Supplementary Information for "Learning Human Feedback from Large Language Models for Content Quality-aware Recommendation"

HUILI WANG, Department of Electronic Engineering, Tsinghua University, China
CHUHAN WU, Huawei Noah's Ark Lab, China
YONGFENG HUANG, Department of Electronic Engineering, Tsinghua University, China and Zhongguancun Laboratory, China
TAO QI*, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China and Tsinghua University, China

ACM Reference Format:
Huili Wang, Chuhan Wu, Yongfeng Huang, and Tao Qi. 2025. Supplementary Information for "Learning Human Feedback from Large Language Models for Content Quality-aware Recommendation". 1, 1 (March 2025), 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

SUPPLEMENTARY MATERIALS

Warning: this document contains example data that may be offensive or harmful. All examples in this document are solely used for academic research purposes and do not reflect our opinions.

Experimental Settings

Experimental Datasets. We construct extensive experiments based on three publicly available real-world content recommendation datasets: MIND, MIND-1M and EB-NeRD. The MIND series datasets are collected from practical Microsoft News platform over a period of 6 weeks and are available at https://msnews.github.io/. The MIND-1M dataset consists of more than 15 million impression logs generated by 1 million users, while the MIND dataset is a randomly chosen subset comprising 50,000 users from this extensive collection. The EB-NeRD is a large-scale Danish dataset created by Ekstra Bladet and is available at https://recsys.eb.dk/dataset/. We use the version of the EB-NeRD which randomly samples 50,000 users and their behavior logs from the full dataset. Specifically, in these datasets, each impression log records the content that a user engaged with, including both clicked and non-clicked content. Detailed results are kept in the code repository: https://github.com/wanghl21/HFAR.

*Corresponding author.

Authors' addresses: Huili Wang, whl21@mails.tsinghua.edu.cn, Department of Electronic Engineering, Tsinghua University, Beijing, China, 100084; Chuhan Wu, wuchuhan15@gmail.com, Huawei Noah's Ark Lab, Beijing, China; Yongfeng Huang, yfhuang@tsinghua.edu.cn, Department of Electronic Engineering, Tsinghua University, Beijing, China, 100084 and Zhongguancun Laboratory, Beijing, China, 100094; Tao Qi, taoqi.qt@gmail.com, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China, 100876 and Tsinghua University, Beijing, China, 100084.

XXXX-XXXX/2025/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Alignment on Content Quality Between LLMs and Human. To further evaluate the capacity of LLM to simulate the human feedback and the effectiveness of HFAR, we conduct an empirical analysis using a small dataset consisting of 5,000 items randomly sourced from MIND. These items are manually annotated by volunteers who assess content quality from three distinct, targeted perspectives, independently of any LLM-generated annotations. Results show that the consistency of quality feedback provided by LLM and human annotation can achieve 96.33% on average (Table 1).

	Clickbait	Racialism	Violence	High-quality	Average
Consis. Rate	95.44	99.72	97.4	92.76	96.33

Table 1. The consistency of human annotation vs human feedback simulated by LLM (GPT-3.5-turbo) on content quality from various perspectives.

Alignment on Content Quality between LLMs. We extend our evaluation to include simulations of human feedback on content quality across widely used LLM models, such as ChatGLM-4 and ChatGPT-4. We then analysis the consistency of feedback on content quality between these models. The consistency rates are shown in Table 2, where we observe that the three LLMs achieve high consistency (over 95% on average) in simulating human feedback. These results demonstrate that different LLMs, which are aligned with human expectations, can provide similar feedback on content quality.

LLMs	Clickbait	Racism	Violence	Average
ChatGLM-4 vs ChatGPT-4o	92.92	99.85	97.23	96.67
ChatGPT-3.5 vs ChatGPT-4o	91.76	99.61	96.02	95.80
ChatGPT-3.5 vs ChatGLM-4	94.62	99.85	97.85	97.44

Table 2. The consistency of different LLMs in simulating human feedback on content quality from various perspectives.

Experimental Metrics Settings. Following previous works, we use four metrics, including *AUC*, *MRR*, *nDCG*@5 and *nDCG*@10, to evaluate the recommendation methods' ability to model and match user information needs.

The *AUC* is commonly assessed using the Area under the Receiver Operating Characteristic (ROC) Curve score.

$$AUC = \frac{\sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} I\left[P(p) > P(n)\right]}{\mid \mathcal{P} \mid\mid \mathcal{N} \mid}, \tag{1}$$

where $P(\cdot)$ is the prediction score of a sample, $\mathcal{P}$ and $\mathcal{N}$ respectively denote the positive and negative sample sets. $I[\cdot]$ denotes an event indicator function.

The Mean Reciprocal Rank (*MRR*) metric is based on the concept of rank, which refers to the position of the first relevant item in the list of returned results. It emphasizes the importance of the first few results, which is critical for user experience.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}, \tag{2}$$

where $|Q|$ is the total number of query and rank_i is the rank of the first relevant item for query i.

The Normalized Discounted Cumulative Gain ($nDCG@K$) is a metric commonly used in information retrieval and recommendation systems to evaluate the effectiveness of a ranked list of items or search results. It takes into account both the relevance of each content in the list and its position.

$$nDCG@K = \frac{\sum_{i=1}^{K} \frac{2^{r_i}-1}{\log_2(1+i)}}{\sum_{i=1}^{N_p} \frac{1}{\log_2(1+i)}}, \tag{3}$$

where N_p is the number of positive samples, and r_i is the relevance score of content with i-th rank, which is 1 for clicked content and 0 for non-clicked content.

Furthermore, we propose a new metric denoted as $RQ@K$ to measure the quality of the recommender system based on the frequency of low-quality information in the top K candidate content.

$$RQ@K = \frac{1}{\mid U_t \mid} \sum_{u \in U_t} \frac{\mid R_u^K \cap T^{\varphi} \mid}{\mid T^{\varphi} \mid}, \quad T^{\varphi} = T^{\varphi_1} \cup T^{\varphi_2} \cup \cdots T^{\varphi_m}, \tag{4}$$

where T^{φ_i} is the content in test dataset which hit the quality factor φ_i and U_t denotes the set of test users. R_u^K is the top K recommended content from the test dataset for user u.

Case Study

Low-quality Recommended Content in Recommendation Platforms. There are some examples that people blame for low-quality recommended content in real-world recommendation platforms collected from social media (Supplementary Fig. 1, Fig. 2, Fig. 3 and Fig. 4). Users express indignation, frustration, and concern over the dubious quality content recommended by real-world information delivery platforms. Besides, most information delivery platform always emphasize the modeling of user preferences while overlooking the intrinsic content quality of items. This oversight not only compromises the user experience but also has the potential to transform the platform into a disseminator of toxic information, thereby impacting societal stability.

The Evaluation of Feedback on Content Quality. In this section, we present several real cases to show the content quality evaluation result from gpt-3.5-turbo (Supplementary Figure. 5). In the first row, we present the prompt template, where $[x]$ and $[y]$ represent the title and content, respectively. We then substitute these symbols with the titles and raw contents from the information delivery platform to obtain the evaluation results. We also highlight the evidence for judging the results. From these cases, we observe that the LLM (gpt-3.5-turbo) is able to correctly understand the title and content. Moreover, it demonstrates alignment with human feedback on content quality and provides accurate justifications for its evaluations.

WORKFLOW OF HFAR

The detailed workflow of HFAR is summarized in Algorithm 1.

Algorithm 1 Pseudo-code algorithm description for HFAR

Require: user behavior record content n, candidate content c, content encoder $f_c(\cdot)$, user encoder $f_u(\cdot)$, LLM π

content representation $\boldsymbol{n} \leftarrow f_c(n), \boldsymbol{c} \leftarrow f_c(c)$
user representation $\boldsymbol{u} \leftarrow f_u([\boldsymbol{n}_1, \boldsymbol{n}_2, \cdots, \boldsymbol{n}_k])$
calculate relevance score S_r based on Eq.3
calculate morality judgement q for content based on LLM π
calculate morality score S_q based on Eq.7
for each sample in D_q **do**
    calculate loss $\mathcal{L}_c$ based on Eq.8 to update the morality encoder $\Psi(\cdot)$ on D_q
    calculate loss $\mathcal{L}_t$ based on Eq.10 to update the morality scoring model $\Phi(\cdot)$ on D_q
    form recommendation training score $\hat{S}_f \leftarrow S_r + \alpha S_q$
    calculate loss $\mathcal{L}_r$ based on Eq.12 to update the relevance modeling module on D_q
    calculate loss $\mathcal{L}_f$ based on Eq.13 to jointly optimize modules in HFAR on D_q
end for
form recommendation inference score $S_f \leftarrow S_r + \beta S_q$

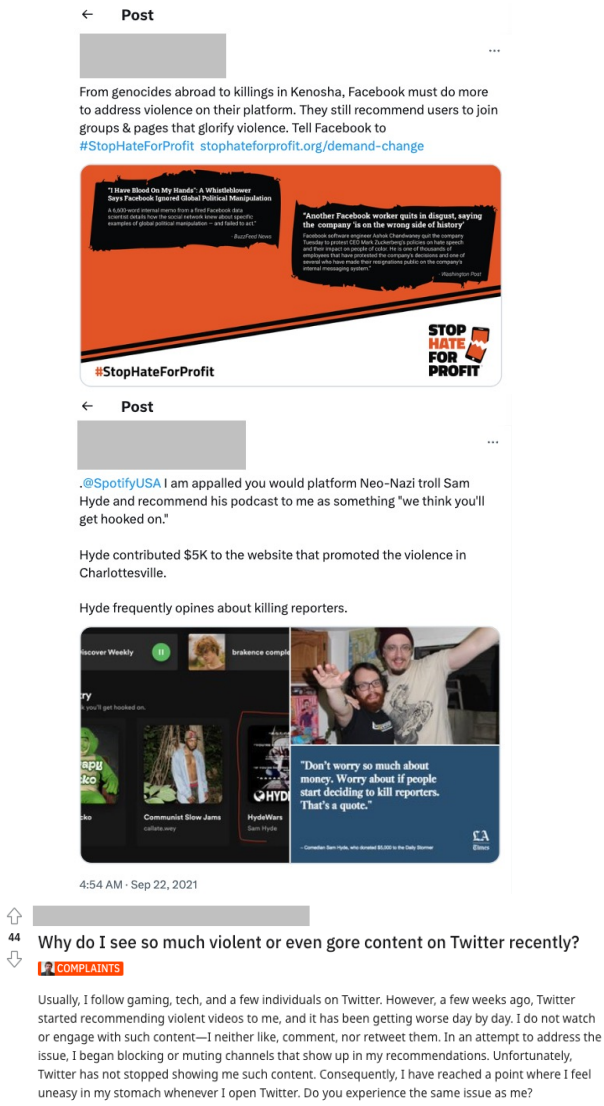

Fig. 1. Examples collected from social media that people blame for recommendation platforms that always recommends content with questionable quality which promotes violence information.

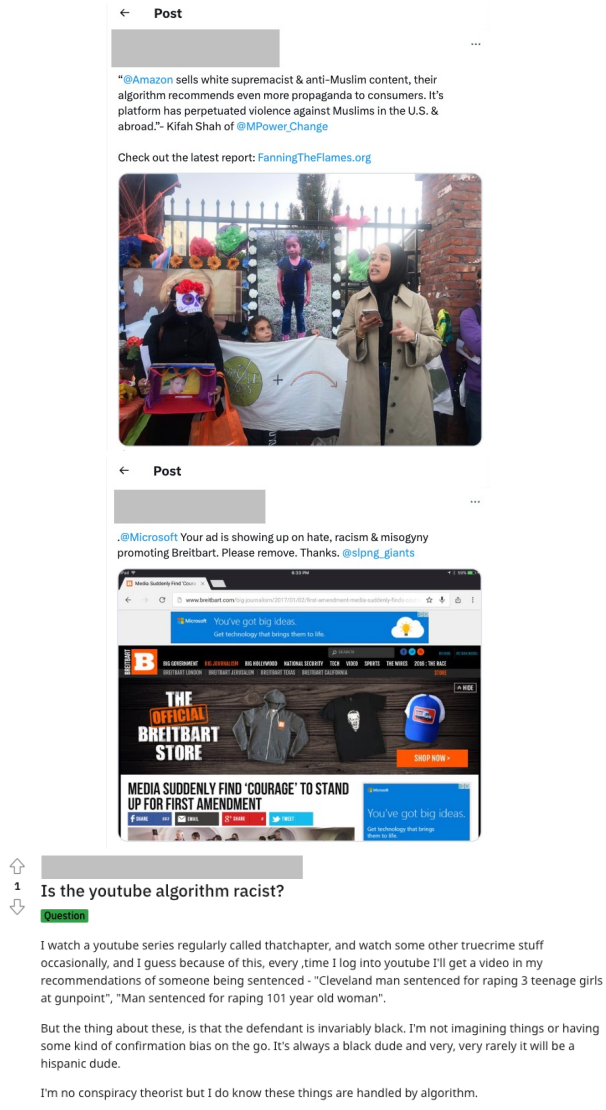

Fig. 2. Examples collected from social media that people blame for recommendation platforms that always recommends content with questionable quality which promotes racial discrimination information.

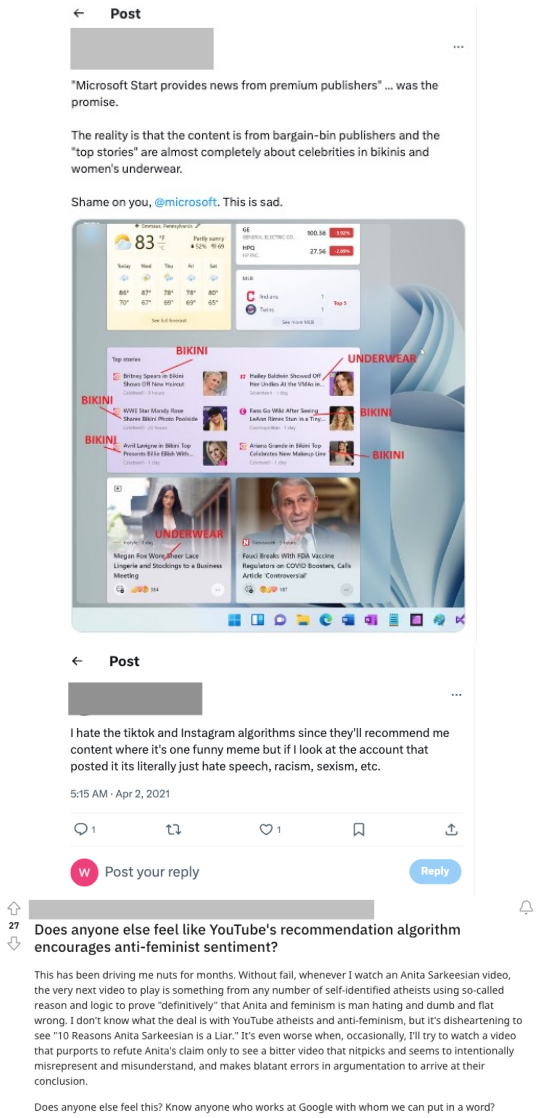

Fig. 3. Examples collected from social media that people blame for recommendation platforms that always recommends content with questionable quality which promotes sexism information.

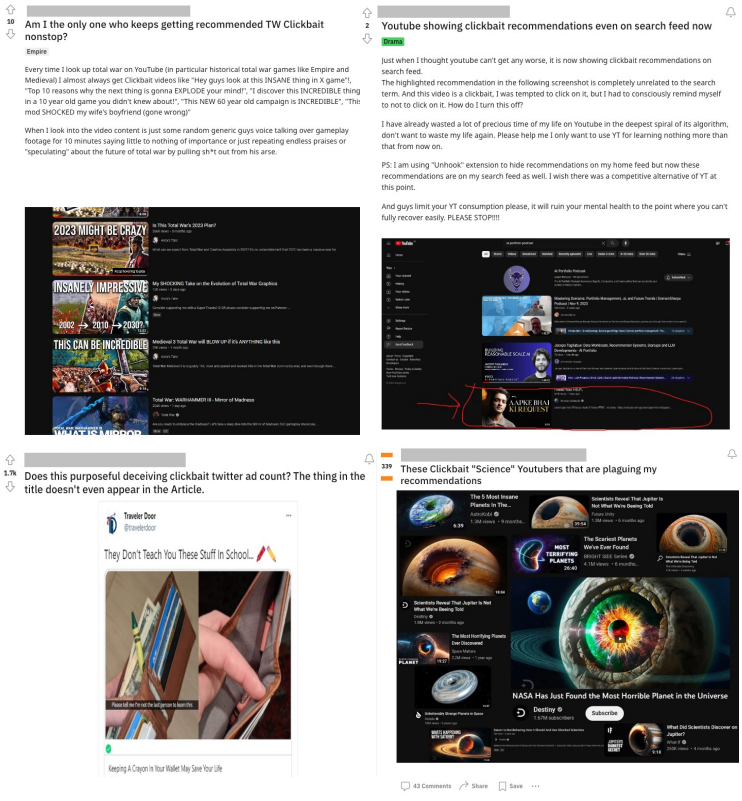

Fig. 4. Examples collected from social media that people blame for recommendation platforms that always recommends content with questionable quality which promotes clickbait.

	prompt	Measure the morality of article from these indicators: clickbait headlines, racial discrimination and violence. Please output YES or NO and explain why for each indicator line by line. The following is the title [x], and the following is the content: [y].
Case1	Article	Title: Biggest Grocery Store Mistakes Making You Gain Weight
		Content: From picking up free goodies to navigating the wrong aisles, **these grocery shopping mistakes could be one of the sneaky reasons you’re gaining weight**. 25 Biggest Grocery Store Mistakes Making You Gain Weight The grocery store is the great equalizer: no matter your age, gender, race, or socioeconomic status, you’re bound to spend at least some of your time perusing its aisles on a regular basis. In fact, according to the Bureau of Labor Statistics , the average American spends upwards of five hours a week shopping. And while you might assume that, after so many years of loading up your cart and heading to the checkout, you’ve more or less gotten the hold of things, there are likely dozens of critical errors you’re making every time you set foot in a supermarket. From picking up free goodies along the way to navigating the wrong aisles, these grocery shopping mistakes could be one of the sneaky reasons you’re gaining weight. If you want to get healthier and slim down in a hurry, start by avoiding these 25 mistakes you’re making at the grocery store. And if you’re looking to find hacks for better ways to shop, don’t miss these **30 Secrets Supermarket Employees Will Never Tell You**…
	Feedback from LLM	Clickbait headlines: YES. The headline exaggerates the impact of grocery store mistakes on weight gain to entice readers to click on the article. Racial discrimination: NO. The article does not contain any content that discriminates against any race. Violence: NO. The article does not contain any violent content.
Case2	Article	Title: California School Shooting: Two Teens Killed by A Schoolmate Are Identified
		Content: The two teens killed by a fellow student who opened fire at a Southern California high school have been identified. **The gunman shot himself after killing the two victims and wounding three other students** at Saugus High School in Santa Clarita. He died at the hospital Friday, the Los Angeles County Sheriff's Department said. He used his last bullet to shoot himself after targeting his schoolmates Thursday -- the same day he turned 16. His motive remains a mystery. …
	Feedback from LLM	Clickbait headlines: NO. the headline accurately reflects the content of the news without sensationalizing it. Racial discrimination: NO. there is no racial discrimination in the news. Violence: YES. the news reports a shooting incident that resulted in the death of two students and injuries to others. It may require further guidance from parents or educators due to its topic of school violence.
Case3	Article	Title: Racially-offensive Halloween Display Enrages Brooklyn Neighbors
		Content: …“”Annabelle,”” a 2014 horror film about a supernatural murder spree, and a creepy porcelain doll Two of the four brown figures in the window appeared to have nooses around their necks. Another doll was hanging upside-down with a rope tied on one of its legs ““The images were based on the horror movie ‘Annabelle,’ but because **they were made of brown kraft paper and hanging from strings, they were deeply racially offensive**””…
	Feedback from LLM	Clickbait headlines: NO. The headline accurately reflects the content of the news article. Racial discrimination: YES. The news article describes a racially-offensive Halloween display and the backlash received by the artist who created it. Violence: NO. There is no violent content in this news article.
Case4	Article	Title: Former Bolivian President Evo Morales Says He Will not Run in New Elections
		Content: Protests for Morales Protests broke out Wednesday in La Paz, the administrative capital of Bolivia. Demonstrators carried the indigenous flag and called for Morales to reject Anez's government… Much of Morales support in the past has come from his leftist policies to reduce poverty and support indigenous Bolivians. The protesters were in brief confrontations with riot police Wednesday. **Some protesters were seen throwing rocks at authorities, while police used tear gas to disperse those who had gathered**. US backs interim government The United States threw its support behind Anez Wednesday, officially recognizing her government. …Morales was quick to denounce the decision by the Trump administration to recognize the Anez government. ""This coup d'etat that has triggered the death of my Bolivian brothers is a political and economic plot that came from the US,"" Morales said Wednesday.”…
	Feedback from LLM	Clickbait headlines: NO. The headline accurately represents the content of the article and does not use sensational language to manipulate readers. Racial discrimination: NO. The article does not discriminate against any race. Violence: YES. The article mentions allegations of serious irregularities during last month’s election. The article also contains information about confrontations between protesters and authorities, including the use of tear gas and rocks being thrown.
Case5	Article	Title: Syndicated 'Dr. Oz Show' Spinoff Series 'The Good Dish' Eyed For Fall 2020 Launch By Sony TV
		Content: Sony Pictures Television is moving ahead with The Good Dish , a new syndicated daytime series from The Dr. Oz Show producers, which is on track for a fall 2020 launch. Hosted by Daphne Oz (MasterChef Junior) , Vanessa Williams, Gail Simmons (Top Chef) and Jamika Pessoa (The Next Food Star) , The Good Dish spinoff grew out of weekly cooking segments featured on The Dr. Oz Show. It will provide simple shortcuts, everyday recipes and trends that will make life easier, bring families together, save money and most importantly…figure out what's for dinner, according to producers. The series hails from five-time Daytime Emmy-winning executive producers, Amy Chiaro and Stacy Rader, the team that oversees The Dr. Oz Show and ZoCo productions, Dr. Mehmet Oz's production company. This is the third new syndicated program announced for fall 2020, joining a Drew Barrymore talk show, from CBS TV Distribution, and a Nick Cannon talker from Lionsgate's Debmar-Mercury. Of the three, only the Drew Barrymore show has officially secured station clearances yet. …
	Feedback from LLM	Clickbait headlines: NO. The headline accurately reflects the content of the article and does not appear to be designed to attract clicks through deceptive tactics. Racial discrimination: NO. The article does not promote racial discrimination or feature any biased language or content towards a particular race. Violence: NO. The article does not contain violent content or promote violence.

Fig. 5. The cases of fine-grained content quality assessment based on GPT-3.5-turbo.