refine this by proposing dynamic training aligned with the inference-time setup, using BERT-based sequential models and revealing trade-offs in noisy data. Later models shift toward deeper architectural representations. Yuan and Singh (2023) adopt a hierarchical transformer design and introduce multitask learning with a novel regression head for time-to-derailment, enhancing interpretability and early warning capacity. Altarawneh et al. (2023) go beyond sequence modeling by introducing a graph-based approach (FGCN), capturing multi-party dynamics through user relationships and community signals such as voting. This model demonstrates state-of-the-art performance, particularly in terms of forecast horizon, as does its knowledge-aware version in Altarawneh et al. (2024). Some approaches use additional information either exploiting aspects of the model itself or adding additional elements to the corpus. Sicilia and Alikhani (2024) leverage uncertainty estimation in derailment forecasting, while Nonaka and Yoshida (2025) add contextual information to the data, including the corpus name with a short description in the prompt.

All current research tends to emphasize the gradual, emergent nature of toxic exchanges and the necessity for models that operate incrementally as conversations unfold. Each approach aims to forecast antisocial outcomes before they manifest explicitly, offering a window for intervention.

Forecast horizon, a metric indicating how early a derailment is predicted, is a key comparative axis. The baseline CRAFT model achieves an average of three turns, while FGCN-T+ extends this to over four turns on CMV. The introduction of dynamic training improves early detection in some datasets (e.g., CGA), but may suffer on noisier data as evidenced by performance drops in BERT.SC+ compared to FGCN-T+.

3 Data and Methodology

Each sample within each dataset is a single conversation containing a set of conversational turns or utterances

$$C = \{t_1 ... t_n\} \tag{1}$$

where n is the number of utterances. Each conversation is marked with at least one indicator of toxicity. The Conversations Gone Awry (CGA) dataset (Zhang et al., 2018) contains conversations from Wikipedia talk pages that derail into personal attacks, with 4,188 conversations. Metadata therein contains a label indicating whether the conversation devolved into a personal attack as judged by 3 crowd-sourced annotators. The Reddit ChangeMyView (CMV) dataset (Tan et al., 2016) contains 6,842 conversations. In this dataset, conversations are considered toxic or derailed in cases where comments were deleted by moderators following Rule 2 of the forum itself which prohibits rude or antisocial behavior. The conversations in those cases only include turns up to, but not including, the toxic comment. We accessed both datasets through the Convokit software developed by Cornell University (Chang et al., 2020).

To answer RQ1, we developed prompts for each dataset in a zero-shot, few-shot, and CoT-reasoning contexts, asking the model to predict whether a conversation will derail given the conversation so far at each utterance. We develop separate prompts with general instructions and dataset-specific instructions for derailment forecasting. The output is a boolean value, True if the conversation is or will be derailed by a personal attack, False otherwise. Examples of the prompts are in the Appendix D. We primarily run four models, GPT-4o (OpenAI, 2024b), GPT-4o-mini (OpenAI, 2024a), Claude 3.5 Sonnet (Anthropic, 2024b) and Claude 3.5 Haiku (Anthropic, 2024a). For completeness, we also evaluate open-source Llama 3.1 (Grattafiori et al., 2024) variants (8B, 70B, 405B); full results and cost analyses are provided in Appendix C.

Our baselines include CRAFT (Chang and Danescu-Niculescu-Mizil, 2019), BERT.SC (Kementchedjhieva and Søgaard, 2021) and the hierarchical transformer (Yuan and Singh, 2023), FCGN (Altarawneh et al., 2023), and KA-FGCN-BRT (Altarawneh et al., 2024).

For each of our models, we report results for the task of predicting the conversational outcome and also the mean forecast horizon associated with that prediction, as our approach to answering RQ2.

Model	CGA				CMV			
	Acc	P	R	F1	Acc	P	R	F1
CRAFT	64.4	62.7	71.7	66.9	60.5	57.5	81.3	67.3
BERT.SC	64.7	61.5	79.4	69.3	62.0	58.6	82.8	68.5
Hierarchical-Multi	65.2	62.3	76.9	68.9	64.2	62.0	73.8	67.4
FGCN	66.9	63.3	80.2	70.8	64.7	60.7	83.3	70.2
KA-FGCN-BRT	67.4	63.7	81.0	71.3	66.6	62.7	82.1	71.1
GPT-4o	**68.3**	63.5	86.4	**73.2**	63.7	59.1	88.7	71.0
GPT-4o-mini	62.6	58.0	91.2	70.9	56.4	53.7	94.2	68.4
Claude 3.5 Sonnet	62.9	58.0	93.3	71.5	**64.7**	59.8	89.3	**71.7**
Claude 3.5 Haiku	63.5	59.1	87.6	70.6	61.7	57.9	85.8	69.1

Table 1: Performance comparison: Accuracy (Acc), Precision (P), Recall (R), F1 score.

Model	CGA	CMV
CRAFT	2.36	4.01
BERT.SC+	2.85	4.06
Hierarchical-Multi	2.98	3.78
FGCN+	2.96	4.12
KA-FGCN-BRT+	3.02	4.16
GPT-4o	3.34	4.47
GPT-4o-mini	**3.89**	**5.10**
Claude 3.5 Sonnet	3.70	4.50
Claude 3.5 Haiku	3.44	4.59
GPT-4o++	3.84	4.65
GPT-4o-mini++	**4.31**	**5.78**
Claude 3.5 Sonnet++	4.18	4.71
Claude 3.5 Haiku++	3.80	4.94

Table 2: Performance comparison: Mean Forecast Horizon (H). "+" denotes dynamically trained models; "++" denotes results without the heuristic to withhold first-utterance predictions.

Details of how we report model forecasting results are found in 4.1.

Finally, we address RQ3 by analyzing the inference-cost trade-offs of our approach on these two benchmark datasets.

4 Results and Discussion

We evaluate performance using standard classification metrics: Accuracy, Precision, Recall, and F1 score. F1 score will serve as our primary metric, as it balances precision and recall and is particularly important when dealing with the asymmetric costs of false positives and false negatives in conversational derailment forecasting. In addition to these standard metrics, we also report the mean forecast horizon (H), defined as the average number of conversational turns before derailment at which the model successfully issues a warning. A longer forecast horizon indicates earlier and more actionable derailment predictions, enabling more effective intervention, so larger values are better. All metrics are calculated following the same evaluation protocol used in prior works to ensure consistency and comparability. Each model is evaluated on the held-out test sets of both the CGA and CMV datasets. Our results are presented in Table 1 and Table 2.

For the CGA dataset, GPT-4o and Claude 3.5 Sonnet both outperform the current best-performing models. For the CMV dataset, Claude 3.5 Sonnet beats the current best-performing model, with GPT-4o showing comparable results to the FGCN baseline. We explored a variety of prompting strategies for each model and report the configurations that achieved the best performance. For GPT-4o, optimal results were obtained using 200-shot general instructions for CGA and 100-shot task-specific instructions for CMV, both without CoT. Claude 3.5 Sonnet performed best with zero-shot general instructions for CGA and 10-shot general instructions with CoT for CMV. GPT-4o-mini showed the highest effectiveness with few-shot prompting — 200 shots for CGA and 100 for CMV — using task-specific instructions without CoT. For Claude 3.5 Haiku, zero-shot general instructions yielded the best results. Evaluation details for open-source Llama 3.1 variants (8B, 70B, 405B) are provided in Appendix C; in brief, Llama-3.1-70B is competitive but trails the strongest proprietary models while offering lower cost.

The main trend we notice is that Claude does better with fewer shots in general than GPT. In a recent study (Shamshiri et al., 2024), zero-shot analysis indicated that GPT-4o excels in simple, short SA tasks across various datasets, while Claude 3.5

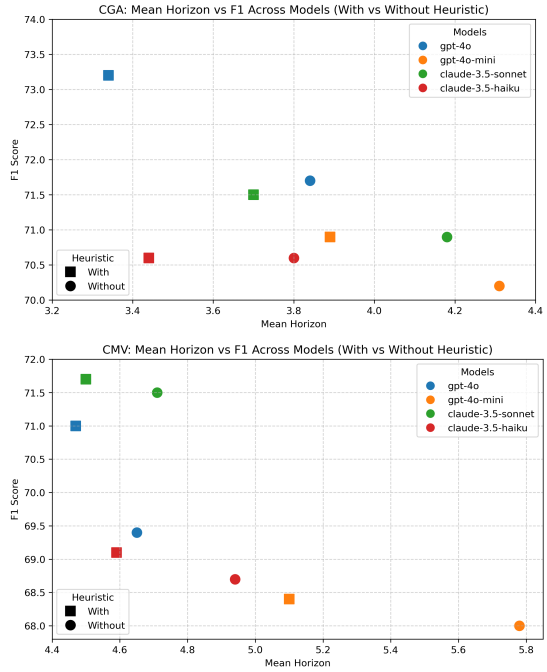

Figure 2: (Top) Mean Forecast Horizon vs. F1 for CGA. (Bottom) Mean Forecast Horizon vs. F1 for CMV.

Sonnet outperforms it in more complex, sentiment-wavering task. This may explain the higher performance of Claude in the CMV dataset, where exchanges tend to be longer and more complex. In the few-shot results in Shamshiri et al. (2024), both models exhibit similar trends, with Claude 3.5 Sonnet achieving superior results on most datasets, but GPT-4o demonstrates greater improvement with more shots. We observe the same pattern here. There is ample evidence in the literature showing differences between the two datasets, in particular, longer sequences and more turns in CMV, which would tend to favor Claude 3.5 Sonnet.

4.1 Horizon Forecasting

For horizon forecasting we include the dynamically trained versions of baseline models. Dynamic training entails mapping a conversation into separate training samples, each representing a different part of the ongoing conversation, but all labeled for whether or not the conversation eventually derails (Kementchedjhieva and Søgaard, 2021). Because the initial utterance in a conversation is almost never a derailment, we also experiment with the heuristic of withholding prediction at the first utterance. We report horizon results with and without this heuristic (Table 2), and general prediction metrics with the heuristic (Table 1).

4.2 Analysis

Claude 3.5 Sonnet shows the highest score with a general prompt on the CMV dataset and GPT-4o shows the highest score on the CGA dataset. All four LLMs consistently beat the baselines on the horizon metric for both datasets, indicating that they are better at predicting a derailment earlier in the conversation, with GPT-4o-mini showing the longest mean forecast horizon. The effect of withholding prediction at the first utterance is generally greater with the larger models, although the effects vary by dataset. We note an increase in F1 when withholding prediction at the first utterance, and an expected decrease in the horizon while doing so. Figure 2 shows the results on each dataset with and without this heuristic. We also perform comparative analysis of forecast horizon distributions of different models in Appendix A. As for why GPT outperforms in general on horizon forecasting, this may be due to lexical and practical differences between the two datasets. Since the Reddit CMV does not contain the actual conversational turn with the toxic comment, it is possible that Claude relies more heavily on the lexicon of toxicity and hate speech than GPT-4o and therefore performs worse when such clues are absent early. In CGA, the conversations are shorter and this may favor GPT-4o on horizon forecasting just as it does for performance on the task in general.

4.3 Cost–Performance Trade-off

To evaluate cost-efficiency (RQ3), we plotted inference cost against F1 performance for large and small variants of GPT-4o and Claude 3.5 on both CGA and CMV datasets (Figure 3). Larger models yield modest F1 improvements (1–3 points) at an order-of-magnitude higher cost; moreover, GPT-4o's reliance on extensive few-shot examples (200 examples for CGA, 100 for CMV) further inflates its cost, and Claude 3.5 Sonnet's 10-shot CoT prompting on the CMV dataset similarly increases its cost. In contrast, smaller models deliver near-competitive performance for a fraction of the price, making them ideal for cost-sensitive deployments, whereas full-size models remain preferable when maximal accuracy is required. Importantly, our approach achieves these trade-offs without any fine-tuning or additional features, demonstrating both robustness and practical deployability.

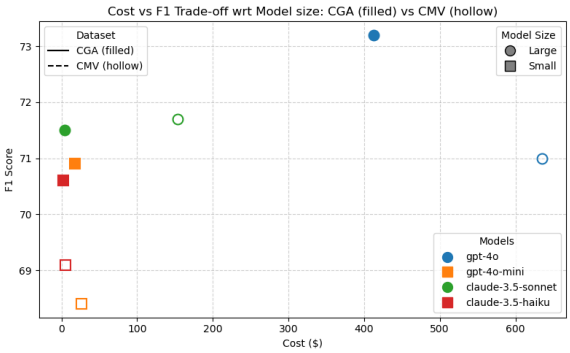

Figure 3: Cost-Performance Trade-off vs. Model Size

4.4 Qualitative Analysis

We also conduct a qualitative analysis of the LLM's predictions to better understand their behavior. We find that models reliably succeed when conversations contain overt incivility or explicit policy-enforcement language, suggesting they have learned strong lexical cues associated with immediate conflict. However, a key failure mode is missing "slow-burn" derailments that begin politely, indicating an over-reliance on early-turn signals rather than cumulative discourse context. We also observed frequent false positives on passionate-but-civil debates, suggesting the models confuse intense tone with toxicity. We provide the example conversations where LLM fails in Appendix E.

5 Conclusion

We have shown that off-the-shelf LLMs, when guided by carefully designed prompts and simple heuristics, can forecast conversational derailment as effectively as specialized models, and more effectively than baselines, while delivering earlier warnings. Our comprehensive evaluation across CGA and CMV benchmarks shows that general-purpose prompts are robust, and that task-specific prompts offer marginal improvements in some settings, while a heuristic of withholding predictions on the first turn improves F1 by reducing false positives. Larger models yield higher F1 and forecast horizons (compared to baselines) but at increased cost, underscoring practical trade-offs in real-world deployments. These findings suggest that general-purpose LLMs, with no additional fine-tuning, can serve as a scalable, vendor-agnostic solution for proactive moderation. We restrict ourselves to zero- and few-shot prompting without fine-tuning, which could further boost performance at the cost of additional data and engineering effort.

Limitations

Our investigation into LLM prompting for derailment forecasting has several caveats: we only benchmarked OpenAI, Anthropic, and Llama (through AWS Bedrock) APIs, requiring third-party data transmission and raising privacy/compliance concerns, and did not evaluate other commercial or open-source models due to cost and time constraints; larger LLMs tend to perform better but incur higher inference costs and latency, so real-world systems may need lightweight heuristics or batching to manage expense and responsiveness; and finally, LLM outputs are sensitive to prompt design and our evaluation focused on forecasting metrics rather than downstream effects like moderator workload, fairness, or user experience, so practitioners should validate prompts, monitor live deployments, and conduct human-in-the-loop trials.

Ethical Considerations

Deploying a conversational derailment prediction model in a real-world environment is an important way to prevent negative interactions from escalating. As such, because models are not perfect, there is some risk in deploying even the highest-quality models, especially if this is done without any human-in-the-loop moderation. Our qualitative analysis shows that such risks include missing "slow-burn" derailments that unfold gradually, as well as mistakenly flagging passionate-but-civil debates as toxic. These errors highlight the dangers of both creating a false sense of security and unintentionally censoring legitimate discourse.

References

Enas Altarawneh, Ameeta Agrawal, Michael Jenkin, and Manos Papagelis. 2023. Conversation derailment forecasting with graph convolutional networks. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 160–169, Toronto, Canada. Association for Computational Linguistics.

Enas Altarawneh, Ameeta Agrawal, Michael Jenkin, and Manos Papagelis. 2024. Knowledge-aware conversation derailment forecasting using graph convolutional networks. *Preprint*, arXiv:2408.13440.

Anthropic. 2024a. Claude 3.5 haiku model overview. https://docs.anthropic.com/en/docs/about-claude/models/all-models.

Anthropic. 2024b. Introducing claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

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.

Yova Kementchedjhieva and Anders Søgaard. 2021. Dynamic forecasting of conversation derailment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenya Nonaka and Mitsuo Yoshida. 2025. Zero-shot prediction of conversational derailment with large language models. *IEEE Access*, 13:55081–55093.

OpenAI. 2024a. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com.

OpenAI. 2024b. GPT-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf.

Alireza Shamshiri, Kyeong Rok Ryu, and June Young Park. 2024. In-Context Learning for Long-Context Sentiment Analysis on Infrastructure Project Opinions. *arXiv e-prints*, arXiv:2410.11265.

Anthony Sicilia and Malihe Alikhani. 2024. Eliciting uncertainty in chain-of-thought to mitigate bias against forecasting harmful user behaviors. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 211–223, Miami, Florida, USA. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16. International World Wide Web Conferences Steering Committee.

Jiaqing Yuan and Munindar P. Singh. 2023. Conversation modeling to predict derailment. *Preprint*, arXiv:2303.11184.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

A Forecast Horizon Distribution

We further analyze the forecast horizon results for CGA. Figure 4 illustrates the distribution of forecast horizon of different models. We notice that LLMs have shifted the horizon towards higher values resulting in earlier predictions as confirmed by Table 2. Moreover, we notice LLMs exhibit the earliest predictions without the withhold first-utterance heuristic (LLM with ++), shifting density from $H < 4$ to $H \geq 4$. This shows that LLMs can effectively forecast derailments well in advance without any type of fine-tuning.

B Model Identifiers

We list the model identifiers for the LLMs used in this work in Table 3.

LLM	Model ID
GPT-4o	gpt-4o-2024-08-06
GPT-4o-mini	gpt-4o-mini-2024-07-18
Claude 3.5 Sonnet	claude3.5-sonnet-20241022
Claude 3.5 Haiku	claude3.5-haiku-20241022
Llama-3.1-405b	Llama-3.1-405b-chat
Llama-3.1-70b	Llama-3.1-70b-instruct
Llama-3.1-8b	Llama-3.1-8b-instruct

Table 3: LLM model identifiers.

C Experiments with Llama 3.1 LLMs

We conducted further experiments with the Llama 3.1 family of models (8B, 70B, and 405B) on both the CGA and CMV datasets. These models allow us to examine performance–cost trade-offs using open-source systems with transparent parameter counts. We present the detailed results in Table 4.

Llama-3.1-70B performs competitively, achieving accuracy and F1 scores close to GPT-4o-mini and Claude 3.5 Haiku, though it remains below GPT-4o and Claude 3.5 Sonnet. At the same time, it offers substantially lower inference cost, highlighting its potential as a practical alternative for derailment forecasting. The smaller Llama-3.1-8B model

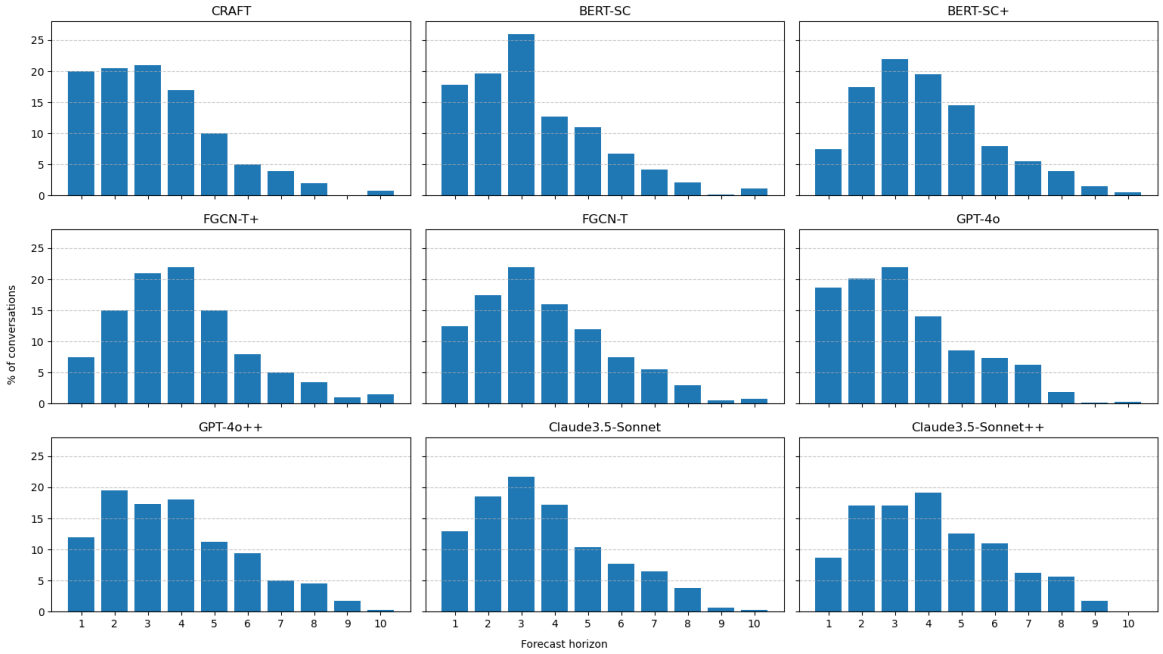

Figure 4: Comparing forecast horizon distribution of different models on the CGA dataset. +: dynamically trained models. ++: without the heuristic to withhold first-utterance predictions.

is the cheapest, demonstrates strong recall but lags in overall accuracy and F1, while the largest Llama-3.1-405B improves recall and F1 but at a considerably higher cost. Collectively, these findings illustrate the trade-offs between scale, accuracy, and efficiency across the Llama family and confirm that effective forecasting can be achieved without exclusive reliance on closed, high-cost models, although proprietary models still hold a modest performance edge.

D Prompt Templates

We list the prompt templates used for all our experiments here.

Prompt 1: General Instruction + Zero Shot

```
System message:

Your input fields are:
1. `conversation` (list[str]): Conversation so far
Your output fields are:
1. `derailment` (bool): Whether the conversation is or will be
↪  derailed
All interactions will be structured in the following way, with
↪  the appropriate values filled in.

[[ ## conversation ## ]]
{conversation}

[[ ## derailment ## ]]
{derailment}        # note: the value you produce must be True
↪  or False

[[ ## completed ## ]]
In adhering to this structure, your objective is:
        Derailment forecasting for a conversation.


User message:

[[ ## conversation ## ]]
{{conversation}}

Respond with the corresponding output fields, starting with the
↪  field `[[ ## derailment ## ]]` (must be formatted as a
↪  valid Python bool), and then ending with the marker for `[[
↪  ## completed ## ]]`.


Response:
```

Prompt 2: CGA-specific Instruction + Zero Shot

```
System message:

Your input fields are:
1. `conversation` (list[str]): List of comments in the
↪  conversation so far, in order
Your output fields are:
1. `derailment` (bool): True if the conversation is or will be
↪  derailed by a personal attack, False otherwise
All interactions will be structured in the following way, with
↪  the appropriate values filled in.

[[ ## conversation ## ]]
{conversation}

[[ ## derailment ## ]]
{derailment}        # note: the value you produce must be True
↪  or False

[[ ## completed ## ]]
In adhering to this structure, your objective is:
        Predict whether a Wikipedia Talk conversation will
        ↪  derail (contain a personal attack).


User message:

[[ ## conversation ## ]]
{{conversation}}

Respond with the corresponding output fields, starting with the
↪  field `[[ ## derailment ## ]]` (must be formatted as a
↪  valid Python bool), and then ending with the marker for `[[
↪  ## completed ## ]]`.
```

Model	CGA						CMV					
	Acc	P	R	F1	H	Cost($)	Acc	P	R	F1	H	Cost($)
Llama-3.1-405b	61.3	57.2	90.0	69.9	3.77	6.00	57.5	54.2	96.2	69.4	4.91	19.00
Llama-3.1-70b	**63.5**	**59.2**	86.7	**70.3**	**3.66**	1.02	**61.8**	**57.6**	89.8	**70.2**	**4.55**	2.58
Llama-3.1-8b	53.8	52.1	**96.7**	67.7	4.03	**0.13**	51.0	50.5	**99.4**	67.0	5.40	**0.30**

Table 4: Performance comparison: Accuracy (Acc), Precision (P), Recall (R), F1 Score.

```
Response:
```

Prompt 3: CMV-specific Instruction + Zero Shot

```
System message:

Your input fields are:
1. `conversation` (list[str]): Ordered list of comments from a
↪  ChangeMyView thread
Your output fields are:
1. `derailment` (bool): True if the conversation is likely to
↪  derail via a moderator-deleted comment (e.g., due to
↪  rudeness/hostility), False otherwise
All interactions will be structured in the following way, with
↪  the appropriate values filled in.

[[ ## conversation ## ]]
{conversation}

[[ ## derailment ## ]]
{derailment}        # note: the value you produce must be True
↪  or False

[[ ## completed ## ]]
In adhering to this structure, your objective is:
        Predict whether a ChangeMyView Reddit conversation will
        ↪  derail due to a rule violation (e.g., rudeness or
        ↪  hostility). Generally under Reddit's Rule: Don't
        ↪  be rude or hostile to other users.


User message:

[[ ## conversation ## ]]
{{conversation}}

Respond with the corresponding output fields, starting with the
↪  field `[[ ## derailment ## ]]` (must be formatted as a
↪  valid Python bool), and then ending with the marker for `[[
↪  ## completed ## ]]`.


Response:
```

Prompt 4: General Instruction + Zero Shot + Chain-of-Thought

```
System message:

Your input fields are:
1. `conversation` (list[str]): Conversation so far
Your output fields are:
1. `reason` (str): Reason for the prediction
2. `confidence` (float): Confidence score for the prediction
3. `derailment` (bool): Whether the conversation is or will be
↪  derailed
All interactions will be structured in the following way, with
↪  the appropriate values filled in.

[[ ## conversation ## ]]
{conversation}

[[ ## reason ## ]]
{reason}

[[ ## confidence ## ]]
{confidence}        # note: the value you produce must be a
↪  single float value

[[ ## derailment ## ]]
{derailment}        # note: the value you produce must be True
↪  or False

[[ ## completed ## ]]
In adhering to this structure, your objective is:
        Derailment forecasting for a conversation.


User message:

[[ ## conversation ## ]]
{{conversation}}

Respond with the corresponding output fields, starting with the
↪  field `[[ ## reason ## ]]`, then `[[ ## confidence ## ]]`
↪  (must be formatted as a valid Python float), then `[[ ##
↪  derailment ## ]]` (must be formatted as a valid Python
↪  bool), and then ending with the marker for `[[ ## completed
↪  ## ]]`.


Response:
```

E Example Conversations

We provide the examples of conversations from our qualitative analysis where the LLM misclassifies dialogue here. In the first two examples, LLM predicts derailment incorrectly and in the third example, LLM misses the derailment altogether. The provided examples are selected from the CGA dataset and the predictions are from GPT-4o.

Restricting Civil Disagreement

```
Turn 1: Please do not continue to remove article cleanup tags
↪  on wiki_link. The notability of the subject is in question,
↪  in part because of the lack of secondary sources
↪  demonstrating notability (only 1 source is cited and it's
↪  an obscure offline source in German), and the tags must
↪  remain in place until the issue is resolved. If you
↪  continue to wiki_link, you may be blocked. Thanks in
↪  advance.

Turn 2: Do not add superfluous tags to the articles I have
↪  written, as all subjects are notable. These tags will be
↪  removed.

Turn 3: The tag is not superfluous and removing it is a breach
↪  of protocol. You are free to discuss the matter on the
↪  article Talk page and present any sources that you may know
↪  of that attest to the subject's notability as per wiki_link
↪  and wiki_link. If you continue to edit war, you may be
↪  blocked. Thanks in advance for your cooperation.

Turn 4: Please wiki_link other editors. If you continue, you
↪  may be wiki_link from editing Wikipedia.
```

Censoring Passionate-But-Civil Debate

Turn 1: OK, I agree with Netoholic on this one. Unless the wayback machine or something else can verify the screenshots, they are pretty worthless. I more object to not discussing this and having revert wars over it. I also object to people blanking the page. Truly, unless someone can give us more reliable data then this stuff shouldn't really be used. –

Turn 2: But we can verify the screenshots exist, we just can't verify they're really of what they claim to be. Analogy time: we can't verify Jesus Christ is the son of God, so should we not mention that in the article? Of course not, we mention who thinks he's the son of God. Same here, we mention that there are some people who feel exit poll data was manipulated, without taking a stance whether it was or not.

Turn 3: Shane, bad example. There is far more evidence that Jesus really existed than there is those screenshots were doctored or altered in some way! –

Turn 4: Good example: can you verify Jesus was the son of God? check netaholic's history, he is allergic to debate.

Turn 5: I'm not saying Jesus didn't exist. I'm saying is Jesus the son of God? You can't prove it, it doesn't mean you shouldn't mention some people believe he is. Likewise, I don't think we can prove those screenshots are real. It doesn't mean we shouldn't report on them if we can find people who believe they are real. See Time Cube for a more extreme and probably less controversial example. Nearly everyone believes it's a complete load of crap crank theory. Doesn't mean we don't report on it. We have a duty to report what people believe. We do '''not''' have a duty to determine whether what they believe is true, that '''would''' be original research!

Turn 6: Look, you're talking about belief in a religious figure. Yes, that can't be proven. This is a different argument however: this is not the same as saying that facts taken from a screen capture of the CNN website, which is highly unreliable, should be included in this article. Either these screenshots or facts, or they aren't. It has nothing to do with my beliefs. Can we verify them or not? –

Missing Slow-Burn Conversation That Starts Politely But Derails

Turn 1: The above author is used as a source for some pretty contentious statements in this article. I've been reading the preview version of the book cited on and off over the last 24 hours and I've got quite a few issues with it. Principally, though:
there are loads of misprints, including sentences that are not completed & could therefore cause confusion/ambiguity etc (I do not include typos/grammar in this description as it is "Indian English")
aside from a lot of bookshops selling the thing, and the usual publicity blurbs etc, I can find nothing to indicate that the author is notable, that the work has been peer-reviewed etc. Anyone can get a book published.
Are we sure about the reliability of this as a source? Is the author an academic or what? Is the publisher known for producing books of a worthy standard? For those unfamiliar with the issue of reliable sources etc, please read WP:RS. Thanks. –

Turn 2: Assuming it's the same guy, according to University of Mumbai Department of History webpage, "Dr. G. M. Moraes was the first Professor and Head of the Department. He had the distinction of being the General President of the Indian History Congress in a premier body of historians in the country." That does seem to indicate that he is a serious academic; however, it doesn't necessarily make everything he ever wrote a reliable source. Have you been able to figure out who printed the book? Self-publishing was far less common in 1964 than today, so it's likely that ''someone'' made a decision to print the book, but it would help to know if it was an academic publisher or some other publisher.

Turn 3: I'm happy with Moraes as a RS, even if there may be other RSs that contradict him. The author I was referring to is the one named in the section heading to this – Vidya Prakesh Tyagi. –

Turn 4: Ah, my apologies. Here is a comment from a new editor on the issue:
I moved this up from the bottom, because it appears to be about the same thing

Turn 5: :The book you want removed appears to be published by a reliable publishing company (Gyan Publishing House), although I'm not totally sure. Do you have some explanation for why it's not a reliable source? If both it and the other you mention are reliable, then policy says we should include both theories and properly attribute them. I will say, though, that your suggested other book is over 100 years old, meaning it may not have up to date information. Again, I'm not saying certainly either way, only that it seems difficult to determine whether one or both meet the reliable sources guidelines.
I've removed my response...I'm not so sure the source is reliable...do we know anything more about Gyan Publishing? I'd already decided that some of the comments in the article regarding status were misquoting the cited sources – "bigging things up", if you like. But Tyagi is used for other assertions also. Given the hotchpotch nature of the book, with its numerous printing errors, I don't think that the ''edition'' can be relied on at all. I have doubts about the author, including for the reasons in the refactored comment above + main at the top of this section, but the ''edition'' is definitely useless in my opinion. When there are printing mistakes, including a cut-off sentence on one of the pages being cited here, who can possibly be sure that important phrases/sentences etc have not been omitted which would completely turn the picture round? –

Turn 6: also these proofs say dat the identity of pandyans is unknown. also refer the wiki pandyan article to get more proofs.the wiki pandyan article look fine. Preceding unsigned comment added by

Turn 7: Please sign your contributions – type 4 tildes (~) at the end of your post. Also, you cannot use another wikipedia article to "prove" that this article (or any other) is wrong, although I agree that there is a contradiction. I'd already seen these issues that you raise but thanks for bringing them to a wider audience. I'm working on things now and, being of European ethnicity, hope to bring some neutrality to this situation. –