Article and Comment Frames Shape the Quality of Online Comments

Anonymous ACL submission

Abstract

Framing theory posits that how information is presented shapes audience responses, but computational work has largely ignored audience reactions. While recent work showed that article framing systematically shapes the *content* of reader responses, this paper asks: does framing also affect response *quality*? Analyzing 1M comments across 2.7K news articles, we operationalize quality as comment health (constructive, good-faith contributions). We find that article frames significantly predict comment health while controlling for topic, and that comments that adopt the article frame are healthier than those that depart from it. Further, unhealthy top-level comments tend to generate more unhealthy responses, independent of the frame being used in the comment. Our results establish a link between framing theory and discourse quality, laying the groundwork for downstream applications. We illustrate this potential with a pro-active frame-aware LLM-based system to mitigate unhealthy discourse.[1]

1 Introduction

Online news platforms have associated comment threads in which readers not only can directly engage with the article, but also with each other. Maintaining constructive dialogue in these spaces, however, is challenging. While extensive computational work has focused on predicting comment toxicity or quality (Pavlopoulos et al., 2017; Founta et al., 2018; Shankaran and Sharma, 2024), it treats all discussions as equivalent, overlooking a fundamental insight from framing theory: the perspective (*framing*) through which issues are presented shapes how audiences respond (Entman, 1993; Scheufele, 1999; Chong and Druckman, 2007).

Traditionally, computational framing research analyzed only source texts (e.g., news articles, political speeches, social media posts) while ignoring how audiences responded to framed content (Card et al., 2015; Field et al., 2018; Liu et al., 2019).[2] Recent work has begun to address this gap by jointly analysing framing in articles and audience responses at scale (Guida et al., 2025). Analyzing news comment sections, they found that on average less than half of the comments retain the dominant article frame—the rest selectively adopt secondary frames or introduce entirely new ones. This proved to be true especially for value-laden frames (Morality, Fairness and Cultural Identity), suggesting that framing shapes response *content*: the perspectives that people adopt when discussing issues. However, a critical question remains: if the choice of frames shape how audiences interpret and reconstruct messages, does it also influence the *quality* of discussions?

We connect framing theory with discourse quality across 1M comments on 2.7K articles from The New York Times and The Globe and Mail. We operationalize discourse quality as **comment health**: the extent to which contributions are made in good faith, invite engagement, and focus on substance rather than hostility (Price et al., 2020). Healthy comments may include robust disagreement but remain constructive in tone. Health is thus different from the widely used concept *toxicity* and the two have been shown to correlate poorly (see Price et al. (2020) and Appendix B). Table 1 shows two examples each for healthy and unhealthy comments, taken from our corpus.

We posit two ways in which frames influence discussion quality. First, article frames may directly influence the health of comments. For instance, a news article framing immigration as a "security threat" versus an "economic opportunity" may activate different affective responses and generate more or less constructive discussions. Second, direct comments to the article serve as "secondary"

[1] Code and data will be released upon acceptance.

[2] See Ali and Hassan (2022) and Otmakhova et al. (2024) for overviews of computational framing analysis.

Healthy	Unhealthy
"How do you spell exploitation? This is a disgusting practice that sanctions abuses of fundamental rights and human decency. We cannot continue to condone it."	"Funny how the 'free' market is not OK in this situation. So 'not OK' that corps lobbied the 'Harper Government' to create a policy to circumvent the free market. How f@#ked is that? The Harper govt must go"
"Excellent article. For too long, the aboriginals' concerns have been treated as a problem to avoid, when it should be seen as a vital part of the whole project."	"Are you kidding? The mandatory native on all the projects I did with the Feds in northern Alberta was a uniformly drunk, indolent or absent 'partner'. What a joke"

Table 1: Examples from our corpus illustrating healthy versus unhealthy comments.

framers who mediate the article's message. Healthy or unhealthy replies to such comments may unfold: a healthy top-level comment using an economic frame may elicit healthier replies than a healthy comment using a moral frame, even when both respond to the same article. We therefore ask:

(RQ1) Do **article frames** influence the health of **top-level comments** through (a) frame type (which frames are used) and (b) frame alignment (whether commentators match or depart from article frames)?

(RQ2) Does the frame of **top-level comments** influence the health of **their replies**?

Importantly, all our analyses control for the confounding factor of article topic. We find that comment health significantly varies as a function of the article frame. Frame alignment also matters: comments that adopt the article frames are significantly healthier than those introducing new perspectives. Analysis of comment threads shows that healthy comments generate healthier replies through a cascade effect that operates consistently across all frame types.

We are the first to show a systematic impact of framing on discussion quality at scale on naturalistic data. Our findings can impact content moderation approaches, suggesting that rather than reacting to malicious content after it appears, platforms could proactively identify high-risk comments based on which frames are used, or whether commentators depart from article frames. We illustrate this through a frame, content and health aware LLM-based system that analyzes article and comments to provide real-time reformulation suggestions, helping commenters express views more constructively.

2 Data and Methods

2.1 Data

Our analysis examines news articles and associated comment threads from The New York Times (NYT), a major U.S. newspaper, and The Globe and Mail (SOCC), Canada's national newspaper. We build on the dataset from Guida et al. (2025), comprising news articles and comments from 2016-2019 across 11 topics (e.g., Immigration, Healthcare, Climate Change).

	NYT	SOCC	Total
Articles	1,671	1,077	2,748
Comments	831.9K	194.8K	1.03M

Table 2: Dataset Statistics.

We extend the data set in two key ways. First, while they sampled only top-level comments, we retrieve *complete comment threads* associated with these articles from the original datasets.[3] This enables us to examine both top-level comment health (RQ1) and how health propagates through reply chains (RQ2). Second, we apply frame and health classifiers to this expanded set of replies. Table 2 summarizes the final dataset.

2.2 Methods

Frame Classification We predict the primary and secondary frames of each article and comment by applying the fine-tuned RoBERTA classifier from Guida et al. (2025) to assign frame labels from a taxonomy of nine generic frames, with an additional Other category. Frame predictions are obtained for all news articles and comments.

Health Classification We assign a health score to each comment following the framework of Price et al. (2020). Under this definition, a *healthy* online conversation is one in which posts and comments are made in good faith, are not overly hostile or destructive, and generally invite engagement. Healthy conversations may include robust debate and disagreement but are typically focused on substance

[3] Kolhatkar et al. (2020) and https://www.kaggle.com/datasets/aashita/nyt-comments

Model	Acc.	Precision	Recall	F1
LLaMA	0.59	0.50	0.50	0.49
ModernBERT	0.68	0.69	0.71	0.67
DeBERTa	0.74	0.72	0.74	**0.72**

Table 3: Accuracy and macro-averaged precision, recall, F1 on the re-balanced UCC test set.

and ideas. Crucially, health does not require posts and comments to be friendly, grammatically correct, well structured, or free of vulgarity.We use their Unhealthy Comment Corpus (UCC) which provides binary labels indicating whether each comment *has a place in a healthy conversation*, annotated by up to five crowd workers with aggregated confidence scores (Price et al., 2020).

The original data are highly imbalanced, with healthy comments comprising over 90% of the corpus. To improve minority class representation, we resample the data by retaining only high-confidence annotations (≥ 0.8) and reducing majority class representation through undersampling of the healthy class. This procedure results in a more balanced dataset of approximately 10k instances (see Appendix Tables 4 and Table 5 for further details).

Models We fine-tuned two transformer-based models for health classification: DeBERTa-v3 (He et al., 2021), ModernBERT (Warner et al., 2024); and fine-tuned LLaMA 3.1–8B (Grattafiori et al., 2024). DeBERTa-v3 and ModernBERT are fine-tuned using class-weighted binary cross-entropy loss to mitigate class imbalance and improve minority-class recall. For LLaMA, we apply LoRA-based instruction fine-tuning (Hu et al., 2022).

Results As shown in Table 3, we obtained robust classifiers capable of distinguishing between healthy and unhealthy comments, with DeBERTa performing best overall. We hence use DeBERTa to predict comment health in the NYT and SOCC.

Three authors of this paper manually annotated a subset of 100 NYT/SOCC comments with moderate inter-annotator agreement (Fleiss' $\kappa = 0.54$). The average agreement between human and model labels was 78%, indicating reliable performance.

NYT and SOCC comment health predictions are skewed (75% healthy). This is expected because both platforms employ manual content moderation,[4] which establishes a relatively high baseline for discourse quality. We still retain over 250k comments labelled as unhealthy to support our main analysis.

3 Results

3.1 RQ1: Article Framing Effects on Comment Health

We examine (1) whether article frame type predicts health, and (2) whether frame alignment (matching the article's primary frame, adopting a secondary frame, or introducing new frames) affects health, while controlling for topic. We fit mixed-effects logistic regression to predict binary top-level comment health with random effects for article IDs and fixed effects for article topic (all models), article framing (RQ1.1) or frame alignment (RQ1.2).[5]

RQ1.1 Article Frame Effects Article framing exerts a significant influence on comment health in both outlets (NYT: $\chi^2(9) = 368.27$, $p < 0.001$; SOCC: $\chi^2(9) = 26.97$, $p = 0.001$). Health and Economic frames consistently elicit the healthiest comments across both platforms (84–87% healthy comments), while Political, Fairness, and Morality frames generate the least healthy discourse (72–77%), indicating that value-laden and polarizing frames provoke more contentious user engagement.

RQ1.2 Frame Alignment Effects The degree of alignment between article and comment frames significantly predicts comment health in both platforms (NYT: $\chi^2(2) = 340.61$, $p < .001$; SOCC: $\chi^2(2) = 48.16$, $p < .001$), after controlling for topic. A clear gradient emerges (Figure 1): comments that match article frames are the healthiest, followed by selective reframing (adopting secondary frames present in the article), with complete reframing (introducing frames absent from the article) exhibiting the lowest health. All three pairwise comparisons between frame alignment conditions are significant at $p \ll .001$ (See Table 10 in the Appendix). This trend holds across topics (see Figure 2 in Appendix C.1).

The selective reframing category is particularly informative. Readers who remain within the article's *frame repertoire*, even when shifting away from its primary emphasis, tend to engage more constructively than those who introduce completely novel perspectives.

[4]Guidelines: `https://help.nytimes.com/hc/en-us/articles/115014792387-Comments` (NYT), https://www.theglobeandmail.com/community-guidelines/ (SOCC).

[5]Full model specifications in Appendix D.1.

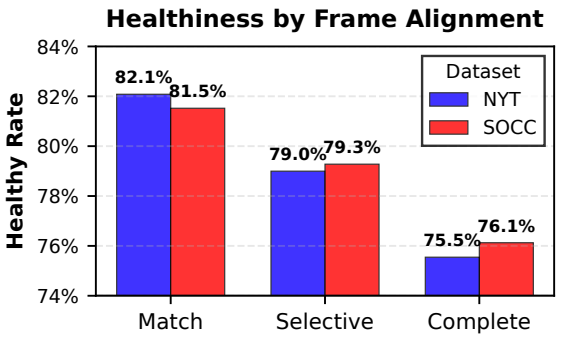

Figure 1: Health by frame alignment across platforms. A clear gradient emerges where comments matching article frames have the highest health rates, followed by selective reframing (adopting secondary frames present in the article), with complete reframing (introducing frames absent from the article) showing the lowest health.

Topic Effects Unsurprisingly, article topic alone also strongly predicts comment health when controlling for frame type (NYT: $\chi^2(10) = 371.85$, $p < .001$; SOCC: $\chi^2(10) = 435.52$, $p < .001$). Some topics (Health, Education) consistently foster more constructive discourse, while others are associated with substantially lower health (Trump). Full results by topic are in Appendix Figure 2.

3.2 RQ2: Comment Framing Effects on Reply Health

We model mean reply health by fitting a linear regression with top-level comment health, top-level comment frame and topic as well as health×frame interactions as predictors. We then test (1) how top-level comment health impacts subsequent discussion and (2) whether this trend differs across frames; while again controlling for topic.[6]

Top-level comment health strongly predicts subsequent reply health in both outlets ($p < 0.001$). This is in line with similar studies on toxicity in Reddit threads (Shankaran and Sharma, 2024). This trend does not significantly vary by top-level comment frame. In other words, healthy top comments initiate healthy discussions and unhealthy top comments initiate unhealthy discussions – largely independent of their topic or frame. Full results are in Appendix Table 12.

Combined with RQ1—where frame type and alignment significantly predicted baseline comment health—this suggests a two-stage model, in which article framing sets discourse environments, while individual comment health governs propagation through reply chains.

3.3 Frame-Aware Content Moderation

Current content moderation systems operate reactively: they detect and remove toxic content after posting. This approach addresses symptoms rather than causes, it penalizes users without helping them improve, and cannot prevent unhealthy discussions before they escalate.

Our findings suggest an alternative: proactive, frame-aware moderation. We illustrate this through a prototype system that analyzes article framing, comment framing, their frame alignment and comment health. Based on these inputs, the system stratifies comments into three risk levels (high, medium and low), and suggests LLM-based reformulations informed by full context. The system can be accessed online at this **link**.[7]

4 Discussion and Conclusion

Our findings suggest that framing and health of online discourse are intertwined. The primary article frame significantly predicts comment health, with value laden frames like Morality and Fairness frames generating the least healthy comments. Furthermore, frame alignment is crucial: departing from the article's frame correlates with decreased health, with complete reframing showing the strongest negative effects. At the reply level (RQ2), we find that comment health is the primary driver of reply quality, operating uniformly across all frame types, with top-level comment health strongly predicting reply health.

We provide computational evidence for framing theory (Neuman et al., 1992; Scheufele, 1999), which posits that audiences actively interpret and reconstruct media frames rather than passively accepting them. We extend this line of work by illustrating that frame reconstruction patterns predict not just *what* audiences say but *how* they engage.

Beyond theoretical significance in the context of framing theory and mechanisms, our findings are of practical importance for maintaining healthy online discourse. We illustrate this potential through a pro-active, frame-aware comment moderation prototype that combines frame classification with health detection to provide constructive reformulation suggestions.

[6] Full modeling details in Appendix D.2.

[7] As it uses serverless GPU infrastructure, initial requests may experience a short delay while models are initiated and loaded. See Appendix E for complete technical specifications.

Limitations

Our analysis examines correlational patterns rather than causal mechanisms. Our binary health operationalization enables large-scale analysis while simplifying the multidimensional nature of discourse quality. We analyze two major English-language news outlets (NYT, Globe and Mail) from 2012-2018, which may reflect outlet-specific moderation practices and temporal context. Our thread analysis focuses on top-level comments and immediate replies (depth ≤ 2); deeper nested dynamics warrant further investigation. The frame and health classifiers achieve robust performance on our datasets, though generalization to other platforms and languages remains to be validated.

Ethical Considerations

Our moderation prototype is intended to support constructive discourse, not suppress dissent. We caution against using frame alignment as a sole criterion for moderation, as novel perspectives that depart from article framing may represent valuable contributions. The operationalization of "health" reflects norms from specific cultural contexts and should be adapted for other communities.

References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dennis Chong and James Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10.

Robert Entman. 1993. Framing: Toward clarification of a fractured paradigm. *The Journal of Communication*, 43:51–58.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Matteo Guida, Yulia Otmakhova, Eduard Hovy, and Lea Frermann. 2025. Retain or reframe? a computational framework for the analysis of framing in news articles and reader comments. *Preprint*, arXiv:2507.04612.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Veselin Kolhatkar, Haofen Wu, Liane Cavasso, and Maite Taboada. 2020. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190. Published: 2 November 2019; Issue Date: June 2020.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

W. Russell Neuman, Marion R. Just, and Ann N. Crigler. 1992. *Common Knowledge: News and the Construction of Political Meaning*. University of Chicago Press, Chicago.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017*

Conference on Empirical Methods in Natural Language Processing, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Dietram A. Scheufele. 1999. Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.

Vigneshwaran Shankaran and Rajesh Sharma. 2024. Analyzing toxicity in deep conversations: A reddit case study. *Preprint*, arXiv:2404.07879.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

A UCC Data Splits

The original train / val / test UCC data splits and our label-balanced data splits are shown in Table 4 and 5, respectively.

	Healthy	Unhealthy	Total
Train	32,848	2,655	35,503
Val	4,091	336	4,427
Test	4,105	320	4,425
Total	41,044	3,311	44,355

Table 4: Original class distribution in UCC splits.

	Healthy	Unhealthy	Total
Train	5,298	2,649	7,947
Val	662	331	993
Test	662	331	993
Total	6,622	3,311	9,933

Table 5: Balanced high-confidence split used for fine-tuning.

B Health and Toxicity

Price et al. (2020) showed that 'health' and 'toxicity' only have a weak correspondence in UCC. We verify this trend on our NYT and SOCC corpus, adding support for our decision to focus on comment health, specifically.

To assess how our healthy discourse classifier relates to toxicity detection, we compared our health predictions with Perspective API toxicity scores on the SOCC and NYT datasets. We binarized toxicity scores at a threshold of 0.5 and calculated Cohen's κ and correlations.

Results show slight agreement ($\kappa = 0.19$–0.21) with moderate negative correlations ($\rho \approx -0.51$, $p < 0.001$). 20–24% of comments classified as unhealthy have low toxicity scores, accounting for 96–97% of instances classified as unhealthy, but with low toxicity scores. Upon manual investigation of a subset of disagreements, we found that these comments are dismissive, unproductive, use sweeping generalizations or stereotypes, and employ condescending or sarcastic tones. As such, these comments undermine constructive and healthy dialogue, however, they do not necessarily use explicit or overly emotional language which would give rise to a toxic label. Examples of such cases are provided in Appendix Table 6.

C Overall Topic Health

Table 7 details the rate of healthy comments by topic. The variation in health by topic is independent of framing and highly significant in both datasets (NYT: $\chi^2(10) = 371.9$, $p < .001$; SOCC: $\chi^2(10) = 435.5$, $p < .001$). Healthcare and Education consistently foster the most constructive discourse, while Trump coverage generates the lowest health rates.

C.1 Frame Alignment Effects by Topic

Figure 2 shows how frame alignment effects vary across topics. The consistent pattern—where matching frames yield healthier comments than selective or complete reframing—holds across most topics.

D Regression Analyses and Full Results

D.1 RQ1: Article effects on top-level comments

For RQ1, we ask how article framing influences comment health. First, we examine whether article frame type predicts health. We fit a mixed-effects logistic regression model that predicts (binary) health of a top-level comment based on article frame, topic as fixed effects and article ID as

Comment Text	Healthy	Toxicity	Characteristics
"A few general comments: did not read the column. Stopped reading Lady Astor long ago; interesting she did not editorialize about Chris Spence, admitted serial plagiarizer, head of TDSB; Really, who still reads her columns? A columnist who's plagiarizes?"	0	0.127	Dismissive, unconstructive
"The number of lies, distortions and exaggerations in this article are many. Wonder where she got them?"	0	0.243	Accusatory
"America: it's an amazing country full of a lot of fine people. But like wayward teenagers you can't tell them anything. They know what's best, even when it's so obvious that their behavior is harming themselves and others. There will be no significant change in America's gun-cult mentality. Not now, nor even if there's 20 more shootings like Newtown. Guns are fetish objects to Americans, and they're worshiped accordingly. Either accept that, or live and visit elsewhere."	0	0.188	Stereotypes, generalization
"Get with the program, Margaret, or you'll be sent to a re-education camp. It's 'climate change.'"	0	0.049	Sarcasm

Table 6: Examples of comments classified as unhealthy (Healthy=0) but with low Perspective API toxicity scores (<0.5).

Topic	NYT Health	SOCC Health
Healthcare	88%	95%
Education	88%	87%
Climate Change	85%	86%
Abortion	83%	86%
Syria	83%	77%
Gun Control	79%	78%
Israel	79%	70%
Russia	78%	72%
Gay Rights	78%	76%
Immigration	78%	80%
Trump	68%	67%

Table 7: Mean comment health by topic, sorted by NYT rate. Differences between topics are statistically significant ($p < 0.001$) for both platforms.

random effect due to multiple measurements (top comments) for the same article. Our variable of interest is article_frame, and we include topic as a control variable:

$$\text{health} \sim \text{article\_frame} + \text{topic} + (1|\text{id}). \quad (1)$$

Second, we test the effect of frame alignment between articles and comments predicts health (frame_condition). Similarly, we fit a logistic regression mixed-effects model replacing article_frame with frame_condition,

$$\text{health} \sim \text{frame\_condition} + \text{topic} + (1|\text{id}), \quad (2)$$

where frame_condition is a three-level factor indicating whether comments match the article's primary frame, selectively reframe by adopting a secondary frame present in the article, or completely reframe by introducing frames not in the article.

Full Results Tables 8 and 9 present the full regression models for frame alignment and article frame effects, respectively. Tables 10 and Table 11 summarize the estimated marginal means.

D.2 RQ2: Top-level comment effects on reply thread health

For RQ2, we model mean reply health (MRH) using ordinary least squares (OLS) linear regression, specifying the model as follows:

$$\begin{aligned} \text{MRH} \sim\ & \text{top\_c\_health} + \text{top\_c\_frame} \\ & + \text{article\_topic} \\ & + \text{top\_c\_health} \times \text{top\_c\_frame}, \end{aligned} \quad (3)$$

where top_c is short for top_comment. We include article_frame again as a control factor and also capture the interaction of health and frame in the interaction term.

Full Results Table 12 present full coefficient estimates, standard errors, t-statistics, and p-values for both platforms.

E Frame-Aware Moderation System: Technical Specification

This section provides complete technical details of the frame-aware moderation prototype described in Section 3.3, accessible at the following link: https://mpprng--comment-moderation-agent-commentmoderationservice-serve.modal.run/.

The prototype system is deployed on Modal[8], a

[8] https://modal.com

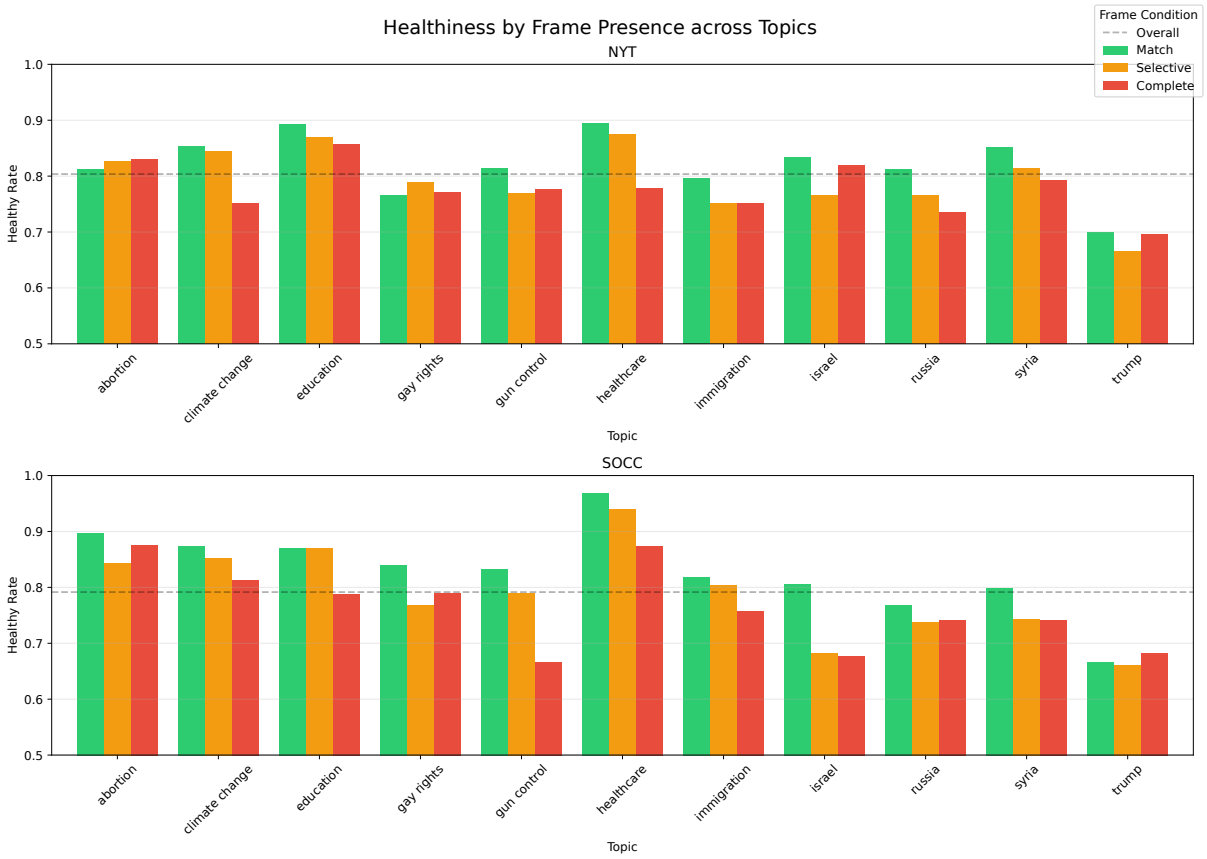

Figure 2: Comment health by frame alignment (Match, Diff in Article, Never in Article) across topics for NYT (top) and SOCC (bottom). The dashed line represents overall health.

serverless GPU infrastructure platform. This deployment approach offers cost efficiency for research prototypes but introduces cold-start latency when the system has been idle.

Cold Start Behavior When no requests have been made for approximately 5–10 minutes, the system enters an idle state. The first subsequent request triggers initialization of:

- Ollama server and Gemma 3:1b model loading
- DeBERTa-v3-base healthiness classifier
- RoBERTa-based frame classifier

Our deployment makes use of NVIDIA A10G GPU with 24GB VRAM.

E.1 System Architecture

Given a user-based keyword input (e.g., "Climate change"), our system scrapes the three most recent articles from the The Conversation website[9] and classifies the article frames, the comment frames, the type of reframing (or match) and the comment health. These components, along with the article and comment text, are parsed by an LLM, which is then prompted to make suggestions based on decision heuristics. Table 13 provides a complete overview of the system pipeline.

E.2 LLM Prompt Structure

For reformulation suggestions, we use Gemma 3:1b deployed via Ollama for local inference. The prompt includes full context (article text, frames, comment, alignment) and requests structured JSON output with risk level confirmation, 2–3 specific suggestions, and an allow/block recommendation. Example prompt structure is shown in Figure 3.

E.3 Interface Example

Figures 4 and 5 illustrate the system interface and moderation workflow, with an example of comment analysis on a news article.

[9] https://theconversation.com/us/, CC BY-ND 4.0

	NYT				SOCC			
Predictor	**Est.**	**SE**	z	p	**Est.**	**SE**	z	p
(Intercept)	1.727	0.099	17.37	< 0.001	1.862	0.173	10.77	< 0.001
Frame Alignment (vs Match)								
Selective	-0.108	0.009	-11.94	< 0.001***	-0.132	0.029	-4.51	< 0.001***
Complete	-0.298	0.018	-16.31	< 0.001***	-0.289	0.044	-6.62	< 0.001***
Article Topic (vs Abortion)								
Climate Change	-0.022	0.113	-0.20	0.843	0.047	0.176	0.27	0.788
Education	0.542	0.131	4.14	< 0.001***	0.131	0.185	0.71	0.480
Gay Rights	-0.328	0.179	-1.83	0.067	-0.268	0.219	-1.22	0.221
Gun Control	-0.288	0.117	-2.45	0.014*	-0.500	0.214	-2.34	0.020*
Healthcare	0.410	0.107	3.84	< 0.001***	0.976	0.209	4.67	< 0.001***
Immigration	-0.380	0.111	-3.42	< 0.001***	-0.316	0.181	-1.74	0.081
Israel	-0.265	0.131	-2.02	0.043*	-0.792	0.191	-4.15	< 0.001***
Russia	-0.374	0.105	-3.55	< 0.001***	-0.744	0.186	-3.99	< 0.001***
Syria	-0.095	0.137	-0.69	0.491	-0.428	0.181	-2.37	0.018*
Trump	-0.862	0.107	-8.09	< 0.001***	-1.103	0.177	-6.24	< 0.001***
Random Effects								
Article ID (Intercept) Variance: 0.369 (NYT), 0.201 (SOCC)								
Article ID (Intercept) SD: 0.608 (NYT), 0.448 (SOCC)								
Model Fit								
NYT: AIC = 363749.3, BIC = 363899.9					SOCC: AIC = 39157.4, BIC = 39277.5			
Overall Effects (Type II Wald χ^2 tests)								
Frame Alignment: $\chi^2(2) = 340.61, p < 0.001$***					Frame Alignment: $\chi^2(2) = 48.16, p < 0.001$***			
Topic: $\chi^2(10) = 627.65, p < 0.001$***					Topic: $\chi^2(10) = 662.46, p < 0.001$***			

Table 8: Mixed-effects logistic regression predicting comment health from frame alignment, NYT and SOCC.

System Instruction:
You are an AI comment moderator. Analyze this comment for health and frame transfer (reframing). Provide constructive suggestions only when the comment is unhealthy or uses a completely different perspective from the article.

CONTEXT:
{context}

Article Text: {article}...

Comment to Analyze: {comment}

Trigger: This comment requires intervention due to: {intervention_trigger}

Task:
Based on health and frame transfer analysis:
1. Confirm the risk level (low, medium, high).
2. Provide 2-3 specific, constructive reformulations that:

- Improve health if unhealthy
- Help align comment with article frames if reframing is detected
- Maintain the core message

3. Determine if the original comment should be allowed.

Provide a JSON response.

Figure 3: LLM prompt used for comment reformulation.

	NYT				SOCC			
Predictor	**Est.**	**SE**	z	p	**Est.**	**SE**	z	p
(Intercept)	1.608	0.108	14.90	< 0.001	1.734	0.196	8.85	< 0.001
Article Frame (vs Cultural)								
Economic	-0.104	0.098	-1.06	0.288	0.185	0.097	1.92	0.055
Fairness	-0.560	0.275	-2.04	0.042*	-0.544	0.195	-2.79	0.005**
Health	0.390	0.085	4.61	< 0.001***	0.051	0.122	0.42	0.673
Legality	-0.262	0.085	-3.06	0.002**	-0.144	0.133	-1.08	0.279
Morality	0.016	0.147	0.11	0.914	-0.125	0.216	-0.58	0.563
Opinion	-0.213	0.157	-1.36	0.174	0.017	0.206	0.08	0.936
Other	0.096	0.083	1.16	0.245	-0.032	0.097	-0.33	0.743
Political	-0.553	0.069	-7.96	< 0.001***	-0.064	0.087	-0.73	0.466
Security	-0.219	0.125	-1.75	0.080	-0.015	0.113	-0.14	0.891
Article Topic (vs Abortion)								
Climate Change	0.167	0.103	1.62	0.104	0.024	0.184	0.13	0.898
Education	0.801	0.121	6.59	< 0.001***	0.177	0.193	0.92	0.360
Gay Rights	-0.051	0.165	-0.31	0.758	-0.155	0.225	-0.69	0.491
Gun Control	-0.033	0.108	-0.30	0.762	-0.433	0.219	-1.98	0.048*
Healthcare	0.547	0.096	5.70	< 0.001***	0.906	0.211	4.30	< 0.001***
Immigration	0.041	0.103	0.40	0.687	-0.314	0.185	-1.70	0.089
Israel	0.023	0.122	0.19	0.847	-0.736	0.198	-3.71	< 0.001***
Russia	0.104	0.098	1.06	0.291	-0.715	0.194	-3.68	< 0.001***
Syria	0.075	0.138	0.55	0.584	-0.384	0.190	-2.03	0.043*
Trump	-0.385	0.099	-3.90	< 0.001***	-0.992	0.185	-5.35	< 0.001***
Random Effects								
Article ID (Intercept) Variance: 0.292 (NYT), 0.195 (SOCC)								
Article ID (Intercept) SD: 0.540 (NYT), 0.442 (SOCC)								
Model Fit								
NYT: AIC = 363757.2, BIC = 363983.1					SOCC: AIC = 39192.5, BIC = 39372.7			
Overall Effects (Type II Wald χ^2 tests)								
Article Frame: $\chi^2(9) = 368.27$, $p < 0.001$***					Article Frame: $\chi^2(9) = 26.97$, $p = 0.001$**			
Topic: $\chi^2(10) = 371.85$, $p < 0.001$***					Topic: $\chi^2(10) = 435.52$, $p < 0.001$***			

Table 9: Mixed-effects logistic regression predicting comment health from article frames, NYT and SOCC.

Alignment	**NYT**	**SOCC**
Match	82.9%	83.1%
Selective	81.3%	81.1%
Complete	78.2%	78.6%
Pairwise Comparisons (Tukey-adjusted):		
Match vs Selective	OR = 1.11***	OR = 1.14***
Match vs Complete	OR = 1.35***	OR = 1.33***
Selective vs Complete	OR = 1.21***	OR = 1.17***

Table 10: Estimated marginal means: Frame alignment effects on health. Matching is the reference level.

Note: Results averaged over topic levels. All pairwise comparisons significant at $p < 0.001$.

Article Frame	**NYT**	**SOCC**
Health	89.2%	82.4%
Other	86.1%	81.2%
Cultural	84.9%	81.7%
Economic	83.5%	84.3%
Opinion	81.9%	81.9%
Security	81.8%	81.5%
Legality	81.2%	79.4%
Morality	85.1%	79.7%
Political	76.3%	80.7%
Fairness	76.2%	72.1%

Table 11: Estimated marginal means: Comment health by article frame.

Note: Results averaged over topic levels. Frames sorted by NYT health rate.

	NYT				SOCC			
Predictor	β	**SE**	t	p	β	**SE**	t	p
(Intercept)	0.706	0.019	37.39	< 0.001	0.633	0.029	21.49	< 0.001
Top Comment Health								
Healthy (vs Unhealthy)	0.095	0.018	5.38	< 0.001***	0.126	0.023	5.40	< 0.001***
Top Comment Frame (vs Cultural)								
Economic	0.026	0.023	1.14	0.256	0.070	0.028	2.48	0.013*
Fairness	−0.003	0.031	−0.09	0.932	−0.102	0.041	−2.51	0.012*
Health	0.006	0.023	0.27	0.787	0.047	0.044	1.07	0.286
Legality	0.020	0.021	0.99	0.323	−0.022	0.048	−0.47	0.642
Morality	−0.034	0.020	−1.75	0.079	−0.046	0.031	−1.48	0.140
Opinion	−0.032	0.044	−0.74	0.458	−0.023	0.048	−0.49	0.626
Other	−0.020	0.017	−1.15	0.251	−0.019	0.022	−0.83	0.406
Political	−0.034	0.016	−2.04	0.041*	−0.023	0.023	−1.03	0.303
Security	0.057	0.045	1.26	0.207	−0.053	0.047	−1.13	0.258
Article Topic (vs Abortion)								
Climate Change	0.030	0.011	2.81	0.005**	0.053	0.021	2.48	0.013*
Education	0.058	0.012	4.75	< 0.001***	0.087	0.022	3.91	< 0.001***
Gay Rights	0.026	0.016	1.61	0.108	0.094	0.027	3.51	< 0.001***
Gun Control	0.028	0.012	2.46	0.014*	0.060	0.027	2.24	0.025*
Healthcare	0.034	0.010	3.30	0.001**	0.122	0.024	5.12	< 0.001***
Immigration	0.017	0.011	1.62	0.105	0.063	0.022	2.87	0.004**
Israel	0.042	0.013	3.21	0.001**	−0.025	0.024	−1.06	0.289
Russia	0.010	0.010	0.99	0.320	0.006	0.023	0.24	0.807
Syria	0.030	0.014	2.17	0.030*	0.038	0.022	1.72	0.085
Trump	−0.024	0.010	−2.29	0.022*	−0.004	0.022	−0.18	0.855
Interactions: Health × Frame								
Healthy × Economic	0.003	0.024	0.13	0.894	−0.044	0.030	−1.44	0.151
Healthy × Fairness	−0.024	0.035	−0.70	0.487	0.024	0.047	0.50	0.614
Healthy × Health	0.039	0.025	1.56	0.118	−0.013	0.047	−0.29	0.776
Healthy × Legality	−0.006	0.022	−0.28	0.782	0.003	0.052	0.07	0.948
Healthy × Morality	−0.012	0.022	−0.54	0.589	−0.006	0.037	−0.15	0.879
Healthy × Opinion	0.039	0.046	0.83	0.406	0.001	0.053	0.02	0.988
Healthy × Other	0.014	0.019	0.74	0.457	−0.004	0.025	−0.15	0.878
Healthy × Political	−0.010	0.018	−0.58	0.563	−0.009	0.025	−0.35	0.727
Healthy × Security	−0.054	0.047	−1.14	0.254	0.048	0.050	0.95	0.340
Model Fit								
NYT: $R^2 = 0.029$; Adj. $R^2 = 0.028$; $F(29, 72{,}800) = 73.82$, $p < 0.001$; Residual SE: 0.342								
SOCC: $R^2 = 0.050$; Adj. $R^2 = 0.049$; $F(29, 20{,}013) = 36.36$, $p < 0.001$; Residual SE: 0.334								

Table 12: Linear regression predicting mean reply health from top comment health, frame, and topic, NYT and SOCC.

Component	Input / Function	Implementation Details
Article Analysis		
Text Processing	Full article text	Sentence tokenization (NLTK `punkt`)
Frame Classification	Sentence-level frame detection	Fine-tuned RoBERTa model (Guida et al., 2025) (10-frame taxonomy)
Output	Sentence-level frame labels	Each sentence annotated as `[{frame, confidence}]`
Comment Analysis		
Text Processing	Comment text	Sentence tokenization (NLTK `punkt`)
Frame Classification	Sentence-level frame detection	Same RoBERTa model as article analysis
Health Prediction	Comment-level quality assessment	Fine-tuned DeBERTa-v3-base (Binary: healthy / unhealthy)
Frame Alignment	Article–comment frame comparison	Primary-frame overlap and divergence measures
Risk Assessment		
Risk Stratification	Health + alignment signals	Rule-based aggregation (Table 14)
Moderation Decision	Risk level	Boolean allow / block decision
Intervention Generation		
Context Construction	Inputs to LLM	Article text, top-5 article frames, comment text, comment frames, alignment status, health score
LLM Inference	Reformulation generation	Gemma 3:1b via Ollama
Output Format	Moderation guidance	Structured JSON: `{risk_level, suggestions[], allow_post}`

Table 13: System architecture for frame-aware comment moderation. Articles and comments are analyzed separately for framing and health, combined into a risk assessment, and used to generate context-aware interventions via an LLM.

Risk Level	Health	Alignment	Action
High	< 0.3	Any	Suggest + Flag
High	< 0.5	Complete	Suggest + Flag
Medium	$[0.3, 0.6)$	Any	Suggest
Medium	≥ 0.6	Selective/Complete	Suggest
Low	≥ 0.6	Match	Allow

Table 14: Risk stratification decision rules.

Figure 4: Landing page with topic search functionality. Users can enter keywords or select predefined topics to retrieve 3 relevant articles from The Conversation. Article view with sentence-level frame analysis. Detected frames are displayed as tags; hovering over a frame highlights corresponding sentences in the article text.

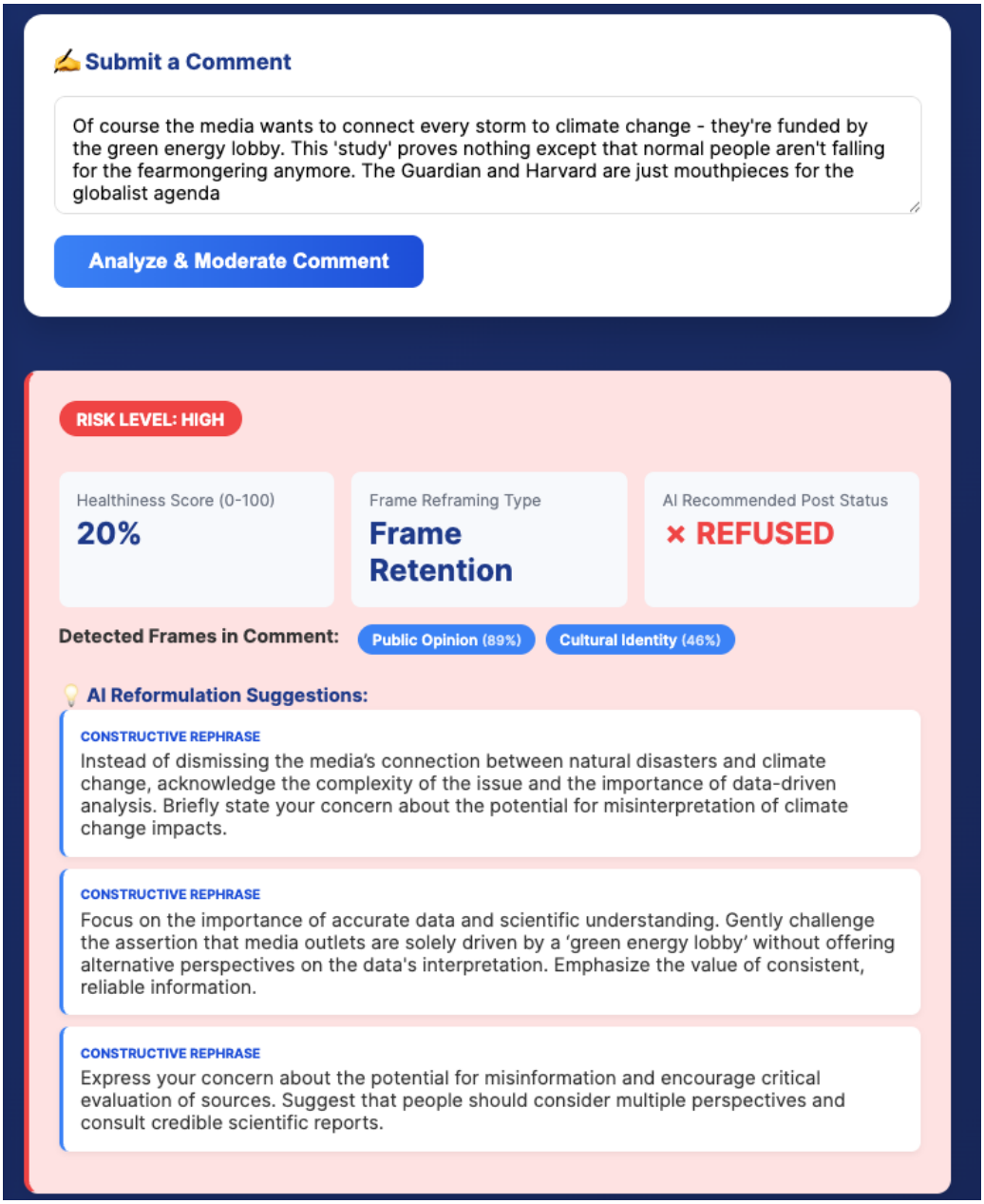

Figure 5: Moderation output for an unhealthy comment expressing climate skepticism. The system assigns a high risk level based on healthiness score, detects frames and retention, and recommends refusing the comment. The LLM generates three ways to reformulate more constructively while preserving the user's core concern and idea.