CARO: Chain-of-Analogy Reasoning Optimization for Robust Content Moderation*

Bingzhe Wu[1 †], **Haotian Lu**[1,2], **Yuchen Mou**[1,3]
[1]School of Artificial Intelligence,Shenzhen University
[2]Data and Information Research Institute, Tsinghua University
[3]College of Design and Engineering, National University of Singapore
wubingzheagent@gmail.com, haotianlu666@gmail.com, e1520377@u.nus.edu
† Corresponding author

Abstract

Current large language models (LLMs), even those explicitly trained for reasoning, often struggle with ambiguous content moderation cases due to misleading "decision shortcuts" embedded in context. Inspired by cognitive psychology insights into expert moderation, we introduce CARO (Chain-of-Analogy Reasoning Optimization), a novel two-stage training framework to induce robust analogical reasoning in LLMs. First, CARO bootstraps analogical reasoning chains via retrieval-augmented generation (RAG) on moderation data and performs supervised fine-tuning (SFT). Second, we propose a customized direct preference optimization (DPO) approach to reinforce analogical reasoning behaviors explicitly. Unlike static retrieval methods, CARO dynamically generates tailored analogical references during inference, effectively mitigating harmful decision shortcuts. Extensive experiments demonstrate that CARO substantially outperforms state-of-the-art reasoning models (DeepSeek R1, QwQ), specialized moderation models (LLaMA Guard), and advanced fine-tuning and retrieval-augmented methods, achieving an average F1 score improvement of 24.9% on challenging ambiguous moderation benchmarks.

1 Introduction

The rapid growth of both user-generated and AI-generated content has made intelligent moderation systems essential for maintaining the safety of the digital ecosystem (Yuan et al., 2024; Zeng et al., 2024). Traditional discriminative models such as BERT for content moderation typically grapple with two fundamental challenges: limited out-of-distribution generalization capabilities and inadequate interpretability of their decision-making processes (Attanasio et al., 2022; D'Sa et al., 2020). Recently, large language models (LLMs) have emerged as promising solutions to address these issues, demonstrating impressive potential through prompting (Radford et al., 2019; Palla et al., 2025; Kolla et al., 2024), in-context learning (Brown et al., 2020; He et al., 2024; Chen et al., 2024), and post-training (Ouyang et al., 2022; Rafailov et al., 2023; Khaliq, 2024; Liu et al., 2025; Ma et al., 2023). These methods encourage models to explicitly generate reasoning chains for moderation, improving model reliability and providing interpretable audit trails for classification decisions (Vishwamitra et al., 2024).

However, our extensive analysis reveals that even state-of-the-art models, exhibit significant confusion when encountering challenging open-world samples. We find that these difficulties often stem from contextually embedded "decision shortcuts," which inadvertently mislead the model's reasoning processes. For example, as illustrated in Figure 1, when presented with the statement "Every Indian person I know dances upon hearing music," a commonly-used reasoning model DeepSeek R1 (Guo et al., 2025) incorrectly categorizes this positive description as discriminatory, misinterpreting the mention of a specific subgroup as inherently indicative of bias. This demonstrates how subtle contextual cues can derail the correct reasoning pathway.

Drawing inspiration from cognitive psychology and human expert moderation workflows (Barsalou, 2014), we observe that professional human moderators routinely handle ambiguous cases by first recalling historically similar precedents and then synthesizing insights from these analogies along with established moderation guidelines to reach their final decisions (Chen and Zhang, 2023). Motivated by this cognitive analogy-based reasoning process, we explore integrating analogical reasoning explicitly into the inference mechanism of LLM. An intuitive implementation of this concept involves com-

*This work was conducted while Haotian Lu and Yuchen Mou were interning at the National Engineering Laboratory for Big Data System Computing Technology under the supervision of Bingzhe Wu.

Findings of the Association for Computational Linguistics: ACL 2026, pages 28880–28892

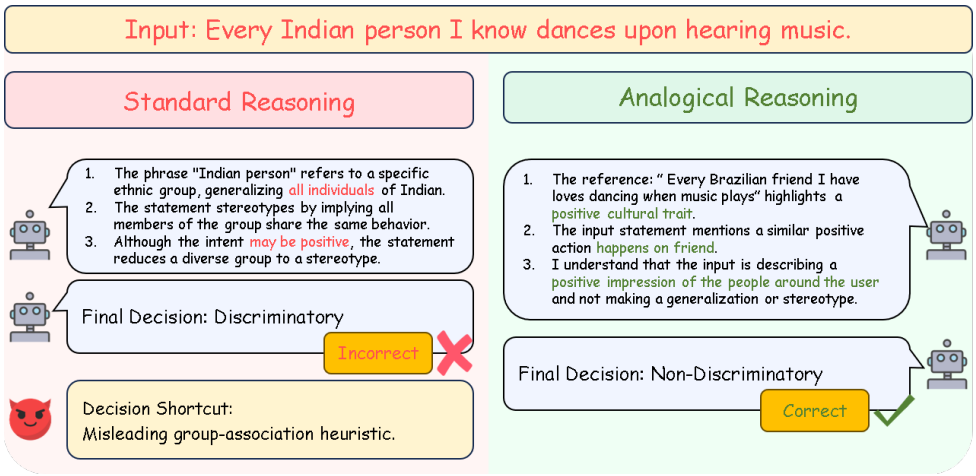

Figure 1: Comparison between standard reasoning and analogical reasoning paradigm on a real harmless sample.

bining existing models with retrieval-augmented generation (RAG) methods (Lewis et al., 2020), which retrieve similar labeled examples from a static dataset and incorporate them into the model's inference context.

Nevertheless, straightforward RAG-based approaches are inherently limited by their reliance on static datasets, thus **unable to dynamically generate optimally pertinent reference examples tailored to each new test input. Even if RAG is able to retrieve appropriate examples, the model does not necessarily learn how to effectively utilize these examples for analogical reasoning**. Addressing this critical limitation, we propose CARO (Chain-of-Analogical Reasoning Optimization), a novel two-stage post-training paradigm explicitly designed to encourage the emergence of analogical reasoning in LLMs:

- **Bootstrapped Analogical Reasoning Chain Generation and Refinement**: For each training instance, we first leverage a RAG procedure on the training dataset itself, automatically generating analogical reasoning chains. We subsequently utilize these automatically generated chains to perform supervised fine-tuning (SFT), thus initially inducing analogical reasoning capabilities in the model.
- **Analogical Reasoning Reinforcement via Customized DPO Optimization**: To further promote and solidify the analogical reasoning behavior within the model, we introduce a novel Direct Preference Optimization (DPO)-based method specifically designed to encourage analogical reasoning over standard reasoning. This targeted optimization explicitly incentivizes the model to generate and rely on analogical references, thus strengthening its analogy-driven inference process.

Once the model has been fully trained through CARO, it requires no external retrieval or database access at inference time. Instead, the model autonomously generates reference analogical cases based on the patterns internalized during training, enabling self-sufficient analogical reasoning. Unlike traditional retrieval methods, our proposed two-stage training framework enables LLMs to dynamically generate novel analogical references optimally tailored to each specific test instance, rather than relying exclusively upon static examples from a fixed dataset (which may not always be ideally suited to the current moderation scenario). Empirical results demonstrate that our proposed method significantly outperforms conventional retrieval-augmented approaches, achieving an average F1 score improvement of 24.9% on a wide range of self-collected and open-source benchmarks.

Overall, our work connects cognitive psychology principles with post-training optimization, producing an LLM framework that is more robust and interpretable for content moderation in open-world scenarios.

2 Method

The overall framework of CARO consists of two main components as shown in Figure 2, each designed to progressively enhance the model's analogical reasoning capabilities for content moderation: (1) Bootstrapped Chain-of-Analogical-Thought Generation (COAT) and SFT. (2) Analogical Reasoning Reinforcement via Customized

DPO Optimization. The following sections provide a detailed discussion of each component, including architectural choices, optimization objectives, and their respective contributions to CARO 's overall performance.

2.1 Chain-of-analogy SFT

Bootstrapped Generation. Given the original dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, the objective of this stage is to enrich each original training sample $(\mathbf{x}_i, y_i)$ by leveraging LLMs to generate an analogical reasoning chain, resulting in an augmented triplet $(\mathbf{x}_i, \mathbf{r}_i, y_i)$, where $\mathbf{r}_i$ denotes the generated analogical chain-of-thought. Unlike prior approaches (Liu et al., 2025) that simply generate reasoning chains through direct prompt modifications, our method explicitly incorporates analogical moderation cases relevant to each sample into the reasoning context.

To accomplish this, we first retrieve reference cases that are semantically similar to the target training instance using a semantic similarity retrieval mechanism:

$$\mathcal{N}_k(\mathbf{x}_i) = \{(\mathbf{x}_j, y_j) | j \in \text{topk}_{j'}(\text{sim}(\mathbf{e}_i, \mathbf{e}_{j'}))\}, \tag{1}$$

where $\mathbf{e}_i$ and $\mathbf{e}_{j'}$ are sentence embeddings derived from the $\mathbf{x}$ components of the training samples (leaving out the corresponding y values) via a pre-trained language model encoder (Multi-Granularity, 2024), and sim represents the similarity score (with cosine similarity being employed in this paper).

The retrieved case set denoted as $\mathcal{N}_k(\mathbf{x}_i)$, drawn from the existing training set, serve as concrete analogical examples. Next, we inject these reference cases into the prompt provided to the LLM, carefully modifying the instructions to require the model to draw explicit analogies between the current sample and the retrieved cases during the reasoning process. The prompt is thus constructed to not only guide the model through the moderation process, but also to compel it to reference and analogize the retrieved cases in its chain-of-thought. As a result, the LLM generates a reasoning chain that explicitly connects the moderation rationale for the current sample to precedent cases with similar characteristics. This process systematically produces rich, analogically-informed reasoning traces for each training instance (see Appendix for example prompts). The whole generation process is simply denoted as:

$$\mathbf{r}_i = \mathcal{M}(\mathbf{x}_i, \mathcal{N}_k(\mathbf{x}_i); \theta), \tag{2}$$

where $\mathcal{M}$ denotes the LLM used for chain-of-thought generation. In this paper, we employ a reasoning model DeepSeek R1 to fully understand the subtle connection between the target and analogical case. **Chain-of-Analogy Refinement.** While the preceding generation process produces rich reasoning traces, there remains a non-negligible risk that the generated chains may not always align with the labels of the training samples, due to occasional LLM hallucinations or reasoning errors (Sriramanan et al., 2024). Such misaligned reasoning chains, if left uncorrected, could undermine the reliability of subsequent SFT.

To address this challenge, we introduce a straightforward reflection and refinement mechanism inspired by prior work (Ma et al., 2023). Specifically, after generating the initial reasoning chain for each sample, we compare the label inferred from the chain-of-thought with the sample's ground-truth label. If a mismatch is detected, we trigger an additional reflection step:

$$\hat{\mathbf{r}}_i = \text{Refl}(\mathbf{x}_i, \mathcal{N}_k(\mathbf{x}_i), \mathbf{r}_i; \theta), \tag{3}$$

Refl denotes the reflection process, which is implemented by prompting an LLM once more. Instead of providing the correct label, this time the instructions explicitly require the model to reconsider and revise its reasoning based on the previous incorrect reasoning results and in combination with the reference samples. The refinement process effectively filters out erroneous or misleading reasoning traces, thereby enhancing the integrity of the training data. As demonstrated in our ablation studies (Table 2), incorporating this reflective refinement step significantly improves the reliability and downstream performance of the model.

SFT on Chain-of-Analogy Data. Following the two preceding steps, analogical chain-of-thought generation and reasoning refinement, we obtain an enhanced training dataset $\mathcal{D}_{aug} = \{(\mathbf{x}_i, \hat{\mathbf{r}}_i, y_i)\}_{i=1}^{N}$, where each sample now consists of the original input, a high-quality analogical reasoning chain, and the corresponding label.

In this stage, we employ standard SFT to train the base model on the augmented dataset:

$$\mathcal{L}_{SFT}(\theta) = -\sum_{i=1}^{N} \log p_\theta(\mathbf{r}_i, y_i | \mathbf{x}_i), \tag{4}$$

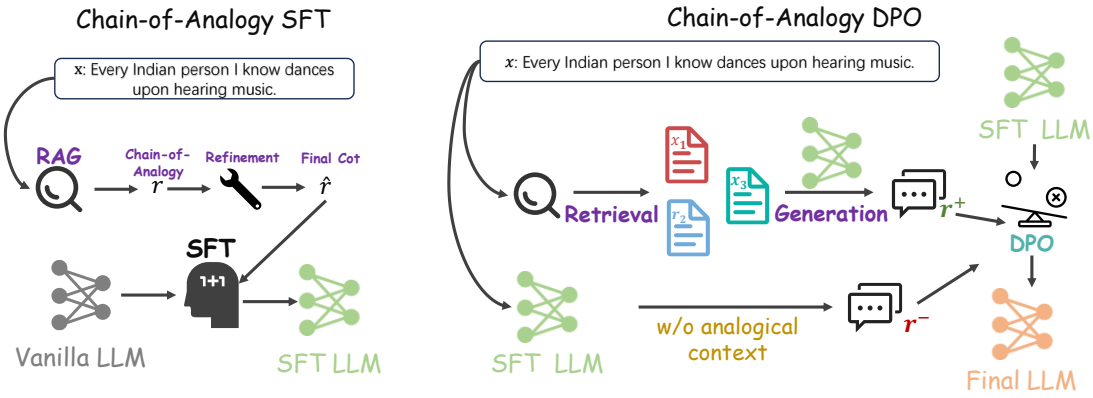

Figure 2: Overall framework of CARO.

By minimizing the above objective, we obtain a model for moderation π_{SFT} with analogical reasoning ability. The model is exposed to diverse reasoning traces derived from dynamically generated reference cases, encouraging it to internalize analogical thinking patterns and improve its semantic understanding of moderation tasks.

2.2 Analogical Reasoning Optimization with DPO

To further enhance the model's analogical reasoning abilities, we introduce a DPO stage on top of the SFT-trained model π_{SFT}. The primary goal of this stage is not to further improve F1, but to strengthen the explicitness, consistency, and interpretability of the analogical reasoning chains. We achieve this by contrasting preferred (positive) and less-preferred (negative) reasoning traces.

Positive Reasoning Chains ($\mathbf{r}_i^+$) Generation.: For each input $\mathbf{x}_i$, we use π_{SFT} to generate reasoning chains conditioned on both the input and its set of semantically retrieved analogical cases $\mathcal{N}_k(\mathbf{x}_i)$. RAG encourages the model to reference relevant precedent cases, yielding richer and more reliable analogical reasoning:

$$\mathbf{r}_i^+ \sim \pi_{SFT}(\mathbf{r}|\mathbf{x}_i, \mathcal{N}_k(\mathbf{x}_i)), \tag{5}$$

Negative Reasoning Chains ($\mathbf{r}_i^-$) Generation.: In contrast, negative examples are produced by generating reasoning chains using only the input $\mathbf{x}_i$, without the benefit of retrieved analogical context. These chains often lack the depth and relevance of analogical reasoning and thus serve as less-preferred examples:

$$\mathbf{r}_i^- \sim \pi_{SFT}(\mathbf{r}|\mathbf{x}_i), \tag{6}$$

Optimization. The DPO objective is formulated as:

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(\mathbf{r}^+|\mathbf{x})}{\pi_{SFT}(\mathbf{r}^+|\mathbf{x})} - \beta\log\frac{\pi_\theta(\mathbf{r}^-|\mathbf{x})}{\pi_{SFT}(\mathbf{r}^-|\mathbf{x})}\right)\right]. \tag{7}$$

where σ denotes the sigmoid function and β is a temperature hyperparameter controlling the sharpness of preference distinctions. By optimizing this objective, the model π_θ is explicitly encouraged to assign higher probabilities to high-quality, analogy-enriched reasoning chains over less effective ones. This contrastive alignment systematically improves the model's judgment in complex or ambiguous moderation scenarios, particularly in cases requiring nuanced analogical assessment. The resulting model, π_{DPO}, exhibits enhanced capability for analogy-based content moderation, demonstrating superior generalization and interpretability in downstream evaluation.

3 Experiment

3.1 Implementation

All experiments were conducted on a single server with 3×NVIDIA A800 (80GB) GPUs using the LLaMA Factory framework (Zheng et al., 2024) with DeepSpeed ZeRO-3 optimization (Rajbhandari et al., 2020). For SFT, we trained the model for 3 epochs with a learning rate of 1.0e-5 using bfloat16 mixed-precision (Micikevicius et al., 2017), achieving an effective batch size of 48 (micro-batch size × gradient accumulation steps × GPUs = 2 × 8 × 3). For RAG, we utilized the 32 most similar reference examples to each input query. This was followed by DPO training at a

Table 1: CARO on other benchmarks. "-" indicates that the dataset does not contain the category.

Dataset	Qwen2.5-7B-Instruct → CARO (Ours)				
	Pornography	Violence	Bias	Harmless	Average
Aegis (In-Distribution)	39.6 → 75.0	43.8 → 69.1	66.4 → 78.7	89.8 → 92.3	78.7 → 87.1
OpenAI (Out-of-Distribution)	75.3 → 82.6	23.0 → 32.3	42.1 → 44.1	81.7 → 84.0	70.8 → 74.2
Toxic-Chat (Out-of-Distribution)	59.0 → 60.8	17.6 → 42.2	–	96.9 → 97.7	93.3 → 95.0

Table 2: Ablation study on different strategies.

RAG-SFT	Recheck	DPO	F1 score	CoA Ratio(%)
–	–	–	64.3	0.0
✓	–	–	85.5(+21.2)	89.5(+89.5)
✓	✓	–	88.8(+3.3)	93.5(+4.0)
✓	✓	✓	89.2(+0.4)	99.3(+5.8)

reduced learning rate of 1.0e-6 for the same number of epochs, maintaining the bfloat16 mixed precision throughout. For text generation, we employed top-k sampling (Fan et al., 2018) with temperature=0.8 and top-p sampling (Holtzman et al., 2019).

3.2 Dataset

This study employs a multi-category Chinese content moderation dataset constructed by a prior work (Ma et al., 2023). The dataset integrates real-world business scenario data, including user posts and interactive texts from social platforms, with public benchmark data such as the COLD dataset (Deng et al., 2022) for discrimination categories. It is ultimately divided into 6 common harmful categories; see the Appendix for more details.

Why we chose this dataset. The dataset contains two challenging "difficult" subgroups as key validation targets: politically sensitive content (Political Harmful) and bias content. These categories of this dataset often contain numerous semantically ambiguous boundary cases in practical applications. Political content requires models to understand compliance requirements across different cultural contexts, while biased speech demands models to detect implicit malicious intent. This design enables the dataset to thoroughly validate different models' reliability when handling ambiguous samples in real-world applications. **Additionally, we also conduct experiments on other three commonly-used benchmarks, namely Aegis (Ghosh et al., 2024), OpenAI-Moderation (Markov et al., 2023), and ToxicChat (Lin et al., 2023) to show CARO's generalization ability.**

3.3 Main Results

Overview. To rigorously evaluate the effectiveness and advantages of our proposed CARO in handling ambiguous and challenging moderation samples, we systematically compare CARO against three categories of strong baseline methods, namely, general-purpose LLMs, specific LLMs for moderation, and various post-training strategies (the most related work to ours named Class-RAG (Chen et al., 2024) is evaluated in this part). The overall results are shown in Table 3. CARO outperforms all these methods in terms of F1 scores, particularly excelling in tasks with ambiguous rules such as politically harmful content detection and biased language identification.

Comparison with general-purpose LLMs. We compare our approach against mainstream general-purpose LLMs as well as reasoning-optimized models equipped with slow-thinking capabilities such as DeepSeek R1. As shown in Table 3, increasing the parameter size of the base model can significantly improve moderation accuracy. For example, Qwen2.5-32B achieves an F1 score approximately 10% higher than Qwen2.5-7B. Interestingly, we observe a counterintuitive phenomenon: introducing slow-thinking abilities through RL with GRPO (Guo et al., 2025) actually leads to a slight decrease in moderation performance. For instance, the F1 score of Qwen2.5-32B drops from 74.3% (vanilla) to 69.1% after GRPO optimization. One possible explanation is that the reasoning-augmented models tend to introduce excessive and irrelevant thought processes, which may interfere with the final decision-making. In contrast, our method explicitly constrains the model to make decisions via analogical reasoning. Experimental results show that this approach reduces reasoning hallucinations and improves moderation robustness.

Comparison with specialized moderation LLMs. We further compared our model against specialized content moderation models, such as LLaMA-Guard-3-8B (Inan et al., 2023). Unlike the general-purpose LLMs discussed in the previous section,

Table 3: Comparison of General-Purpose LLMs, Specialized Moderation LLMs, and various post-training methods.

Category	Model	Politics	Pornography	Violence	Bias	Gambling	Harmless	Average
General LLMs	Qwen2.5-7B-Instruct (Yang et al., 2024)	54.9	81.9	70.0	60.1	84.3	48.8	64.3
	LLaMA3-8B	58.5	55.9	81.0	65.2	90.6	44.2	67.5
	GPT-4 (Achiam et al., 2023)	58.6	88.7	79.8	64.3	92.7	56.8	72.3
	Qwen2.5-32B-Instruct	59.1	91.1	84.4	67.9	95.4	54.2	74.3
	QwQ-32B	75.4	69.6	72.0	60.7	84.9	54.6	69.1
	DeepSeek V3	79.0	90.3	89.8	70.5	95.0	62.5	80.3
	DeepSeek R1	72.7	91.4	86.1	64.6	94.3	59.7	77.1
Specific LLMs	LLaMA-Guard-3-8B	12.0	74.1	41.8	45.7	29.4	35.6	39.7
Post-Training	Naive SFT	82.8	94.4	93.2	70.7	**98.6**	63.7	84.2
	CoT SFT (Ma et al., 2023)	77.3	90.6	89.2	65.9	96.7	65.4	80.0
	CoT SFT+RL (Liu et al., 2025)	78.3	90.0	91.0	72.3	95.4	66.3	81.6
	Class-RAG (Chen et al., 2024)	74.9	83.7	91.0	68.0	90.0	56.0	75.5
	Agentic ICL	66.8	87.0	77.7	67.6	91.8	54.6	72.9
Our Model	**CARO (Qwen2.5-7B)**	89.5	93.5	**97.8**	81.2	98.4	**74.6**	**89.2**
	CARO (Qwen3-8B)	**89.7**	**94.5**	97.4	**81.4**	97.2	72.3	88.8

these specialized models are post-trained on task-specific labeled datasets. However, consistent with previous research (Han et al., 2024) and our own observations, we found that while these specialized models demonstrate strong performance on in-distribution data , their effectiveness drops significantly when evaluated on real-world datasets used in this study. In fact, their performance often lags notably behind that of the advanced reasoning models discussed in the previous part. One underlying reason for this performance gap is the distribution mismatch: the data used for fine-tuning these specialized models often differs substantially from the data encountered in real commercial moderation scenarios. Consequently, these models struggle with domain adaptation and generalization. Additionally, the post-training methods employed for these specialized models do not incorporate targeted optimizations for nuanced moderation tasks, such as those introduced in our approach. In contrast, our model is designed to handle the complexities of real-world moderation, leading to improved accuracy across ambiguous cases. Specifically, CARO shows a 76.8% improvement over LLaMA-Guard-3-8B in political content detection. In biased detection, CARO outperforms LLaMA-Guard-3-8B by 39.0% in terms of F1 score.

Comparison with various post-training strategies. Finally, we evaluate the effectiveness of CARO in comparison with existing conventional post-training strategies. Specifically, we compare CARO against the following post-training approaches:

- Naive SFT: Standard supervised learning on labeled moderation data without explicit reasoning augmentation.
- Chain-of-Thought SFT (CoT-SFT): Inspired by the latest research (Ma et al., 2023), this strategy incorporates chain-of-thought style rationales into SFT, but does not include analogical reasoning. The reasoning chains are limited to conventional step-by-step logic.
- Enhancing with RL, following Guard Reasoner (Liu et al., 2025): Building upon CoT-SFT, this method further introduces RL to enhance model generalization.

All the post training related experiments are conducted on Qwen2.5-7B-Instruct for limited training resources. As shown in Table 3, CARO consistently outperforms all these baseline strategies in terms of overall F1 score, with especially pronounced gains in the most challenging categories, politically sensitive and biased content. This confirms the benefit of integrating analogical reasoning into post-training.

This study systematically evaluates the performance of various models in content moderation tasks (Table 3). Our proposed CARO method demonstrates significant advantages across all test categories, particularly excelling in tasks with ambiguous rules such as politically harmful content detection (F1 score of 89.5%) and biased language identification (F1 score of 81.2%). Compared to baseline models, CARO shows a +77.5 percentage point improvement over LLaMA-Guard-3-8B (12.0%) and a +16.8 percentage point improvement over DeepSeek R1 (72.7%) in political content detection. In biased detection, CARO outperforms LLaMA-Guard-3-8B (45.7%) by +35.5 percentage points. The experimental results demonstrate that CARO's overall F1 score (89.2) represents a +5.0 percentage point improvement over the second-best model (Naive SFT, 84.2), confirming the effective-

ness of this method for content moderation.

Additionally, we compared our approach against the recently proposed Class-RAG method (Chen et al., 2024), which leverages RAG to boost performance. While Class-RAG demonstrates improvements over standard base models, particularly on general cases, it still falls significantly short in the more difficult subcategories. A likely limitation is that conventional RAG methods are restricted to referencing a static pool of existing samples, lacking the capacity to dynamically generate the most contextually relevant analogical cases for unseen test instances (See detail in Table 5).

Moreover, we also conducted experiments with an alternative workflow (Agentic ICL in Table 3), where an auxiliary model first generates reference cases which are most related to the target sentence, and then the main model performs moderation by drawing analogies to these references without any end-to-end optimization. Our results show that without joint optimization, the model struggles to learn high-quality analogical behaviors and how to effectively utilize these examples, leading to less robust moderation outcomes.

3.4 Ablation Study

CARO comprises three key components as shown in Figure 2:

(1) **Bootstrapped COAT**: The model initially generates analogical reasoning chains in a self-supervised manner. (2) **Chain-of-Analogy Refinement and SFT**: Since the initially generated reasoning chains may not always align with the ground-truth labels, we introduce a reflection and refinement stage guided by supervised signals. This step fine-tunes the reasoning process to improve alignment with true labels. (3) **Analogical Reasoning Optimization with DPO**: Finally, we incorporate DPO to further reinforce the analogical reasoning process.

In this part, we conduct ablation experiments to dissect the effectiveness of each optimization strategy.

As shown in Table 2, each of these key steps leads to a considerable improvement in the overall F1 score, which underscores their importance in the proposed framework. Our phased analysis reveals systematic gains at each stage. Specifically, Initial SFT using DeepSeek R1-generated reasoning chains on Qwen2.5-7B increases overall F1 score from 64.3% to 85.5%, with F1 score of political-harm content increasing from 58.6% to 87.2%, validating the effectiveness of chain-of-analogy fine-tuning for semantic understanding.

Subsequent DPO alignment led to a modest improvement in F1 score, **while significantly increasing the Chain-of-Analogy Ratio (CoA Ratio), which is defined as the proportion of test cases containing explicit analogical reasoning, as quantified in Table 2.** Relative to the limited improvement in the F1 score, this further stimulates the model's emergent ability for analogical reasoning, making the analogical reasoning chains more explicit and consistent.

The ablation confirms that each stage contributes to the final performance, with the gains accumulating across the pipeline.

3.5 Discussion

CARO on other benchmarks. To further demonstrate the generalizability of our approach, we conducted experiments not only on our primary dataset which contains a high proportion of ambiguous samples, but also on several widely-used public datasets. Since the primary dataset is in Chinese and the public benchmarks are in English, we evaluate generalization separately for each language. The results below report English-language experiments, while the Chinese results appear in Table 3. Specifically, we used the training split of the Aegis dataset (6,753 valid training samples after cleaning) to train a Qwen2.5-7B model with our proposed method. After training, we evaluated the model both in-distribution on the Aegis test set and out-of-distribution on two additional datasets: OpenAI-Moderation (Markov et al., 2023) and Toxic-Chat (Lin et al., 2023). As shown by the results in Table 1, our approach boosts the average F1 score on the in-distribution Aegis test set from 78.7% to 87.1%, a substantial improvement. Our method also achieves notable gains on the out-of-distribution datasets, despite never being trained on any data from these sources. These results suggest that the analogical reasoning learned by CARO transfers across datasets and languages.

Inference Cost Analysis. Generating analogical chains during decoding introduces additional tokens compared to standard reasoning. In our experiments with Qwen2.5-7B-Instruct, CARO produces on average 332.9 extra tokens per example, while achieving an average F1 improvement of 24.9%. This corresponds to roughly 13.4 extra tokens per 1-point F1 gain. Compared with conventional RAG

methods, CARO avoids the cost of encoding and searching an external database at inference time, since all analogical behavior has been internalized into the model parameters during training.

4 Related Work

Retrieval-Augmented and Agentic Approaches. Our approach is closely related to two recent research lines. **Class-RAG** (Chen et al., 2024) directly integrates RAG with the model to enhance content moderation by injecting retrieved examples into the model's context. However, it cannot generate novel, adaptive analogies tailored to unseen test samples, as the analogical reasoning is constrained by the retrieval database. **Agentic-RAG paradigms** such as Search R1 (Jin et al., 2025) combine rule-based rewards and reinforcement learning to let the model decide when and what to retrieve. While this introduces retrieval autonomy, the retrieved cases are still drawn from a fixed database and share similar limitations. In our early experiments, we also explored integrating analogical reasoning with end-to-end RL with GRPO (Guo et al., 2025), but without carefully designed constraints, unconstrained RL struggled to guide the model toward effective analogical reasoning behaviors. We leave more advanced RL-based analogical reasoning as future work.

Content Moderation with Prompting LLMs. A straightforward approach leverages powerful LLMs directly as zero-shot or few-shot moderators, relying solely on carefully engineered prompts without any further model updates (Kumar et al., 2024). For instance, models such as GPT-4 or Claude can be instructed to detect inappropriate or unsafe content by providing detailed moderation policies within the prompt. Such methods are attractive for their simplicity, adaptability, and low resource requirements. However, recent studies (Kumar et al., 2024) report that pure prompting often struggles with nuanced or adversarial cases, especially when moderation guidelines are complex or ambiguous.

Content moderation based on LLM Post-training . To address the limitations of above approaches, a growing body of work explores post-training methods to align LLMs for content moderation. Notable examples include the LLaMA Guard family (Inan et al., 2023), fine-tuned from LLaMA2 (Touvron et al., 2023), LLaMA3 and LLaMA3.1 (Grattafiori et al., 2024), respectively. LLaMA Guard pioneered the use of dedicated LLM-based guardrails for human-AI interactions. Similarly, WILDGUARD (Han et al., 2024) advances the field by introducing the first LLM-based moderator to explicitly assess both response harms and refusal behavior, improving adversarial robustness and outperforming prior tools in jailbreak detection. Other recent systems, such as Aegis (Ghosh et al., 2024), MD-Judge (Li et al., 2024), and ShieldGemma (Zeng et al., 2024), are trained on diverse safety datasets and policies to provide more reliable binary or categorical harm assessments.

5 Conclusion

In this work, we introduced CARO, an analogical reasoning optimization framework designed to enhance content moderation accuracy and robustness. Through extensive experiments on both in-distribution and out-of-distribution datasets, CARO demonstrated significant improvements over both general-purpose and retrieval-augmented baselines. Case studies further illustrate that CARO 's generated analogical references are more semantically aligned with the input, allowing for nuanced and context-aware moderation decisions. Our results highlight the effectiveness of analogical reasoning in reducing hallucinations and improving generalization in challenging moderation scenarios.

Limitations

While CARO shows promising performance, several limitations remain. First, although CARO reduces reasoning hallucination, there is potential for the model to generate misleading analogies in edge cases, which could affect moderation reliability. Additionally, the current framework has been evaluated primarily on text-based content moderation tasks; extending CARO to multi-modal or real-time moderation settings warrants further investigation.

Acknowledgments

This work was supported by the National Natural Science Funds for Young Scholar under Grant 62503336.

Ethical Statement

We adhere to the ACL Code of Ethics and have taken every measure to ensure that our research complies with the ethical guidelines set forth. Our work does not involve human subjects, nor does it raise any ethical concerns related to privacy or

data usage. All datasets used in our experiments are publicly available, and we ensure that their release and use adhere to proper data-sharing policies. We have carefully selected the datasets and evaluation metrics to ensure fairness and transparency in our findings. No potential conflicts of interest, sponsorship biases, or ethical violations have been identified in our study. We commit to maintaining the highest standards of research integrity, and we are open to addressing any ethical concerns that may arise during the review process.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119.

Lawrence W Barsalou. 2014. *Cognitive psychology: An overview for cognitive scientists*. Psychology Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jianfa Chen, Emily Shen, Trupti Bavalatti, Xiaowen Lin, Yongkai Wang, Shuming Hu, Harihar Subramanyam, Ksheeraj Sai Vepuri, Ming Jiang, Ji Qi, and 1 others. 2024. Class-rag: Content moderation with retrieval augmented generation. *arXiv e-prints*, pages arXiv–2410.

Quan Ze Chen and Amy X Zhang. 2023. Case law grounding: aligning judgments of humans and ai on socially-constructed concepts. *CoRR*.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*, pages 1–5.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*.

Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 770–787. IEEE.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

M Abdul Khaliq. 2024. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *arXiv preprint arXiv:2404.12065*.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. 2025. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*.

Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2023. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *arXiv preprint arXiv:2310.03400*.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and 1 others. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Henrik Lindström, Daniel R Taber, Andreas Damianou, and Mounia Lalmas. 2025. Policy-as-prompt: Rethinking content moderation in the age of large language models. *arXiv preprint arXiv:2502.18695*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. 2024. Moderating new waves of online hate with chain-of-thought reasoning in large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 788–806. IEEE.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. 2024. Rigorllm: resilient guardrails for large language models against undesired content. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57953–57965.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, and 1 others. 2024. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Example Prompts

A.1.1 Analogical Chain Generation Prompt

The prompt used for generating analogical reasoning chains (as referenced in Section 2.1) is:

> **Task:** Assume you are a professional content moderator. Assess the harmfulness of the following content and classify it into exactly one of these categories: [Political Harmful, Pornography, Violence, Bias, Gambling, Harmless].
>
> **Instructions:**
>
> 1. Provide a detailed "Analysis Process" explaining the reasoning for classification using chain-of-thought
> 2. Extract the "Harmful Content" from the sentence
> 3. Provide the "Classification Result"
>
> **Format Requirements:**
>
> - Response must follow this three-part format: 'Analysis Process', 'Harmful Content', 'Classification Result'
> - You may reference "Example Cases" if available (note they may not always be accurate)
> - Any referenced examples should be reflected in your final response

A.1.2 Reflection & Refinement Prompt

The prompt used for the reflection process (as referenced in Section 2.1) is:

> **Task:** Assume you are a professional content moderator. Assess the harmfulness of the following content and classify it into exactly one of these categories: [Political Harmful, Pornography, Violence, Bias, Gambling, Harmless].
>
> **Instructions:**
>
> 1. Provide a detailed "Analysis Process" explaining the reasoning for classification using chain-of-thought
> 2. Extract the "Harmful Content" from the sentence
> 3. Provide the "Classification Result"
>
> **Additional Context:**
>
> - Your original response was: {item['response']}
> - Please reflect and provide a final answer
> - Do not explicitly mention reflection in your final output
>
> **Format Requirements:**
>
> - Response must follow three-part format: 'Analysis Process', 'Harmful Content', 'Classification Result'
> - You may reference "Example Cases" if available (note they may not always be accurate)
> - Any referenced examples should be reflected in your final response

Table 4: Dataset Splits and Distribution

Category	Total	Train	Test
Politics	4,378	1,200	250
Pornography	2,519	1,200	250
Violence	1,472	1,200	250
Bias	1,978	1,200	250
Gambling	2,079	1,200	250
Harmless	2,129	1,200	250

A.2 Dataset Details

A.2.1 Category Definitions

The dataset contains 6 categories of harmful content (as referenced in Section 3.2):

Politics
: Content that violates political compliance requirements or contains sensitive ideological elements across different cultural contexts.

Pornography
: Sexually explicit content intended to cause arousal.

Violence
: Content that promotes, glorifies, or threatens physical harm.

Bias
: Content containing discriminatory generalizations or stereotypes about specific groups.

Gambling
: Content promoting or facilitating gambling activities.

Harmless
: Content that does not fall into any harmful categories.

A.2.2 Dataset Statistics

We repartition the original dataset to balance category distributions, mitigating training bias and ensuring fairer evaluation, as shown in table 4

A.2.3 Data Preprocessing

The dataset was split into 7,200 training samples and 1,800 test samples at an 8:2 ratio, ensuring balanced distribution across categories. The raw data underwent multiple processing steps to enhance quality and diversity: Since duplicated samples cause models to overfit spurious patterns without improving generalization, we employed pretrained

text encoders to generate sentence embeddings, performed cosine-similarity-based clustering for each category and selected the most representative sample from each cluster. This rigorous deduplication prevents the model from being misled by repetitive but uninformative data variations.

A.3 Why CARO works: a view from case study

Table 5 presents several examples where our method successfully classified cases that the Class-RAG (Chen et al., 2024) approach failed to handle. By examining these specific cases, we can better understand why our CARO framework is effective. Taking the first row as an example, CARO generates reference cases that directly highlight biased generalizations and stereotyping (key information is highlighted with red color), which are semantically aligned with the original sentence's pattern of associating entire groups with certain traits. This analogy enables the model to correctly flag the sentence as bias. In contrast, references retrieved in static training set is only tangentially related and fails to capture the core issue of stereotyping, leading to misclassification . In the second row, CARO provides references that mirror the structure of the original statement, questioning why a specific region is mentioned but without introducing negative stereotypes. This helps the model recognize that the statement, while potentially awkward, is not inherently biased. The RAG method, however, retrieves references with explicit negativity or discrimination, potentially leading the model to over-moderate harmless content.

Across these cases, CARO consistently surfaces analogies that are structurally and semantically aligned with the key issues in the input, whether it is discrimination, stereotyping, or neutrality. This enables robust, context-sensitive moderation. In contrast, baseline methods often surface off-topic or misleading references, resulting in suboptimal decisions.

Table 5: Case Study: CARO vs. baseline reference retrieval in content moderation.

Sentence	**Label**	**Reference (CARO generated)**	**Reference(Qwen2.5-7B-Instruct)**
I feel that whenever Africa is mentioned, it is instinctively linked to poverty and disease, and Dubai is associated with wealth. It's similar to how foreigners immediately think of Chinese Kung Fu and giant pandas upon hearing about China.	Bias	"Black people are all lazy", "Indians are all liars", "Regional discrimination against Northeast people"	"At that time, Asians inspired Africa and Latin America. In Egypt, yellow - skinned people were respected for defeating white - skinned people."
Just talk about the roommate. Why bring up the roommate from Inner Mongolia? What's the point? This can easily come across as overgeneralizing and unfriendly...	Harmless	"Why don't you say it's from Guizhou?", "Isn't it from Jiangxi?"	"There are many black intermediaries and annoying second-landlords from Northeast China in Beijing", "Aren't people from Northeast China generally straightforward? Why is it like this...?"
The accent of the Chinese Korean ethnic group is closer to the Korean language of North Korea rather than that of South Korea, right [Surprise]	Harmless	"The Chinese pronunciation of North Koreans is related to the Korean language"	"Don't be overly feminist about everything", " South Korean international students have extremely low quality", "Some South Koreans indeed have extremely low quality"
The Anti Bar leans to the left. Why are there so many posts in the Anti Bar that look down on black people? Isn't discriminating against black people a typical characteristic of the right wing?	Bias	"Why are black people discriminated against?", "Black people are all lazy"	"At that time, Asians inspired Africa and Latin America. In Egypt, yellow-skinned people were respected for defeating white-skinned people."