Eliciting Uncertainty in Chain-of-Thought to Mitigate Bias against Forecasting Harmful User Behaviors

Anthony Sicilia **Malihe Alikhani**
Khoury College of Computer Sciences
Northeastern University
sicilia.a@northeastern.edu

Abstract

Conversation forecasting tasks a model with predicting the outcome of an unfolding conversation. For instance, it can be applied in social media moderation to predict harmful user behaviors before they occur, allowing for preventative interventions. While large language models (LLMs) have recently been proposed as an effective tool for conversation forecasting, it's unclear what biases they may have, especially against forecasting the (potentially harmful) outcomes we request them to predict during moderation. This paper explores to what extent model uncertainty can be used as a tool to mitigate potential biases. Specifically, we ask three primary research questions: 1) how does LLM forecasting accuracy change when we ask models to represent their uncertainty; 2) how does LLM bias change when we ask models to represent their uncertainty; 3) how can we use uncertainty representations to reduce or completely mitigate biases without many training data points. We address these questions for 5 open-source language models tested on 2 datasets designed to evaluate conversation forecasting for social media moderation.

1 Introduction

Conversation forecasting – where a model predicts the outcome of a partial conversation – is useful across many domains, e.g., see research on negotiation dynamics (Sokolova et al., 2008), mental health monitoring (Cao et al., 2019a), and social media moderation (Zhang et al., 2018). For instance, in online moderation, the forecasting task may be to predict whether a harmful behavior (like digital bullying) will eventually occur in an unfolding conversation, allowing moderators to intervene to prevent these behaviors. Recently, Sicilia et al. (2024) demonstrate pre-trained language models are relatively effective conversation forecasters, setting themselves apart because they do not require copious amounts of domain-specific training data prior to inference time. Yet, it remains unclear what biases these systems may hold, especially in digital media contexts, where they are specifically asked to predict outcomes that may be harmful to the parties involved (see Figure 1).

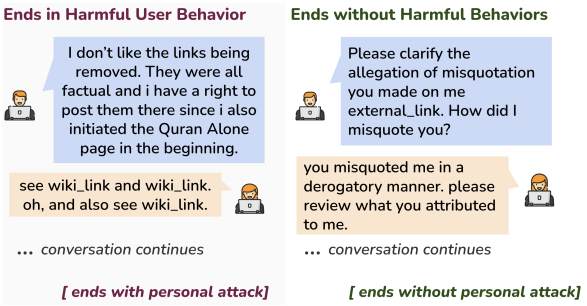

Figure 1: Two difficult social media moderation examples. Both instances appear as if they may derail, leading to harmful user behaviors. Yet, only one does. These are real examples from the moderation corpora we study, identified using this online tool.

Indeed, the data used in common instruction-tuning algorithms – e.g., RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2024) – are designed to align language models with human values, and subsequently, avoid any propagation of harm. Meanwhile, the motivating tasks of this paper draw a fine line between propagation and prediction. Surely, "predicting" a harmful outcome is not "speaking into existence" but it's unclear whether this distinction is lost on "aligned" language models. Or, if it is not lost, whether underlying data bias (i.e., against harmful outcomes) predisposes language models to propagate this bias when forecasting harmful outcomes.

While the role of alignment mechanisms in producing model bias is difficult to confirm,[1] our own empirical results, and those of previous work (Sicilia et al., 2024), indicate current language models are indeed biased against predicting harmful out-

[1]For instance, pre-training data could also play a role.

Proceedings of the Third Workshop on NLP for Positive Impact, pages 211–223

comes. Aptly, this paper is interested in mitigating these biases, and we approach this task using uncertainty estimation.

By its nature, conversation forecasting is a highly uncertain task. For instance, two seemingly similar conversations can end with opposite outcomes (e.g., a personal attack vs. an amicable resolution, as in Figure 1). While modeling this uncertainty has independent motivations besides the study of bias (Sicilia et al., 2024), we are specifically interested in how considering uncertainty effects the "reasoning" process of language models.[2] Indeed, neuroscience (both cognitive and computational) recognizes the role uncertainty plays in human decision-making, wherein the brain is understood to both predict and process different forms of uncertainty (Bland and Schaefer, 2012). We hypothesize language models may benefit from utilizing similar patterns of reasoning, having learned these (statistical) patterns from the human-generated text on which they are trained. In particular, we hypothesize elicitation of uncertainty can mitigate bias in model predictions.

In studying this broader hypothesis, we focus on three central research questions:

1. how does the forecasting accuracy of a language model change when it is prompted to reflect uncertainty in it's prediction;
2. how does the bias of a language model's forecasts (i.e., against harmful outcomes) change when it is prompted to reflect uncertainty;
3. and, how can we use a language model's predicted uncertainty to mitigate any such biases.

We address these questions for 5 open-source language models tested on two datasets from the conversation forecasting corpora proposed by Zhang et al. (2018), specifically tailored towards harmful behaviors (i.e., personal attacks) in social media.

2 Background

2.1 Conversation Forecasting Setup

We work within the conversation forecasting framework established by Sicilia et al. (2024), wherein the model is tasked with predicting a conversation's outcome. For instance, it may need to predict whether a personal attack will occur (or not). Since the conversation provides only a limited glimpse into the underlying reality, unknown factors like future developments or unobservable mental states introduce an element of randomness, making it challenging to determine the outcome with certainty based solely on the available information.

Task For a set of natural language tokens $\mathcal{T}$, we assume observation of a partial multi-party dialogue $D \in \mathcal{T}^*$ consisting of K turns. Following Sicilia et al. (2024), the length K is a uniform random number between 2 and the full dialogue length, simulating the "partial" property of the dialogue.[3] These conversations appear unfinished to the model, but in reality, have an eventual ground-truth outcome $O \in \{0, 1\}$, indicating whether a personal attack occurs or does not occur. The task of the model is to predict O given D – that is, to predict whether a personal attack will occur given the partial conversation.

Metrics Sicilia et al. (2024) evaluate the quality of a model's uncertainty estimates when conversation forecasting (i.e., using a metric called the Brier score). We focus on different evaluation metrics, selected to properly answer our distinct research questions. Given a model prediction $\hat{O}$ for O, we evaluate the model using the **accuracy** of the prediction: $\mathbf{E}[\hat{O} = O]$. Besides accuracy, we also report the **F1 score** to capture both precision and recall. To measure the bias of the predictions, we report the **statistical bias**: $\mathbf{E}[\hat{O} - O]$, which is traditional measure of systematic error in an estimator. Specifically, this captures the average trend of the model's errors: whether it *over*-estimates (bias is positive) or *under*-estimates (bias is negative) on average. This type of bias is seemingly different from common quantitative notions of social bias in a model's outputs; e.g., see Gallegos et al. (2024). In reality, this (older) measure of bias is a special case of *accuracy parity* (Zhao and Gordon, 2022) where the group trait of interest, or "protected attribute," is the occurrence of a personal attack.

Corpora We consider two corpora in this work:

1. (`wiki`) a corpus of conversations from Wikipedia's *talk* page, proposed by Zhang et al. (2018), in which authors discuss edits to Wikipedia articles; and
2. (`reddit`) a corpus of conversations from the subreddit ChangeMyView, proposed by Chang et al. (2019), in which redditors try to

[2] We do not intend to imply that language models conduct any human-like forms of reasoning. Yet, changing prompts to elicit focus on uncertainty innately changes the tokens on which we condition language model outputs; this is the statistical process which we intend to study.

[3] Turns are marked by unique token sequences; e.g., "Speaker 4: ..."

convince each other to change their position on an (often contentious) issue.

Both corpora come with labels of whether a personal attack eventually occurs. The portion of each dataset we use in this paper contains 100 instances without a personal attack and 100 instances with a personal attack, following the (nearly) even distribution of positive/negative instances in the original data. The average number of tokens in each dataset are 387 and 624, respectively; this is checked *after* we prune turns to simulate partial conversations.

2.2 Other Related Work

Conversation Forecasting As noted, Zhang et al. (2018) and Chang et al. (2019) provide early investigations and data for forecasting personal attacks during dialogue to proactively moderate online forums. Using the same data, Kementchedjhieva and Søgaard (2021); Altarawneh et al. (2023) propose new models, capitalizing on temporal and social aspects of dialogue. Meanwhile, forecasting of other conversation outcomes includes task-success (Walker et al., 2000; Reitter and Moore, 2007), mental health codes (Cao et al., 2019b), emotions (Wang et al., 2020; Matero and Schwartz, 2020), situated actions (Lei et al., 2020), and financial events (Koval et al., 2023). Among these, our work is uniquely positioned by its focus on the relationship between uncertainty and bias when using modern language models for this task. Broadly, studying how language models perform at this task is an important research direction because they promise a pipeline that requires very limited labeled data relative to other, previous directions of study. At the same time, these pre-trained models may have unknown biases, calling for the direction of study proposed in the current paper.

Uncertainty Estimation with LMs Modern "aligned" language models have been shown to be capable at representing uncertainty in their responses to factual queries, even with minimal supervision (Kadavath et al., 2022). Meanwhile, uncertainty has also been well studied in models without alignment to human preferences (Desai and Durrett, 2020; Jiang et al., 2021; Dan and Roth, 2021; Kong et al., 2020; Zhang et al., 2021; Li et al., 2022) Unlike existing work, ours is interested in how fine-tuning for alignment to human preferences might bias the model against predicting adverse outcomes. As far as how we extract uncertainty estimates from the language model, our work is most in line with that of Lin et al. (2022); Mielke et al. (2022); Tian et al. (2023) who all suggest "direct forecasts" or uncertainty estimates directly specified in the sampled tokens of the model. These estimates are considered best out-of-the-box for the types of models we study (Sicilia et al., 2024).

3 Methods

3.1 Forecasting with Language Models

Here, we describe prompts used to elicit conversation forecasts. A full example is in the Appendix.

Traditional CoT Classification To predict conversation outcomes with language models, we simply provide the language model with the partial conversation segment and prompt the language model to predict the outcome. There are some key components to precisely detail our strategy.

1. **Role Play**: As part of the system prompt, we give the language model a "name" and "skill set" to direct the language model to mimic a task expert. This is a common prompt engineering technique. We use a similar role description as (Sicilia et al., 2024), emphasizing skills like Theory of Mind and the ability to predict actions/thoughts of different interlocutors.
2. **Output Format**: To conclude the system prompt, we direct the model to use an easy-to-parse format; e.g., `ANSWER = 1` for $O = 1$.
3. **Context**: To start the user prompt, we explain the context of the conversation; e.g., "The speakers are discussing edits to a Wikipedia article." We then provide context for predicting this specific instance. These include the partial conversation segment (delimited using special token sequences) and the question of interest. Specifically, we ask "Will a personal attack occur at the end of the conversation?".
4. **Chain of Thought**: We conclude the user prompt with a chain-of-thought trigger phrase. Specifically, we use "Let's think step by step, but keep your answer concise (less than 100 words)." This encourages the model to output reasoning for it's answer and has been shown to improve performance (Kojima et al., 2022).

Uncertainty-Aware CoT Classification We use largely the same prompting strategy as traditional classification. Instead of asking for an answer directly, we instruct the model to report it's answer on a 10 point Likert scale where 1 indicates "not likely at all" and 10 indicates "almost certainly."

After parsing the answer (with the same regular expression), we set $O = 1$ if the score is greater than 5. We set $O = 0$ otherwise. This allows the model to explicitly consider "uncertainty" in it's answer as well as the "reasoning" process triggered by the chain-of-thought prompting technique.

Post-hoc Intervention for Bias Mitigation Besides our initial hypothesis – that considering "uncertainty" in the inference step may improve chain of thought reasoning and subsequent performance – outputting certainty in the answer allows us to tune the model's answer to our data source. Rather than data- and compute-expensive fine-tuning of model weights, we suggest **post-hoc forecast scaling**, which is a variant of Platt Scaling, proposed to improve the forecasts of language models by Sicilia et al. (2024). If $\hat{P}$ is the parsed and normalized Likert score (i.e., divided by 10), which signals model uncertainty, we use parameters τ and β to scale:

$$\begin{aligned} \hat{Z} &\leftarrow \log \hat{P}/(1-\hat{P}) \\ \tilde{Z} &\leftarrow \hat{Z}/\tau - \beta \\ \hat{P}_{\text{new}} &\leftarrow 1/(1+\exp(-\tilde{Z})). \end{aligned} \tag{1}$$

$\hat{P}_{\text{new}}$ is then used as the new (normalized) Likert score for confidence; i.e., if $10 \times \hat{P}_{\text{new}} > 5$ we set $O = 1$. Parameters are learned by MLE (n=50), treating $\hat{P}_{\text{new}}$ as likelihood for the ground-truth outcome. While this method is known to improve uncertainty estimates, it's not yet been studied in the current paper's context; i.e., exploring its impact on forecasting accuracy or model bias.

Models We test these prompting and scaling techniques on Llama 3.1 8B and 70B (AI@Meta, 2024), Mistral 7B v0.3 and Mixtral 8x22B (Jiang et al., 2023, 2024), and Qwen2 72B (Yang et al., 2024). All models are instruction-tuned variants. We use the default sampling parameter settings for Llama as provided in the official Llama GitHub repository (temp = 0.6, top p = 0.9). For all other models, we use temp = 0.7 and top p = 1. We access models via the together AI API.

3.2 Semi-Automated Topic Analysis

Method One aspect we explore empirically is the relationship between a model's forecasting bias and the topic of the conversation. This can give us a more fine-grained view of how a model is biased in the context of social media moderation. We use a semi-automated pipeline to predict topics using a large language model. Specifically, we use Meta's Llama 3.1 405B. Our strategy is as follows:

1. Prompt the language model to provide a noun phrase describing the topic of each instance.
2. Prompt (the same model) to collect the list of sub-topics into higher-level categories.
3. Iterate step two if the model misses any sub-topics. This process is accelerated with a programmatic check on the model outputs. We re-prompted (in the same conversation context) to tell the model which noun phrases were left out of the current category list.
4. Manually inspect the final model-generated categories. To improve the categories, we reorganize, combine, and remove small categories (less than 10 instances).
5. Ask the model to analyze it's own (author adjusted) categories and provide descriptions.

Topics This process only worked well for the `reddit` corpus (as manually evaluated by the authors based on diversity and correctness). It produced the following categories (and descriptions):

- **Social Issues**: "This category encompasses a wide range of topics related to social justice, equality, and human rights. It includes discussions on discrimination, feminism, LGBTQ+ rights, racism, and other forms of social inequality. Sub-topics also explore issues related to family and relationships, such as marriage, child abuse, and parental leave."
- **Politics and Law**: "This category delves into the realm of governance, policy-making, and the legal system. It covers topics such as gun control, immigration, free speech, and electoral politics, as well as issues related to national security, terrorism, and international relations. Sub-topics also examine the role of government, the judicial system, and the relationship between citizens and the state."
- **Economics**: "This category focuses on the production, distribution, and exchange of goods and services. It includes discussions on trade deficits, minimum wage, labor unions, and regulation, as well as emerging topics like cryptocurrency and digital goods. Sub-topics also touch on social welfare and the economic aspects of family relationships, such as alimony and child support."
- **Health**: "This category explores topics related to physical and mental well-being, including vaccination, mental health, and substance use. It also covers issues related to healthcare policy, medical

ethics, and the intersection of health and society, such as prostitution and sexting laws. Sub-topics also examine lifestyle choices, such as veganism and vegetarianism."

- **Culture and ID**: "This category examines the complex and multifaceted nature of identity, culture, and society. It includes discussions on cultural identity, feminist terminology, indigenous rights, and the Israeli-Palestinian conflict, among others. Sub-topics also explore the intersection of culture and politics, including the role of historical figures, social movements, and cultural protests.
- **Tech and Ent**: "This category delves into the world of technology, entertainment, and media. It covers topics such as ad blocking, game streaming, journalism, and social media, as well as issues related to censorship, art, and sports. Sub-topics also examine the impact of technology on society, including privacy concerns and the ethics of online behavior."
- **Ethics and Morality**: "This category grapples with fundamental questions about right and wrong, morality, and ethics. It includes discussions on free will, animal rights, organ donation, and evidence-based reasoning, among others. Sub-topics also explore the nuances of human behavior, including discipline, gift giving, and historical judgment."

Descriptions were judged to be accurate by the authors. The full list of sub-topics and super-topics are in the Appendix, along with key prompts.

4 Experiments

In general, we use Hoeffding's Inequality to test statistical significance at level $\alpha = 0.05$. It provides a versatile (albeit, conservative) confidence interval with limited assumptions, making it applicable to accuracy (**ACC**) *and* statistical bias (**SB**).

4.1 Uncertainty and Forecasting Performance

RQ1: How does uncertainty-aware inference impact the forecasting performance of language models?
A: Some language models, especially those that perform poorly initially, benefit from considering uncertainty.

Forecasting Accuracy Results Table 1 shows forecast accuracy across models and datasets with and without the uncertainty-aware prompt strategy. For 3 out of 5 models, the uncertainty-aware strategy leads to improved performance on average. Average increases in accuracy range from 1% up to 5.25%, which on our dataset corresponds to about 3 to 13 more correct predictions, respectively. The Llama 3.1 series (8B and 70B) are the only models which do no benefit from the uncertainty-aware strategy. For the 8B model, performance is unchanged (averaged across datasets). For the 70B model, performance is reduced by nearly 4%. For both datasets, the uncertainty-aware strategy lead to improved performance (on average). Average increases are near 1% for the Wikipedia corpus and the Reddit corpus. The only statistically significant improvement in performance comes when we apply the uncertainty-aware strategy to Mixtral.

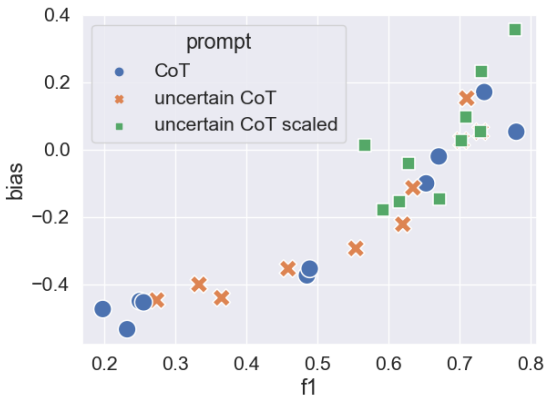

Figure 2: F1 v. Bias for all models / datasets with different inferences strategies. CoT refers to our standard conversation forecasting prompt (i.e., which uses CoT), while uncertain CoT ask the model to represent it's uncertainty in place of direct classification. Scaling refers to post-hoc scaling and is only applicable to the former strategy. It is best to have near 0 bias and high F1 score.

Forecasting F1 Results Table 2 shows F1 scores for forecasts across models and corpora. When considering precision and recall of inferences (F1 is their harmonic mean), we find results are largely consistent with those reported for accuracy. Three of five models show improvement, meanwhile both datasets show improvement. Relative performance of models is also consistent: Qwen2 does worst, is improved by the Mistral models, and further improved by the Llama 3.1 series.

Discussion Findings indicate that considering uncertainty in the LM forecast either has little impact (on average) or a slight positive one, for certain models. One observation is that the best performing models (the Llama 3 series) are either unaf-

	Llama 3.1 8B		Llama 3.1 70B		Mistral v0.3 7B		Mixtral 8x22B		Qwen 72B		mean **ACC**	
uncertainty	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
`wiki`	67.5	68	64	62	51.5	54	53	58	53.5	54.5	57.9	**59.3**
`reddit`	58	57.5	66.5	61.5	52	51.5	54	59.5	43.5	48.5	54.8	**55.7**
mean **ACC**	62.75	62.75	65.25	61.75	51.75	**52.75**	53.5	**58.75***	48.5	**51.5**	56.35	**57.5**

Table 1: Accuracy of different models at forecasting personal attacks with (✓) and without (✗) uncertainty-aware prompting strategy. Accuracy is reported on a 100pt scale. **Bold** shows improvement from incorporating uncertainty for model/data averages. An asterisk is used to denote statistically significant results (among the averages).

	Llama 3.1 8B		Llama 3.1 70B		Mistral v0.3 7B		Mixtral 8x22B		Qwen 72B		mean **F1**	
uncertainty	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
`wiki`	0.692	0.698	0.621	0.6	0.185	0.258	0.266	0.4	0.243	0.305	0.401	**0.452**
`reddit`	0.702	0.699	0.747	0.712	0.461	0.497	0.494	0.61	0.199	0.383	0.521	**0.580**
mean **F1**	0.697	0.699	0.684	0.656	0.323	**0.378**	0.38	**0.505**	0.221	**0.344**	0.461	**0.516**

Table 2: F1 scores of different models at forecasting personal attacks with (✓) and without (✗) uncertainty-aware prompting strategy. F1 ranges from 0 to 1. **Bold** shows improvement from incorporating uncertainty.

fected by the change in prompt (in case of the 8B model) or negatively effected by the prompt (in case of the 70B model). Although, the negative result is not statistically significant. We hypothesize a saturation effect may occur for these high performing models, where there is little additional predictive power to be gained through simple means like prompt engineering. Comparing these results to related literature suggests this may be the case. Indeed, in a similar experimental setup (albeit, slightly easier) an average accuracy near 64% is achieved by a specialized model *which is trained on the dataset* (Altarawneh et al., 2023), showing (potentially) that waning amounts of insight can be gained on this highly uncertain task once accuracy reaches a certain threshold. On the other hand, for models with a worse baseline accuracy, considering uncertainty in the prompt does seem to offer some benefit to the inference process. As we note previously, we hypothesize this is due the interaction between the chain-of-thought "reasoning" and the answer-format (which represents model uncertainty). Considering uncertainty may tap into patterns of "reasoning" learned from the training data that are overall beneficial.

4.2 Uncertainty and Forecasting Bias

RQ2: How does uncertainty-aware inference impact forecasting bias?
A: While some language models consistently under-predict the occurrence of personal attacks, considering uncertainty is able to partially reduce this bias.

Forecasting Bias Results Table 3 shows statistical bias of language model forecasts with and without consideration of uncertainty at inference-time. Uncertain inferences reduce bias for three out of five models. Again, the Llama 3.1 series are the only models that do not show any benefit. In this case, bias is relatively consistent with/without uncertainty (unlike the drop in accuracy for the 70B model observed in Table 1). Bias was often negative, *indicating that models typically under-predict the occurrence of a personal attack*; i.e., on average, they predict no personal attack when an attack does in fact occur. Only the Llama 3.1 series showed any sign of positive bias (specifically, on the Reddit corpus). Reductions in bias range from 0.05 up to 0.09. In our context, this means use of uncertainty corrected 5 out of 100 or 9 out of 100 false negatives, respectively. For some models (Mixtral and Qwen2), this reduction is statistically significant. Both datasets also experience reduction in bias on average, with 3 out of 100 and 7 out of 100 less false negatives for the Wikipedia corpus and Reddit corpus, respectively. These reductions were not statistically significant.

Interactions Between Topic and Bias Figure 3 shows the relationship between bias and different topics identified using the method from § 3 applied to the Reddit corpus. We limit consideration to the Mixtral models and Qwen2, since these models exhibited consistent negative bias (i.e., systematic under-prediction of personal attacks). For traditional inference without uncertainty (traditional

	Llama 3.1 8B		Llama 3.1 70B		Mistral v0.3 7B		Mixtral 8x22B		Qwen 72B		mean **SB**	
uncertainty	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
`wiki`	-0.03	0.01	-0.12	-0.12	-0.48	-0.45	-0.44	-0.37	-0.46	-0.42	-0.30	**-0.27**
`reddit`	0.21	0.20	0.10	0.11	-0.34	-0.27	-0.32	-0.20	-0.53	-0.40	-0.18	**-0.11**
mean **SB**	0.09	0.11	-0.01	-0.01	-0.41	**-0.36**	-0.38	**-0.29***	-0.49	**-0.41***	-0.24	**-0.19**

Table 3: Statistical bias of models forecasting personal attacks with (✓) and without (✗) uncertainty-aware prompting strategy. **SB** ranges between -1 and 1 with closer to 0 being best. **Bold** shows improvement from incorporating uncertainty. An asterisk is used to denote statistically significant results (among the averages).

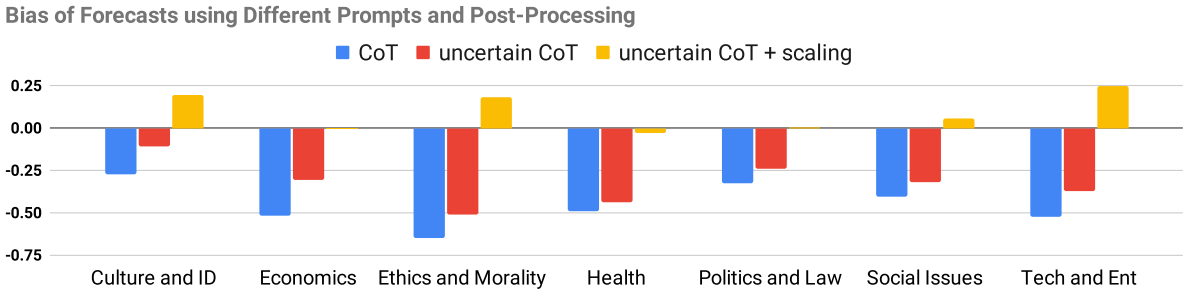

Figure 3: Statistical Bias of Forecasts on Reddit for Mistral models and Qwen2. Language models either use uncertainty estimates to report inferences (uncertain CoT) or make traditional binary decsions (CoT). Impact of post-hoc scaling is also shown for the former of these methods. Topics are determined using the method from § 3.

CoT), bias is most prominent on Reddit conversations about "Ethics and Morality" followed by conversation about "Economics" or "Tech and Entertainment." When uncertainty is considered during inference (uncertain CoT), bias is reduced for all topics. One of the biggest reductions occurs for the "Economics" topic. For both forecasting methods, the topics with the lowest overall bias are "Culture and Identity" and "Politics and Law."

Discussion Findings indicate that most language models exhibit negative statistical bias (systematic under-prediction) when forecasting personal attacks. This lends evidence to our over-arching hypothesis – that AI alignment mechanisms can bias language models against predicting harmful outcomes – since under-prediction of a personal attack is indeed a harmful outcome. Of course, it is difficult to confirm this idea without transparent access to training data and methods (for alignment) as well instruction-tuned models, which are guaranteed to be "un-aligned" along the dimensions of interest. In any case, findings also indicate that uncertainty-aware inference with language models is able to reduce negative bias. As before, the impact of uncertainty-aware inference is not consistent across models: the more biased models experience the greatest degrees of bias reduction. For two models, this reduction was even statistically significant. We hypothesize the disparity across models again may be due to a saturation effect, as models which are not consistently biased do not have consistent patterns of "reasoning" that can be modified by consideration of uncertainty. We also observe that bias is not uniform across topics, nor is bias reduction (by uncertain CoT). We do not find any consistent properties among topics, which cause more/less bias. Yet, if our overarching hypothesis is correct – that AI alignment is a cause of bias – then this non-uniformity may be related to the types/amounts of data used during alignment.

4.3 More Benefits of Uncertainty: Scaling

RQ3: Can post-hoc scaling of uncertainty estimates further mitigate bias without impacting accuracy?
A: Yes. Scaling consistently produces the least biased and most accurate forecasts.

Forecasting Accuracy Results Table 4 shows F1 scores for language model forecasts with and without post-hoc scaling of uncertainty estimates. Note, this implies we use the uncertain CoT strategy, since scaling is not possible with traditional CoT. Scaling improves F1 scores by almost 20 pts (out of 100) for Mistral models and more than 30 pts for Qwen2. The Llama 3.1 series remain as the "odd-models-out" with their high performance

	Llama 3.1 8B		Llama 3.1 70B		Mistral v0.3 7B		Mixtral 8x22B		Qwen 72B		mean **F1**	
scaling	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
`wiki`	0.703	0.703	0.634	0.628	0.274	0.593	0.459	0.671	0.333	0.731	0.481	**0.665**
`reddit`	0.710	0.709	0.730	0.730	0.554	0.567	0.620	0.779	0.365	0.615	0.596	**0.680**
mean **F1**	0.707	0.706	0.682	0.679	0.414	**0.580**	0.539	**0.725**	0.349	**0.673**	0.538	**0.673**

Table 4: F1 scores of different models with (✓) and without (✗) post-hoc scaling; i.e., so all models are prompted to express uncertainty. Post-hoc scaling uses a 50 sample dev. set and results are reported on remaining (held out) data. **Bold** shows improvement from incorporating uncertainty for model/data averages.

	Llama 3.1 8B		Llama 3.1 70B		Mistral v0.3 7B		Mixtral 8x22B		Qwen 72B		mean **SB**	
scaling	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
`wiki`	0.03	0.03	-0.11	-0.04	-0.45	-0.18	-0.35	-0.15	-0.40	0.23	-0.26	**-0.02***
`reddit`	0.15	0.10	0.05	0.05	-0.29	0.01	-0.22	0.36	-0.44	-0.15	-0.15	**0.07**
mean **SB**	0.09	**0.06**	-0.03	**0.01**	-0.37	**-0.08***	-0.29	**0.11***	-0.42	**0.04***	-0.20	**0.03**

Table 5: Statistical bias of different models with (✓) and without (✗) post-hoc scaling; i.e., so all models are prompted to express uncertainty. Post-hoc scaling uses a 50 sample dev. set and results are reported on remaining (held out) data. **Bold** shows improvement from incorporating uncertainty for model/data averages.

being maintained after the application of scaling. All datasets also show substantial improvements in F1 score after application of scaling.

Forecasting Bias Results Table 5 shows statistical bias with and without post-hoc scaling. Scaling is able to reduce the magnitude of bias for all models, including three (out of five) statistically significant reductions (i.e., all models except the Llama 3.1 series). Average reduction in bias across datasets is also consistent with statistically significant reduction on the Wikipedia corpus. From Figure 3, we more easily see that scaling tends to lead to slight positive bias (less in magnitude then the original negative bias).

Interaction Between Forecasting Bias and Accuracy Figure 2 shows bias and F1 score simultaneously via a scatter plot, for all models/data, organized by prompt strategy and use of scaling. Reductions in bias generally correlate with improved accuracy (an apparent quadratic relationship). Use of all proposed methods (uncertainty-aware CoT with scaling) creates a unique cluster of data points with near 0 bias and high F1 score.

Discussion Findings show that using a small amount of data for post-hoc scaling consistently improves both F1 score and bias by a relatively large magnitude. We remark, this is a benefit of using uncertainty estimates to make predictions, since post-hoc scaling is not possible for traditional CoT classification. One interesting point is that the Llama 3.1 series remains relatively unaffected by any of our modifications. Again, we believe this to be an effect of saturated (high) performance out-of-the-box. We can understand why scaling works from a mathematical perspective. In particular, the parameter β acts to remove systematic biases from the latent score $\hat{Z}$ in Eq. (1). If latent scores are typically higher than they should be (i.e., leading to higher forecast confidence, and thus, over-prediction), the MLE optimization uses β to lower these latent scores systematically across all predictions. We hypothesize the reason this correction sometimes leads to positive bias is from over-fitting to the small data sample used for MLE.

5 Conclusions

This paper studies three research questions about the interaction between uncertainty estimation and forecast bias for social media moderation using language models. Briefly, our findings show how asking language models to represent their uncertainty when forecasting personal attacks can reduce bias and increase accuracy, especially if a small amount of data is available to fine-tune these inferences.

One interesting point, which we are unable to address, is the root cause of the biases observed. We speculate this is a result of alignment mechanisms biasing language models against predicting the harmful outcomes we wish to forecast (i.e., personal attacks). Yet, more transparency in language model training is needed to investigate this issue.

Limitations

As noted in our conclusions, some key hypotheses of our work remain under-explored. Specifically, the cause of observed biases in the language models we study. Working with open-source language models that have closed-source training pipelines makes this a difficult research question to definitely handle. On the other hand, the research questions we *do* answer may also have limited interpretation outside of the contexts in which we study them; i.e., the specific models and datasets explored in § 4. A compounding issue of our analysis is the relatively small test sets we explore (200 instances, due to paper budget) which limited the statistical power of our study, as highlighted by the relatively few statistical significant results.

Ethics Statement

While the focus of this work is on analyzing (and mitigating) the bias of the language models we study, we emphasize that models which employ our proposed techniques still incur some bias. This can have direct, negative impact on users if these models are used for social media moderation in a automated pipeline without appropriate human checks. Even with human checks, if these models are used for decision-making, they may influence their human users in unknown ways, which can have unknown (and vast) negative impacts on online communities where they are deployed. Not to mention, we have only explored a very small subset of the potential biases these pre-trained models can possibly have. Other (social) biases may also exist in these models, which our methods are not explicitly designed to counteract and which can also have negative impacts on (vast) numbers of users if used for semi-automated decision-making. These caveats should be carefully considered and studied before systems like the language models we study are used for any automated moderation decisions.

One additional issue is the broader of role content moderation on the internet, and how decisions in content moderation can broadly impact online discourse. The question of who makes moderation decisions, how these decisions are made, and whether moderation should occur at all are each important issues of social debate, which we do not address in this paper. Tacitly, the datasets we study make some claim about what behaviors should be allowed (or not allowed) on online forums, as annotated by human moderators and crowd-workers. We emphasize these distinctions are for the purpose of research study alone, and the content of this data (used for learning and evaluation) should be carefully considered prior to it's use to make decisions or deploy models in real online communities.

Acknowledgements

This research was supported in part by Other Transaction award HR0011249XXX from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. Thanks to Jack Hessel for helpful discussion about early versions of this work.

References

AI@Meta. 2024. Llama 3 model card.

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.

Amy R Bland and Alexandre Schaefer. 2012. Different varieties of uncertainty in human decision-making. *Frontiers in neuroscience*, 6:85.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019a. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *Computational linguistics-Association for Computational Linguistics*.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019b. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and . 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.

Soham Dan and Dan Roth. 2021. On the effects of transformer size on in- and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online. Association for Computational Linguistics.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022. Calibration meets explanation: A simple and effective approach for model confidence estimates. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2784,

Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Matthew Matero and H Andrew Schwartz. 2020. Autoregressive affective language forecasting: a self-supervised task. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2020, page 2913. NIH Public Access.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

Anthony Sicilia, Hyunwoo Kim, Khyathi Raghavi Chandu, Malihe Alikhani, and Jack Hessel. 2024. Deal, or no deal (or who knows)? forecasting uncertainty in conversations using large language models. *arXiv preprint arXiv:2402.03284*.

Marina Sokolova, Vivi Nastase, and Stan Szpakowicz. 2008. The telling tail: Signals of success in electronic negotiation texts. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen L Gorin, and Diane Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: experiments with how may i help you? In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Zhongqing Wang, Xiujun Zhu, Yue Zhang, Shoushan Li, and Guodong Zhou. 2020. Sentiment forecasting in dialog. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2448–2458, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1958–1970, Online. Association for Computational Linguistics.

Han Zhao and Geoffrey J Gordon. 2022. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26.

A Appendix

A.1 Forecasting System Prompt Example

You are TheoryOfMindGPT, an expert language model at using your theory-of-mind capabilities to predict the beliefs and actions of others in human conversations. You will be given an unfinished conversation between two speakers. Put yourself in the mindset of the speakers and try to reason about the requested conversation outcome. Use the keyword "ANSWER" to report your prediction for the outcome of interest. Report your answer on a scale from 1 to 10 with 1 indicating "not likely at all" and 10 indicating "almost certainly". For example, "ANSWER = 7" would mean you think the outcome is fairly likely.

A.2 Forecasting User Prompt Example

In the following conversation segment, the speakers are negotiating how to allocate available resources among themselves.

[SEGMENT START]
Speaker 0: Hello how are you?
Speaker 1: Hello! I am doing well. How about you?
Speaker 0: I'm doing well. I'm trying to prepare for this camping trip.
Speaker 1: Me too.
Speaker 0: What are you looking for?...
[SEGMENT END]

Now, fast-forward to the end of the conversation. Will both speakers be satisfied at the end of the conversation? Let's think step by step, but keep your answer concise (less than 100 words).

A.3 Topic Model System Prompt

You are TopicClassifierGPT, an expert language model at assigning topics to conversations across the internet. Try to categorize the topic of the conversation using only one or two words, so that your categories can be automatically grouped and analyzed later. Topics should be nouns or noun phrases that provide an answer to the question: "What are the speakers discussing?" Use the keyword "ANSWER" to report your predicted category. For example, "ANSWER = Religion" could be used for a conversation that is broadly about religion.

A.4 Topic Model User Prompt

In the following conversation segment,

... {*same as forecasting prompt*}

[SEGMENT END]

What is the topic of the conversation?

A.5 Topics

- "Social Issues": ["homophobia", "transgenderism", "transgender issues", "transgender rights", "lgbt rights", "islamophobia", "racism", "sexism", "discrimination", "feminism", "social justice", "equal pay", "body image", "objectification", "rape", "sexual assault", "hate speech", "slurs", "marriage pressure", "alimony", "child support", "parental leave", "child abuse", "bullying", "polygamy"],
- "Politics and Law": ["politics", "gun control", "immigration ban", "judicial bias", "free speech", "affirmative action", "abortion", "censorship", "media bias", "socialism", "communism vs capitalism", "electoral college", "government", "nationalism", "patriotism", "travel ban", "us-saudi relations", "terrorism", "military draft", "war", "nuclear power", "capital punishment", "self-defense", "gun ownership", "gun rights", "gun regulation", "gun violence", "dueling laws", "prison", "corporal punishment", "death penalty", "military spending", "immigration", "don't ask don't tell (dadt)", "immigration enforcement", "immigration policy"],
- "Economics": ["economics", "cryptocurrency", "digital goods", "trade deficits", "minimum wage", "labor unions", "regulation", "social welfare", "alimony", "child support"],
- "Health": ["mental health", "vaccination", "vaccines", "cannabis", "marijuana", "opium trade", "prostitution", "sexting laws", "necrophilia", "veganism", "vegetarianism", "gmos"],
- "Culture and ID": ["cultural identity", "feminist terminology", "islam", "indigenous rights", "israeli-palestinian conflict", "israel", "jordan peterson", "hillary clinton emails",

"donald trump", "trayvon martin case", "kavanaugh nomination", "russian investigation", "cults vs religion", "historical figures", "metoo movement", "flag protest", "pride", "racial protests", "diversity debate", "transgender identity", "pronouns", "transgender dating", "transgender athletes", "transgender youth", "pride parades", "race genetics"],

- "Tech and Ent" : ["ad blocking", "game streaming", "journalism", "media bias", "censorship", "art censorship", "social media", "adblocking", "privacy", "american football", "college football", "sports", "star trek", "transgender athletes"],
- "Ethics and Morality": ["morality", "ethics", "free will", "circumcision", "animal rights", "organ donation", "evidence", "argumentation", "discipline", "historical judgment", "merging", "gift giving", "tipping", "hunting", "protected classes"]