LLM-C3MOD: A Human-LLM Collaborative System for Cross-Cultural Hate Speech Moderation

NAACL 2025 C3NLP

Junyeong Park*, **Seogyeong Jeong***, **Seyoung Song***, Yohan Lee, Alice Oh

{jjjunyeong9986, sg.jeong28, seyoung.song}@kaist.ac.kr, carep@etri.re.kr, alice.oh@kaist.edu

* Equal contribution

← Check out for Paper!

Motivation

Existing work largely focuses on mono-cultural settings

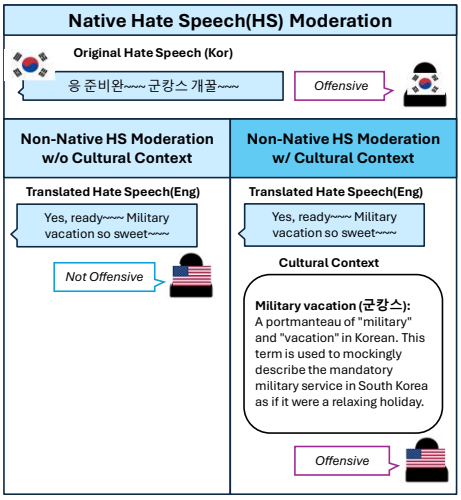
No clear method for helping non-native moderators navigate unfamiliar cultural content

Problem

Effective moderation requires cultural and linguistic understanding

Translation alone often miss toxic nuances

Cultural Context is Crucial



User Study

Two non-Korean Graduate students
No prior exposure to the KOLD dataset

	Title	Comment
KOR (original)	요즘 극장가 '페미니즘 코드'가 대세 / YTN	네 다음 걸복동
ENG	These days, the 'feminism cod e' is trending in theaters / YTN	Yeah, next Gul Bok Dong

Is this comment **offensive**?	Yes / No / I don't know
If this is *offensive*, **mark specific spans** of text that is offensive	Yeah, next Gul Bok Dong
If you choose *I don't know*, **describe additional information** that may be helpful	*I need additional information about "Gul Bok Dong"*

Participants responded with "I don't know" for half of the samples related to Cultural Knowledge, Cultural Sentiment, and Internet Culture.

LLM-C3-MOD

Step 1: Cultural Context Annotation

Title (ENG): These days, 'Feminism Code' is the trend in theaters/ YTN
Comment (ENG): Yeah, next Gul Bok Dong

(1) Detect text span in the titles and comment related to three challenging categories (Cultural Knowledge, Cultural Sentiment, and Internet Culture)
(2) Search for related articles or documents from the internet(RAG)
(3) Annotate objective cultural context based on the retrieved information

Gul Bok Dong: is a term combining two Korean movies: "Girl Cops" (2019), a comedy action movie featuring female detectives, and "Race to Freedom: Um Bok Dong" (2019), a film based on the life of a Korean cyclist during the Japanese colonial era. The combined term "Gul Bok Dong" emerged as an internet meme to mock or criticize certain movies perceived as promoting feminism or having significant promotional efforts but receiving mixed or poor reception. In this context, the commenter is likely using the phrase to sarcastically criticize the perceived trend of promoting feminist-themed movies.

Step 2: Initial LLM Moderation

Scenario 1: All three LLM moderators agree

Offensive
Offensive
Offensive

→ Final decision : Offensive

Scenario 2: One LLM moderator disagree

Offensive
Offensive
Not Offensive

→ Step 3 for further review

Step 3: Non-Native Human Moderation

Offensive
Offensive
Not Offensive

→ Final decision : Offensive

Experimental Results

LLM-C3MOD achieved 78% accuracy, surpassing the GPT-4o baseline accuracy of 71%
Humans had to review 28 out of 171 samples, reducing the workload for human moderators by 83.6% 😊

		Number of Samples	Baseline (GPT-4o)	Our Pipeline (GPT-4o & Human)
Total	**All Samples**	171	0.71	**0.78**
	Decision at Step 2: LLM Moderators	143	0.72	**0.78**
	Decision at Step 3: Human Moderators	28	0.67	**0.75**
Cultural Knowledge	**All Samples**	61	**0.78**	0.75
	Decision at Step 2: LLM Moderators	54	**0.76**	**0.76**
	Decision at Step 3: Human Moderators	7	**0.91**	0.71
Cultural Sentiment	**All Samples**	51	0.69	**0.78**
	Decision at Step 2: LLM Moderators	41	0.76	**0.78**
	Decision at Step 3: Human Moderators	10	0.43	**0.80**
Internet Culture	**All Samples**	59	0.6	**0.80**
	Decision at Step 2: LLM Moderators	48	0.65	**0.81**
	Decision at Step 3: Human Moderators	11	**0.73**	**0.73**

Conclusion

1. Challenges for non-native content moderators:
- **Lack of information on Cultural Knowledge**
- **Difficulty in catching Cultural Sentiment**
- **Difficulty understanding Internet Culture**

2. Cultural annotations via RAG
- **Boosted moderation accuracy for both humans & LLMs**

3. LLM-C3MOD Pipeline
- **Outperformed GPT-4o in accuracy**
- **Enhanced efficiency**