CHAIRO: Contextual Hierarchical Analogical Induction and Reasoning Optimization for LLMs*

Haotian Lu[1,2*], **Yuchen Mou**[1,3*], **Bingzhe Wu**[1 †]
[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
haotianlu666@gmail.com, e1520377@u.nus.edu, wubingzheagent@gmail.com
* Equal contribution † Corresponding author

Abstract

Warning: This paper may contain content that could be disturbing or offensive.

Content moderation in online platforms faces persistent challenges due to the evolving complexity of user-generated content and the limitations of traditional rule-based and machine learning approaches. While recent advances in large language models (LLMs) have enabled more sophisticated moderation via direct prompting or fine-tuning, these approaches often exhibit limited generalization, interpretability, and adaptability to unseen or ambiguous cases.

In this work, we propose a novel moderation framework that leverages analogical examples to enhance rule induction and decision reliability. Our approach integrates end-to-end optimization of analogical retrieval, rule generation, and moderation classification, enabling the dynamic adaptation of moderation rules to diverse content scenarios. Through comprehensive experiments, we demonstrate that our method significantly outperforms both rule-injected fine-tuning baselines and multi-stage static RAG pipelines in terms of moderation accuracy and rule quality. Further evaluations, including human assessments and external model generalization tests, confirm that our framework produces rules with better clarity, interpretability, and applicability. These findings show that analogical example-driven methods can advance robust, explainable, and generalizable content moderation in real-world applications.

1 Introduction

The exponential growth of online content has made content moderation an indispensable component in maintaining healthy, safe, and compliant digital environments (Yuan et al., 2024). Automated content moderation systems are increasingly relied upon to filter out harmful, illegal, or inappropriate material on social media platforms, forums, and other user-driven services (Zeng et al., 2024). As large language models (LLMs) have demonstrated remarkable progress across various natural language understanding tasks, deploying LLMs for content moderation has become a promising direction (Kolla et al., 2024; Nghiem and III, 2024; Wu et al., 2024). Recent studies have explored the application of LLMs to content moderation tasks through diverse strategies, including post-training (Ouyang et al., 2022; Rafailov et al., 2023; Khaliq et al., 2024; Liu et al., 2025; Ma et al., 2024) and prompt engineering (Radford et al., 2019; Palla et al., 2025; Kolla et al., 2024; Brown et al., 2020; Chen et al., 2024a), yielding promising progress in both moderation accuracy and reasoning abilities (Kumar et al., 2024; Vishwamitra et al., 2024).

However, despite their impressive capabilities, even state-of-the-art LLMs often struggle in scenarios characterized by contextual ambiguity or vague moderation criteria (Masud et al., 2024; Huang, 2025; Keluskar et al., 2024). For example, when moderation rules are implicit, incomplete, or open to interpretation, LLMs may produce inconsistent or erroneous judgments, undermining the reliability of automated moderation systems. As shown in Figure 1, Chain of thought (CoT) relies solely on explicit standards such as the absence of insults, incitement, or attacks on specific groups, and fails to recognize the underlying discriminatory logic of the statement "low scores equal low ability." As a result, it incorrectly classifies the statement as "Safe." This approach lacks the deep semantic understanding needed to address finer-grained, metaphorical, or indirect discriminatory content. Conversely, when explicit and precise moderation rules are incorporated into the context, models demonstrate significantly improved accuracy and interpretability, as Figure 1 illustrates. This obser-

*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.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36526–36537

vation highlights the importance of well-defined moderation rules, which improve both moderation precision and transparency of the model's decision-making process (Rebedea et al., 2023; Kumar et al., 2024; Wu et al., 2025).

Nevertheless, identifying or constructing the most appropriate moderation rule for a given content instance remains a challenging problem. Existing solutions typically fall into two categories: (1) manually defined high-level rules, such as those targeting broad categories like "sexual content." While effective to some extent, such rules often fail to account for the nuanced differences among fine-grained instances or across diverse application scenarios, making it virtually impossible to exhaustively enumerate all necessary rules (Chandrasekharan et al., 2019; He et al., 2024). (2) LLM-driven adaptive rule discovery, which leverages the model's world knowledge and prompt engineering to synthesize rules on the fly (Kumar et al., 2024). However, these approaches frequently overlook domain-specific expertise and the rich experience accumulated by human moderators, relying instead on generic or coarse-grained priors.

To address these challenges, we propose leveraging the inductive power of LLMs to generalize rules from analogous instances within the same moderation context. Our key insight is that by systematically analyzing similar content samples and their corresponding moderation outcomes, models can distill more robust and contextually relevant rules that generalize better across instances. A straightforward approach to realizing this high-level idea is to first retrieve samples similar to the current instance from an existing database and then employ an auxiliary LLM to induce and analyze the relevant rules based on these retrieved examples. However, this pipeline separates the processes of rule generation and content moderation, and thus may lose fine-grained cues most pertinent to the current sample during rule induction.

To address this limitation, we introduce CHAIRO, a **c**ontextual **h**ierarchical **a**nalogical **i**nduction and **r**easoning **o**ptimization framework, which jointly optimizes the case retrieval and rule induction process. By jointly optimizing these components, our method ensures that the induced rules are grounded in highly relevant examples and better tailored to each moderation instance. This end-to-end approach allows the model to more effectively utilize the annotated data and uncover hidden expert knowledge from human moderators.

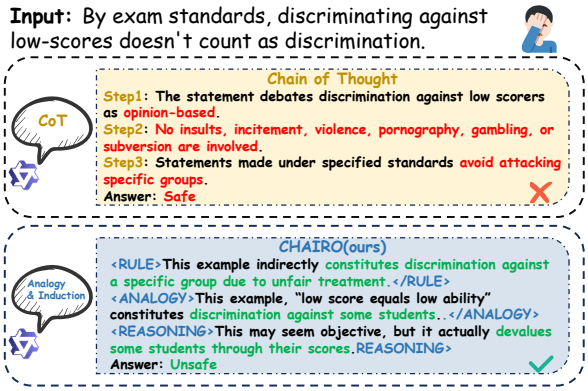

Figure 1: Comparison of the Chain of Thought (CoT) method and the CHAIRO framework for content moderation. CoT fails to identify the implicit bias and classifies it as 'Safe,' while CHAIRO, leveraging analogical reasoning and explicit rule induction, correctly identifies it as 'Unsafe' by recognizing the underlying discriminatory logic.

Concretely, our post-training framework is designed to endow the model with an integrated three-stage reasoning capability: example generation, rule induction, and final moderation. Our framework consists of three critical steps: we first fine-tune the base LLM with a chain-of-analogy approach on the training set, enabling the model to autonomously generate the most relevant analogical cases for each content sample. Second, we employ an auxiliary rule-generation module that synthesizes explicit moderation rules by analyzing the commonalities between the original and analogous cases. Finally, these generated rules are injected back into the LLM's moderation reasoning chain during a second round of fine-tuning, equipping the model with both exemplar-based and rule-inductive reasoning capabilities.

Through comprehensive experiments, we demonstrate that our framework leads to more accurate and robust moderation outcomes while enhancing the interpretability and adaptability of LLM-based moderation systems across diverse, real-world scenarios.

2 Method

Our proposed framework, CHAIRO, leverages a systematic three-stage pipeline to enhance the reasoning capabilities of LLMs in content moderation tasks. Specifically, our approach comprises the following stages as shown in Figure 2: First, we introduce a self-augmented analogical chain-of-thought generation strategy, enriching training data to em-

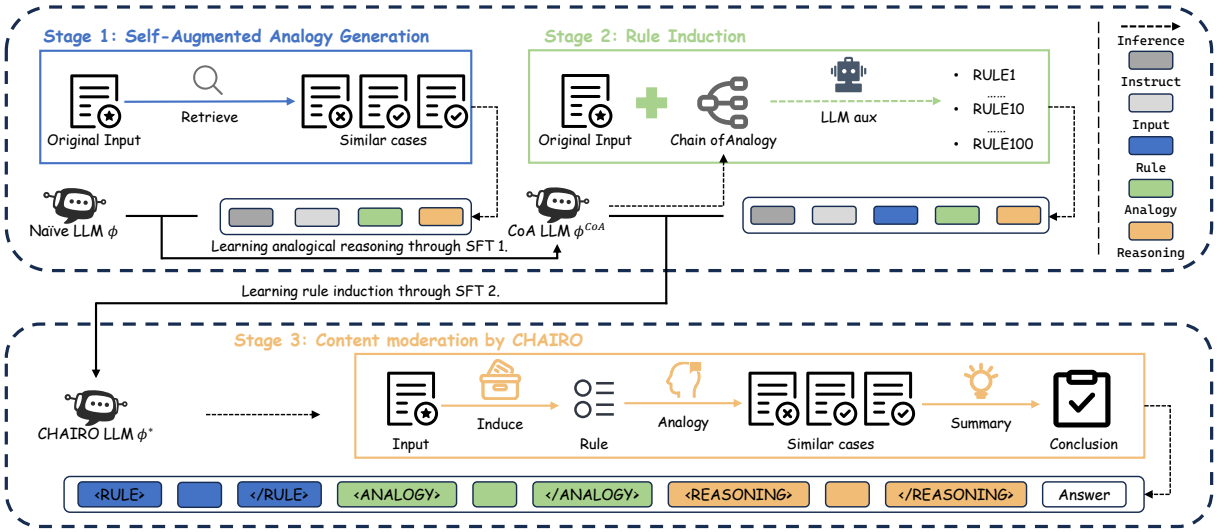

Figure 2: The workflow of the CHAIRO framework.

power the model's capability of adaptively retrieving relevant analogical moderation cases. Second, an auxiliary reasoning LLM performs inductive reasoning over these retrieved examples, extracting explicit moderation rules tailored to specific moderation contexts. Finally, we inject these induced moderation rules back into the reasoning chains and fine-tune the model again, integrating exemplar-based and rule-inductive reasoning capabilities. This structured approach effectively leverages annotated moderation data and human expertise, significantly improving moderation performance and interpretability.

2.1 Self-augmented Analogical Chain-of-Thought Generation

In this initial stage, we augment existing labeled moderation data through a bootstrapping retrieval-enhancement procedure following prior work (Ma et al., 2024). Specifically, denoting the initial base LLM's parameters as ϕ. For each labeled moderation instance x_i with corresponding moderation decision y_i from the training set, we first employ BGE-M3 (Chen et al., 2024b) to derive semantic embeddings for all instances. Subsequently, we compute the cosine distance between every pair of samples as the metric for semantic similarity. Finally, we retrieve a set of analogical moderation examples $\mathcal{A}(x_i)$ for each training instance x_i. Prompted with the current sample and the labels of its analogous examples, the model parameterized by ϕ generates an augmented analogical moderation reasoning chain $\hat{c}_i^{\text{aug}}$:

$$\hat{c}_i^{\text{aug}} = \text{LLM}_\phi(x_i, y_i, \mathcal{A}(x_i)). \tag{1}$$

Applying this augmentation procedure systematically across the entire training set yields an enriched training dataset:

$$\mathcal{D}_{\text{aug}} = \{(x_i, y_i, \hat{c}_i^{\text{aug}})\}_{i=1}^{N}. \tag{2}$$

Finally, we perform supervised fine-tuning (SFT) of the base model parameters ϕ using the augmented dataset $\mathcal{D}_{\text{aug}}$, resulting in updated model parameters ϕ^{CoA}:

$$\phi^{\text{CoA}} = \arg\max_{\phi} \sum_{(x_i, y_i, \hat{c}_i^{\text{aug}}) \in \mathcal{D}_{\text{aug}}} \log P_\phi(y_i, \hat{c}_i^{\text{aug}} \mid x_i). \tag{3}$$

Through this fine-tuning step, the model acquires enhanced analogical reasoning capabilities and becomes proficient in adaptively generating relevant analogical cases for unseen moderation instances. This adaptive analogical reasoning capability lays a solid foundation for subsequent rule induction.

2.2 Rule Induction via Auxiliary Reasoning Model

In this stage, we perform explicit rule induction to extract moderation rules from analogous examples generated by the previously trained analogical reasoning model. Specifically, for each training instance x_i, we first utilize the fine-tuned analogical reasoning model (parameterized by ϕ^{CoA}, already trained in the previous subsection to generate analogical chains of thought) to generate virtual analogical samples $\mathcal{A}(x_i)$:

$$\mathcal{A}(x_i) = \text{LLM}_{\phi^{CoA}}(x_i). \quad (4)$$

Then, leveraging an auxiliary reasoning model (denoted as LLM_{aux}), we perform inductive reasoning on the retrieved analogous instances and the target instance x_i, synthesizing explicit moderation rules r_i (a simple text). Specifically, we use QwQ-32B as LLM_{aux} to generate analogy-based reasoning chains and induce explicit rules from the analogous examples. For quality control, we automatically verify that the category descriptions in each reasoning chain match the target labels and discard samples with inconsistencies. A random subset is further checked manually to validate the reasoning quality. Formally, the generation of the moderation rule r_i can be expressed as:

$$r_i = \text{LLM}_{\text{aux}}(x_i, \mathcal{A}(x_i); \text{prompt}_{\text{rule}}), \quad (5)$$

where $\text{prompt}_{\text{rule}}$ denotes the carefully designed prompts used for rule induction (see detailed prompts in the Appendix).

This explicit rule induction step systematically captures shared characteristics and moderation criteria across similar moderation instances, providing contextually precise and interpretable moderation rules. These induced rules serve as critical input for the subsequent reasoning chain refinement and final moderation decision-making stage.

2.3 Rule Injection and Final Model Refinement

In the final stage, we inject the moderation rules derived from the previous rule induction step back into the reasoning process to further enhance the moderation capabilities of the model. Specifically, given the training instance x_i, the corresponding analogical examples $a_i = \mathcal{A}(x_i)$ obtained from the previous stages, and the explicit moderation rule r_i generated by the auxiliary reasoning model, we employ an additional reasoning model (LLM_{aux}) to synthesize these components along with the instance's label y_i, producing a comprehensive moderation reasoning chain c_i':

$$c_i' = \text{LLM}_{\text{aux}}(x_i, a_i, r_i, y_i; \text{prompt}_{\text{reasoning}}), \quad (6)$$

where $\text{prompt}_{\text{reasoning}}$ represents carefully engineered instructions guiding the synthesis process.

We then structure each moderation instance's reasoning chain in a hierarchical format using special tokens to clearly delineate different reasoning components, forming a enhanced hierarchical moderation chain $\hat{c}_i^{\text{refined}}$ as shown below:

$$\hat{c}_i^{\text{refined}} = \left\langle \begin{array}{c} \texttt{<RULE>}\, r_i \,\texttt{</RULE>} \\ \texttt{<ANALOGY>}\, a_i \,\texttt{</ANALOGY>} \\ \texttt{<REASONING>}\, c_i' \,\texttt{</REASONING>} \end{array} \right\rangle \quad (7)$$

Finally, leveraging this hierarchical, structured reasoning chain, we conduct an additional round of SFT, updating the model parameters from ϕ^{CoA} to the final refined parameters ϕ^*:

$$\phi^* = \arg\max_{\phi'} \sum_{(x_i, y_i, \hat{c}_i^{\text{refined}}) \in \mathcal{D}_{\text{refined}}} \log P_{\phi'}(y_i, \hat{c}_i^{\text{refined}} \mid x_i), \quad (8)$$

where

$$\mathcal{D}_{\text{refined}} = \{(x_i, y_i, \hat{c}_i^{\text{refined}})\}_{i=1}^{N}. \quad (9)$$

Through this rule injection and hierarchical refinement process, the model effectively integrates analogical reasoning, explicit rule induction, and label-guided reasoning into a unified moderation reasoning capability, significantly enhancing both interpretability and moderation accuracy.

3 Experiment

3.1 Settings

All experiments were conducted on a single server with 4×NVIDIA H20 (96GB) GPUs using the LLaMA Factory framework (Zheng et al., 2024) with DeepSpeed ZeRO-3 optimization (Rajbhandari et al., 2020). For First-stage SFT, we trained the model for 1 epoch with a learning rate of 1.0e-5 using bfloat16 mixed-precision (Micikevicius et al., 2018), achieving an effective batch size of 64 (micro-batch size × gradient accumulation steps × GPUs = 2 × 8 × 4). For Second-stage SFT, the settings were the same as those for First-stage SFT. For Retrieval-Augmented Generation (RAG), we utilized the 32 most similar reference examples to each input query. For text generation, we employed top-k sampling (Fan et al., 2018) with temperature=0.8 and top-p sampling (Holtzman et al., 2020).

3.2 Main Results

3.2.1 Overview

In this section, we systematically investigate the effectiveness of the proposed framework by addressing the following key research questions:

- **RQ1:** Can introducing explicit rules into the LLM-based moderation process improve moderation performance?
- **RQ2:** Does incorporating analogical examples enhance the quality of generated moderation rules?
- **RQ3:** Does the end-to-end two-stage optimization proposed in our framework lead to improved classification reliability?
- **RQ4:** Are the moderation rules generated by our framework of higher quality and better generalizability compared to those produced by single-instance approaches?

To comprehensively answer these research questions, we conduct extensive comparative experiments using Qwen3-8B (Yang et al., 2025) as our base language model on a series of standard benchmarks commonly adopted in the content moderation domain.

Specifically, we choose a fine-grained content moderation dataset proposed by the prior work (Ma et al., 2024). According to prior studies, this dataset is particularly valuable as it originates from realistic moderation scenarios and includes challenging subcategories like politically sensitive content, where context ambiguity commonly leads to difficulty in accurate moderation. Hence, it is ideally suited to rigorously validate the effectiveness and practical utility of the rules generated by our framework.

Additionally, an experimental result on another widely-used moderation benchmark are summarized in Table 3. These results collectively demonstrate the generalizability and robustness of our proposed approach across diverse moderation scenarios and datasets.

3.2.2 Analysis of RQ1: Effectiveness of Introducing Explicit Rules

To address the first research question, we compare our proposed method CHAIRO against two baseline approaches: (1) Naive SFT, where the model is fine-tuned on data containing moderation reasoning chains but without explicit rules, and (2) Standard prompting method, which directly queries the pre-trained LLM without fine-tuning or rule injection.

The experimental results clearly demonstrate the substantial benefits of explicitly incorporating moderation rules into the LLM-based moderation workflow. Specifically, our proposed method consistently surpasses baseline approaches across various moderation categories, significantly improving the moderation F1 scores. For instance, our method achieves an improvement of 5.3% in F1 scores compared to the naive SFT baseline. Such pronounced performance gains strongly validate the importance and utility of injecting explicit moderation rules into the LLM moderation process, effectively resolving ambiguities and enhancing decision-making precision in challenging moderation scenarios.

3.2.3 Analysis of RQ2: Effectiveness of Analogical Examples in Enhancing Rule Generation Quality

To address Research Question 2, we compare our method against a rule-injected SFT baseline, in which explicit moderation rules are similarly injected into the fine-tuning process. The key difference is that these rules are generated by the LLM based on individual target instances alone, without leveraging analogical examples.

The experimental results clearly demonstrate that our analogical example-based rule induction significantly enhances the quality of moderation rules compared to those generated from single-instance contexts alone. By leveraging analogical examples, our method effectively captures broader contextual nuances and moderation criteria, resulting in more comprehensive and generalizable moderation rules. For instance, our proposed framework achieves an improvement of 4.5% in F1 scores over the baseline approach using rules generated from single-instance prompting. These findings strongly confirm the benefit and effectiveness of incorporating analogical examples in moderation rule generation.

3.2.4 Analysis of RQ3: Impact of End-to-End Two-Stage Optimization on Classification Reliability

To investigate Research Question 3, we compare our proposed framework against a static RAG baseline. The Static RAG baseline follows a multi-stage moderation pipeline: it first retrieves relevant ex-

amples via static retrieval, then induces moderation rules using an additional LLM separately, and finally injects these rules into the moderation context for classification decisions.

Experimental results clearly indicate that our proposed end-to-end two-stage optimization significantly enhances moderation reliability compared to the static multi-stage RAG baseline. By jointly optimizing the analogical retrieval, rule induction, and moderation decision-making processes, our approach effectively reduces the cumulative errors and inconsistencies that arise from separately optimized stages. Specifically, our method achieves an improvement of 2.3% in moderation accuracy compared to the static RAG baseline. This outcome underscores the advantage and necessity of employing our unified, end-to-end optimization strategy to achieve more reliable and consistent moderation decisions.

3.3 Discussion

The experimental results presented in the preceding section demonstrate that our proposed CHAIRO framework, by integrating analogical reasoning, explicit rule induction, and hierarchical reasoning chain optimization, significantly enhances the performance of LLM-based content moderation systems across various content moderation tasks. In this section, we delve deeper into the contributions of individual components, provide case studies for qualitative insights, and present human evaluations of the generated rule quality to further elucidate the framework's effectiveness and practical implications.

3.3.1 Ablation Study

To isolate the contributions of key components in our CHAIRO framework, we conduct ablation experiments by systematically removing core modules and measuring the resulting performance degradation, as reported in Table 2.

Role of the Retrieval Module. In the "w/o k-NN Retrieval" setting, we replace the k-NN-based analogical retrieval in Stage 1 with random sample selection. This leads to a 2.3% drop in the overall F1 score, with particularly notable declines in politically sensitive content (-3.9%) and gambling-related content (-3.5%). The result underscores the critical role of retrieving relevant analogical cases in capturing subtle moderation criteria, especially in domains where rules are implicit or context-dependent. Without the most relevant analogical examples, the model struggles to generalize beyond superficial patterns, often leading to oversimplified judgments.

Importance of the Second-Stage Fine-Tuning. In the "w/o Second-stage SFT" setting, we skip the rule injection and second-round fine-tuning in Stage 3, so the model relies solely on Stage 1 analogical reasoning. This leads to a 1.2% decline in overall F1 score, with the most significant drops observed in political content (-3.2%) and pornography (-2.2%). The result highlights the value of combining explicit rules with analogical reasoning. The second-stage fine-tuning ensures that rules are dynamically adapted to specific content contexts, bridging the gap between general guidelines and case-specific nuances.

These results confirm that the retrieval and rule injection components complement each other: retrieval supplies relevant examples, and rule induction turns them into generalizable moderation standards.

3.3.2 Results on Other Dataset

We further evaluate model performance on additional datasets to assess the cross-dataset generalization and harmful content recognition capabilities of each method. As summarized in Table 3, there are clear differences in F1 scores across categories and models. Our proposed method, CHAIRO achieves the highest overall average F1 score of 68.0, outperforming SFT (61.6), Qwen3-8B (52.9), and RAG (54.7).

A closer look at category-level results reveals that CHAIRO consistently achieves the best or second-best F1 scores in key harmful content categories, including "Hate" (81.5), "Sexual" (81.6), "Confessions" (82.5), "Harassment" (40.7), and "Profanity" (53.9). Notably, the improvements in the "Hate" and "Sexual" categories are particularly pronounced compared to other methods.

Overall, these results demonstrate that CHAIRO exhibits stronger and more comprehensive performance in most harmful content recognition categories, validating its effectiveness and generalizability in cross-dataset scenarios. This further highlights the practical value of our approach for robust and reliable harmful content detection in diverse real-world settings.

3.3.3 Rule Quality Evaluation

How effective are the moderation rules generated by our proposed method in practical moderation

Category	Model/Method	Average	Politics	Pornography	Violence	Gambling	Bias	Harmless
General LLMs	GPT-4	72.3	58.6	88.7	79.8	92.7	64.3	56.8
	DeepSeek R1	77.1	72.7	91.4	86.1	94.3	64.6	59.7
	DeepSeek V3	80.3	79.0	90.3	89.8	95.0	70.5	62.5
	Qwen2.5-32B-Instruct	74.3	59.1	91.1	84.4	95.4	67.9	54.2
	QwQ-32B	69.1	75.4	69.6	72.0	84.9	60.7	54.6
	LLaMA3-8B	67.5	58.5	55.9	81.0	90.6	65.2	44.2
Specific LLMs	LLaMA-Guard-3-8B	39.7	12.0	74.1	41.8	29.4	45.7	35.6
Proposed Methods	Rule Impact (RQ1)	83.9	83.8	67.5	92.2	73.7	90.6	95.7
	Analogy-Rule Quality (RQ2)	84.7	84.4	68.4	92.8	75.4	90.3	96.9
	E2E Reliability (RQ3)	86.9	88.3	93.2	96.0	98.0	82.2	63.7
	CHAIRO (Ours)	**89.2**	**89.3**	**71.5**	**97.8**	**82.0**	**96.1**	**98.6**

Table 1: Moderation F1 Scores for General LLMs, Specific LLMs and CHAIRO

Setting	Overall F1	Politics	Pornography	Violence	Gambling	Bias	Harmless
w/o k-NN Retrieval	86.9 (-2.3)	85.4 (-3.9)	69.0 (-2.5)	96.6 (-1.2)	78.5 (-3.5)	93.6 (-2.5)	98.2 (-0.4)
w/o Second-stage SFT	88.0 (-1.2)	86.1 (-3.2)	69.3 (-2.2)	97.2 (-0.6)	81.2 (-0.8)	95.8 (-0.3)	98.4 (-0.2)
CHAIRO (Ours)	**89.2**	**89.3**	**71.5**	**97.8**	**82.0**	**96.1**	**98.6**

Table 2: Ablation Study on Model Components with Performance Change

Categories	Qwen3-8B	RAG	SFT	CHAIRO (Ours)
Hate	67.2	62.3	74.7	**81.5**
Sexual	65.2	72.3	71.4	**81.6**
Confessions	66.4	68.3	80.0	**82.5**
Harassment	25.8	27.6	26.7	**40.7**
Profanity	40.0	42.9	**55.0**	53.9
Average F1-Score	52.9	54.7	61.6	**68.0**

Table 3: F1-Score Comparison on Aegis Dataset

Model	F1	Human (%)
Qwen2.5-32B-Instruct	74.3	-
Simple Rule	75.1	15
RQ4	88.7	85

Table 4: Rule Quality Evaluation Results. The F1-score reflects the generalization ability of rules when applied to an external moderation model. The "Human (%)" denotes the preference rate derived from a double-blind comparison of 100 test cases by three annotators with content moderation experience, where rules are assessed for contextual relevance, completeness, and alignment with human judgment criteria.

scenarios?

To address Research Question 4, we evaluate the quality of the moderation rules themselves, beyond end-to-end moderation accuracy. We compare rules generated by our analogical approach against those produced by single-instance LLM prompting, using two complementary evaluations summarized in Table 4.

Human Evaluation. We invited three annotators with practical experience in content moderation to independently rate the quality, clarity, and usefulness of the moderation rules produced by our method and the baseline. Each annotator independently reviewed 100 randomly ordered pairs of rules from different methods in a double-blind setting, and selected the one they judged more contextually relevant, complete, and reliable. The annotators consistently preferred rules generated by our analogical approach, highlighting their clarity and practical utility. Across the 100 test cases, the three annotators preferred our rules in 85% of cases, compared to only 15% for the simple rule-based baseline. The Qwen2.5-32B-Instruct baseline was not included in the human preference comparison. These results indicate that analogically generated rules align more closely with human moderation judgment.

Generalization Assessment with External Models. We further evaluated the generalizability of the generated rules by injecting them into an external model distinct from our base model Qwen3-8B. Rules generated by our method achieve an F1-score of 88.7 when applied to the external model, outperforming both the Qwen2.5-32B-Instruct baseline (F1 = 74.3) and the simple rule-based approach (F1 = 75.1). This represents an absolute improvement of 14.4 and 13.6 percentage points respectively, confirming that the generated rules transfer well beyond the original modeling context.

3.3.4 Adaptability to Evolving Standards

Moderation standards shift over time and differ across cultures and platforms. The out-of-distribution results in Table 3 show that CHAIRO

retains strong performance under distributional shift, suggesting that the combination of analogical reasoning and explicit rules provides robustness beyond the training distribution. Moreover, the modular design of our framework supports lightweight fine-tuning on recent data when the shift is moderate, without requiring a full retraining of the pipeline.

3.3.5 On the Choice of Supervised Fine-Tuning

We explored reinforcement learning as an alternative to SFT in preliminary experiments. However, binary reward signals based on label correctness proved too coarse for this task, and the model quickly learned superficial shortcuts instead of performing genuine analogical reasoning and rule application. We attribute this to the multi-step nature of our reasoning chain, where intermediate quality matters but is not captured by a single binary reward. Designing reward functions that also account for reasoning chain quality is a promising direction for future work.

4 Related Work

4.1 LLMs for Content Moderation

In recent years, LLMs have garnered significant attention in content moderation research due to their strong natural language understanding capabilities. Existing LLM-based moderation approaches can be broadly categorized into two paradigms:

Prompt Engineering-driven Direct Moderation: Models such as GPT-series can directly generate moderation decisions from carefully designed prompts (Radford et al., 2019). This approach does not require large-scale annotated data and can inherently provide detailed reasoning processes, enhancing interpretability (Zhan et al., 2025; Li et al., 2025). However, when moderation criteria are inherently ambiguous, model predictions become sensitive to subtle variations in prompt wording, leading to inconsistent moderation outcomes (Röttger et al., 2022; Gligoric et al., 2024).

Fine-tuning-based Domain Adaptation: Post-training large pretrained models on annotated moderation datasets has been explored extensively. For example, contrastive fine-tuning approaches leverage labeled datasets containing both compliant and violating content to enhance models' sensitivity to specific moderation rules (Devlin et al., 2019). While fine-tuning clearly improves performance on explicitly annotated rules, it suffers from data-dependency issues: high-quality annotated moderation datasets are costly to acquire, and fine-tuned models often fail to generalize to previously unseen moderation rules or subtle semantic nuances (Jha et al., 2024; Inan et al., 2023).

4.2 Generation and Optimization of Moderation Rules

The explicitness and clarity of moderation rules are critical for ensuring reliable moderation decisions. However, existing methods for constructing moderation rules have notable shortcomings:

Manually-defined High-level Rules: Current moderation systems commonly utilize abstract categories like hate speech and adult content as moderation guidelines. However, such broad categories often fail to adequately cover fine-grained real-world scenarios. For instance, subtle forms of harassment or discrimination expressed through metaphor, euphemism, or implicit linguistic cues cannot be effectively captured by simple keyword-based rules (Mei et al., 2024; Wang et al., 2024; Palla et al., 2025). Furthermore, exhaustively enumerating all possible instances or patterns of problematic content through manual effort alone is practically infeasible.

LLM-driven Adaptive Rule Generation: Recent studies have explored prompting-based approaches that use LLMs to automatically generate moderation rules by summarizing characteristics of violating content (Franco et al., 2023). However, this strategy relies heavily on the generic world knowledge encoded in large models and often neglects domain-specific expert knowledge such as platform-specific community guidelines. Additionally, current LLM-driven rule generation processes are typically independent from the moderation decision-making stage. As a consequence, generated rules lack the flexibility and context-awareness required to dynamically adapt to specific content instances, limiting the practical effectiveness of such approaches in realistic moderation scenarios (Masud et al., 2024).

5 Conclusion

We presented CHAIRO, a framework that jointly optimizes analogical retrieval, rule induction, and hierarchical reasoning for LLM-based content moderation. On both in-distribution and out-of-distribution benchmarks, CHAIRO outperforms

SFT and RAG baselines by a clear margin. Human annotators also preferred the rules induced by our method over single-instance alternatives in 85% of cases. The modular design also makes each moderation decision traceable to explicit rules and analogical evidence, improving interpretability for practical deployment.

Limitations

Although this study aims to provide a reliable and robust moderation approach for harmful content on real-world online platforms, several important limitations remain.

First, this study primarily focuses on the textual modality and has not yet extended the proposed reasoning paradigm to multimodal large language models . In the contemporary context where audio-visual content is increasingly prevalent, moderation of multimodal content, including short videos and live streaming, is an equally pressing challenge. A key direction for future work is to extend the proposed paradigm to content moderation tasks involving multimodal data.

In addition, although this study conducts systematic experimental analyses across several real-world datasets, the effectiveness and robustness of the model in actual platforms and complex application scenarios remain to be further validated, particularly in contexts involving interactive dialogues, contextual dependencies, and user diversity.

Additionally, while our framework shows robustness to moderate distributional shifts, drastic changes in moderation standards across time or culture may still require retraining or rule reconstruction (see Section 3.3.4 for further discussion). Similarly, extending beyond supervised fine-tuning to reinforcement learning remains an open challenge, as discussed in Section 3.3.5.

Acknowledgments

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

Ethical considerations

Exposure to Offensive Content: During this study, we encountered and curated a substantial amount of offensive content for the purpose of constructing the research dataset. All authors were fully aware of the nature of the study and consented to review such materials. It is important to note that, owing to the use of an isolated experimental protocol, no individuals other than the authors were exposed to these materials.

References

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.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30.

Jianfa Chen, Emily Shen, Trupti Bavalatti, Xiaowen Lin, Yongkai Wang, Shuming Hu, Harihar Subramanyam, Ksheeraj Sai Vepuri, Ming Jiang, Ji Qi, Li Chen, Nan Jiang, and Ankit Jain. 2024a. Class-rag: Real-time content moderation with retrieval augmented generation. *Preprint*, arXiv:2410.14881.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2023. Analyzing the use of large language models for content moderation with chatgpt examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 1–8.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5942–5959.

Zihao He, Jonathan May, and Kristina Lerman. 2024. Cpl-novid: Context-aware prompt-based learning for norm violation detection in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 569–582.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Tao Huang. 2025. Content moderation by llm: From accuracy to legitimacy. *Artificial Intelligence Review*, 58(10):1–32.

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

Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. 2024. MemeGuard: An LLM and VLM-based framework for advancing content moderation via meme intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8084–8104, Bangkok, Thailand. Association for Computational Linguistics.

Aryan Keluskar, Amrita Bhattacharjee, and Huan Liu. 2024. Do llms understand ambiguity in text? a case study in open-world question answering. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7485–7490. IEEE.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

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.

Yaqiong Li, Peng Zhang, Hansu Gu, Tun Lu, Siyuan Qiao, Yubo Shu, Yiyang Shao, and Ning Gu. 2025. Demod: A holistic tool with explainable detection and personalized modification for toxicity censorship. *Proc. ACM Hum. Comput. Interact.*, 9(2):1–24.

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. *Preprint*, arXiv:2501.18492.

Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu. 2024. Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. *Preprint*, arXiv:2310.03400.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of llms in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 15847–15863. Association for Computational Linguistics.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. 2024. Hiddenguard: Fine-grained safe generation with specialized representation router. *CoRR*, abs/2410.02684.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *Preprint*, arXiv:1710.03740.

Huy Nghiem and Hal Daumé III. 2024. Hatecot: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 5938–5956.

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, Andreas Damianou, Henrik Lindström, Daniel R. Taber, and Mounia Lalmas. 2025. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2025, Athens, Greece, June 23-26, 2025*, pages 840–854. ACM.

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.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. Nemo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 431–445.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

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 *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, pages 788–806. IEEE.

Peiran Wang, Qiyu Li, Longxuan Yu, Ziyao Wang, Ang Li, and Haojian Jin. 2024. Moderator: Moderating text-to-image diffusion models through fine-grained context-based policies. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pages 1181–1195.

Jialin Wu, Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Jiayang Xu, Xinfeng Li, and Wenyuan Xu. 2024. Legilimens: Practical and unified content moderation for large language model services. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pages 1151–1165.

Mengyang Wu, Yuzhi Zhao, Jialun Cao, Mingjie Xu, Zhongming Jiang, Xuehui Wang, Qinbin Li, Guangneng Hu, Shengchao Qin, and Chi-Wing Fu. 2025. Icm-assistant: Instruction-tuning multimodal large language models for rule-based explainable image content moderation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 8413–8422.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

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, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative ai content moderation based on gemma. *Preprint*, arXiv:2407.21772.

Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2025. Slm-mod: Small language models surpass llms at content moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 8774–8790. Association for Computational Linguistics.

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 Prompts

We modularly synthesize each constituent of the proposed specialized chain-of-thought, thereby both transparently elucidating the contribution of every component to the overall reasoning trajectory and progressively enhancing the chain's logical coherence and robustness through iterative leverage of a more capable large language model, as shown in Figure 3.

- Chain of Thought: In this stage, we generate a chain-of-thought-style reasoning trajectory for each input instance, thereby furnishing the foundational inferential logic that underpins both the focal sample and its reference exemplars.

- Chain of Analogy: In this stage, we augment existing labeled moderation data through a bootstrapping retrieval-enhancement procedure, thereby endowing the model with an analogical-reasoning epistemic modality.

Prompt for Chain of Thought

You are a security moderator. After analysis, classify the following content into only one of these categories: politically harmful, pornographic, violent, gambling-related, discriminatory/abusive, or harmless. Your response should be provided in the format <REASONING>...</REASONING>, and directly after </REASONING> state the final classification result.

Prompt for Chain of Analogy

You are a security moderator. After analysis, classify the following content into only one of these categories: politically harmful, pornographic, violent, gambling-related, discriminatory/abusive, or harmless. Your response should be provided in the format <ANALOGY>...</ANALOGY><REASONING>...</REASONING>, where the ANALOGY section is dedicated to citing analogous reference cases, and the REASONING section is dedicated to comprehensive analysis. Directly after </REASONING>, state the final classification result.

Prompt for Rules

You are a security moderator. After analysis, extract general rules from the following content without explicitly referencing cases. Return only the extracted rules in the format <RULE>...</RULE>.

Prompt for CHAIRO

You are a security moderator. After analysis, classify the following content into only one of these categories: politically harmful, pornographic, violent, gambling-related, discriminatory/abusive, or harmless. Your response should be provided in the format <RULE>...</RULE><ANALOGY>...</ANALOGY><REASONING>...</REASONING>, where the RULE section is dedicated to rule citation, the ANALOGY section to citing analogous reference cases, and the REASONING section to comprehensive analysis. Directly after </REASONING>, state the final classification result.

Figure 3: Prompt for synthesizing the chain of analogical inductive reasoning.

- Induction Rules: In this stage, we perform explicit rule induction to extract moderation rules from analogous examples generated by the previously trained analogical reasoning mode, thereby furnishing a principled data substrate for subsequent inductive-capability infusion.
- CHAIRO: In this stage, we inject the moderation rules derived from the previous rule induction step back into the reasoning process to further enhance the moderation capabilities of the model.