RAG-Safe: A Recall-First Safety Framework Comparing Open-Source and Commercial LLM Moderation Pipelines

Alejandro Salinas-Medina[†,‡*], Xue Liu[†, ‡, ◇]
[†] School of Computer Science, McGill University, Montreal, Quebec
[‡] Mila - Quebec AI Institute, Montreal, Quebec
[◇] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Abstract

False negatives—missed unsafe content—are the dominant risk in safety-critical moderation. We study *recall-first* moderation under a deployment-realistic retrieval-augmented classification (RAG) protocol and evaluate two end-to-end stacks on the Jigsaw Unintended Bias in Toxicity Classification corpus using a leakage-resistant *unseen-article* split (no `article_id` overlap between train and test). On a capped test set of n=8,000 (8% `FLAGGED`), Pipeline A (commercial embeddings + hosted inference) attains Accuracy 0.761 with $\text{Recall}_{\texttt{FLAGGED}} = 0.631$ at ReviewRate 0.260. Pipeline B (open-source embeddings + FAISS + local LLM inference via `llama.cpp`) reaches substantially higher $\text{Recall}_{\texttt{FLAGGED}}$ (0.903), at the cost of elevated review burden (ReviewRate 0.613, FPR 0.588). A retrieval-diversity sweep further reveals that graph-based selection yields +19.5 point recall gains over MMR (0.855 vs. 0.660), with retrieval strategy—not threshold tuning—emerging as the primary control mechanism for navigating the recall–review frontier. Baseline comparisons against state-of-the-art RAG systems, kNN and logistic regression classifiers on the same embeddings reveal that Pipeline B achieves the highest recall among all text-based methods, while simpler supervised approaches can match or exceed Pipeline A on recall at lower false positive cost—highlighting that RAG-based moderation does not uniformly dominate traditional classifiers. Bootstrap confidence intervals and paired McNemar testing confirm the difference in error profiles. Subgroup analysis using Jigsaw identity annotations further shows elevated false positive rates for identity-associated text, highlighting the need to evaluate recall jointly with review capacity and fairness-sensitive burden. A component ablation further reveals scale-dependent effects: policy calibration achieves $\text{Recall}_{\text{FLAGGED}} \approx 0.94$ immediately at any scale, while committee retrieval reduces review burden only with sufficiently large indices (FPR: $0.59 \rightarrow 0.42$ at N=16000).

Keywords: Content moderation, Recall-first classification, Distribution-preserving data augmentation, Committee-based retrieval, Retrieval-augmented large language models, Safety-critical AI

1. Introduction

Automated content moderation has become a critical application area for large language models (LLMs), due to their ability to interpret nuanced context and language. Recent studies have explored using LLMs to assist or even replace traditional classifiers in detecting harmful or policy-violating content [1, 2]. Compared to static fine-tuned models, LLMs offer greater flexibility and understanding in borderline cases. OpenAI, for example, has proposed using GPT-4 as a content moderator to achieve more consistent policy enforcement [3]. However, deploying LLMs for moderation also introduces new challenges: LLMs may reflect majority cultural biases [4], struggle with ambiguous "hard cases" that require contextual judgment [5], and can be resource-intensive to query repeatedly [6]. Critically, a key risk in *safety-critical moderation* is *false negatives*—failing to flag harmful content. Recent work emphasizes the need to prioritize recall for unsafe content, even if it means tolerating some false positives [2, 5]. Our work addresses this need by combining *retrieval augmentation* and *contrastive data augmentation* to boost the detection of subtle policy

* alejandro.salinas@mail.mcgill.ca

violations while maintaining overall accuracy in imbalanced datasets, as real-world data is typically distributed this way [7, 8].

Main Contributions. We propose a novel LLM-based moderation pipeline that integrates *diverse retrieval* and *contrastive augmentation* to enhance recall of harmful content. First, we introduce a distribution-preserving paraphrase augmentation strategy that generates additional training examples which are statistically indistinguishable from the original data distribution [9]. Unlike standard augmentation (e.g., synonym replacement or back-translation [10]), our method produces paraphrases that closely match the original length and semantics but include *hard positives/negatives*—subtle rephrases of flagged content and safe content—to better expose decision boundaries. This approach builds on the idea of counterfactual and adversarial augmentation in Natural Language Processing (NLP) [11, 12], but focuses on *safety-critical* categories. Second, we design a *committee-based retrieval* mechanism to supply the LLM with diverse, relevant context at inference time. Instead of retrieving nearest neighbors with a single embedding model, our pipeline combines multiple retrievers and reranking strategies—dense semantic search, Maximal Marginal Relevance (MMR) diversification, and graph-based selection—to ensure retrieved examples are informative and non-redundant [13–15]. Finally, we implement a *recall-oriented decision policy*: the model casts votes across multiple retrieved contexts, and we apply a calibrated threshold that leans on the side of flagging whenever there is reasonable doubt. By aggregating judgments and lowering the threshold for the positive class, we explicitly favor high recall for unsafe content—a property critical for safety filters [5]. To our knowledge, this is the first framework to unify *distribution-preserving augmentation* with *retrieval-diverse RAG* for moderation, showing that careful retrieval diversity and thresholding can push an LLM moderator into a safer operating regime. In addition to its methodological contributions, we use the framework to *benchmark commercial RAG systems against open-source implementations* and to conduct a controlled comparison of open-source versus proprietary subscription-based models under a unified protocol.

Stack comparison scope. Our A/B comparison is *systems-level*: it evaluates two deployment-realistic stacks under a unified, leakage-safe protocol, and is not intended as a claim of vendor-category superiority. We therefore emphasize *operating-point trade-offs* (miss rate vs. review burden and subgroup burden) and report paired uncertainty/statistical tests rather than attributing performance differences to any single component.

2. Related Work

Our work integrates three complementary research directions—LLM-based moderation, retrieval augmentation, and contrastive data augmentation—into a unified framework that explicitly prioritizes recall in safety-critical moderation scenarios. In this section we examine how prior work has applied these approaches in practice.

LLMs for Content Moderation. Chen, Shen, Bavalatti, Lin, Wang, Hu, Subramanyam, Vepuri, Jiang, Qi, Chen, Jiang, and Jain [2] introduce Class-RAG, which enhances robustness against adversarial prompts by retrieving safe and unsafe examples for classification. Kolla, Salunkhe, Chandrasekharan, and Saha [16] demonstrate how GPT-based assistants can support human moderators, while Franco, Gaggi, and Palazzi [17] investigate workflows where LLMs generate policy-grounded explanations. At the same time, concerns remain regarding cultural bias and the dominance of majority perspectives in LLM moderation [4]. To address this, Nguyen, Suresh, and Shieh [4] propose *Mod-Guide*, a retrieval-augmented system designed to surface minority viewpoints. Our approach complements these efforts by placing explicit emphasis on recall-oriented moderation.

Retrieval-Augmented Classification and Diversity. Retrieval-augmented generation (RAG) has become central to knowledge-intensive NLP tasks [18, 19]. In moderation, retrieval provides LLMs with contextual policy text or representative examples [2]. However, the effectiveness of retrieval hinges on both relevance and diversity. Techniques such as maximal marginal relevance (MMR) [20], clustering-based selection [13], and ensemble retrievers [14, 15] have been shown to improve coverage. Zhang, Jin, Cheng, Yu, and Xu [13] demonstrate that redundancy undermines recall, while Chernogorskii, Averkiev, Kudraleeva, Martirosian, Tikhonova, Malykh, and Fenogenova [14] introduce *DRAGON*, a retrieval training method that explicitly promotes diversity. Building on these insights, we employ an inference-time ensemble that integrates dense, MMR, and graph-based retrieval to ensure coverage of multiple facets of potentially harmful content.

Data Augmentation for Moderation. Data augmentation has proven effective for improving generalization, particularly under conditions of class imbalance. Traditional methods include synonym replacement, back-translation, and noise injection [10]. More recently, generative augmentation with LLMs has been investigated [6, 16], though unconstrained paraphrasing risks shifting away from the original distribution. Papakipos and Bitton [9] emphasize the importance of maintaining realism in augmented data. Alternative strategies, such as counterfactual augmentation [11] and adversarial generation for toxicity detection [12], have demonstrated strong improvements in recall. Extending this line of work, we introduce a lightweight paraphrase generator that produces *hard positives and negatives*, thereby enhancing boundary-focused exposure for classifiers.

3. Methods

3.1. Overview

We study *recall-first* moderation: settings where the dominant risk is a false negative (unsafe content predicted safe), and operationally the system is permitted to escalate uncertain cases for human review. Our goal is therefore to characterize (i) **miss rate** on unsafe content, and (ii) the **review burden** induced by conservative behavior. We propose a recall-oriented moderation framework that integrates complementary components into a unified retrieval-augmented classification (RAG) pipeline. *Committee-based retrieval* leverages diverse retrieval strategies (dense, MMR, and graph-based) to construct more informative and less redundant neighborhoods at inference time. We validate this augmentation strategy on auxiliary corpora (Appendix A); our headline Jigsaw evaluation uses original data to ensure comparability with prior work.

Our headline experiments are conducted on the **Jigsaw Unintended Bias in Toxicity Classification** corpus (1.8M+ rows). We evaluate two deployment-realistic retrieval-augmented classification (RAG) pipelines under a unified protocol:

- **Pipeline A (Commercial)**: API-provided embeddings + hosted inference under a deterministic prompt.
- **Pipeline B (Open-source)**: local embeddings + FAISS [21] retrieval + local LLM inference via `llama.cpp`.

We emphasize that these are *systems-level* comparisons of two realistic stacks, therefore we report a detailed operational trade-off analysis and statistical uncertainty rather than attributing improvements to a single component, Figure 1 provides a schematic overview of the recall-oriented moderation framework.

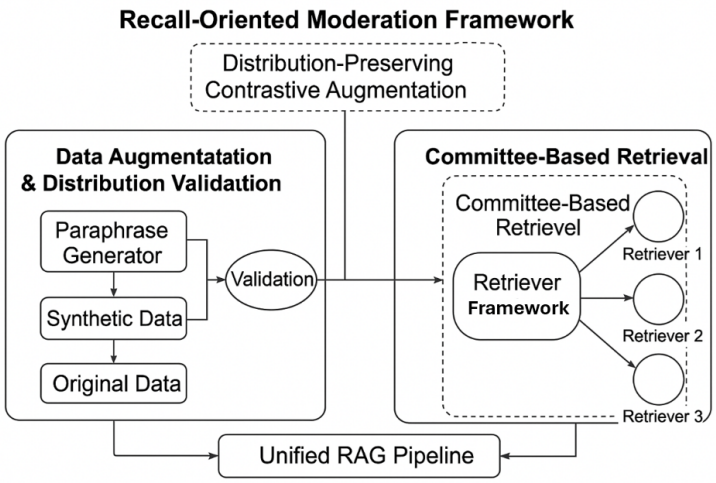

Figure 1. **Overview of our recall-oriented moderation framework.** The system integrates *distribution-preserving contrastive augmentation* and *committee-based retrieval* into a unified RAG pipeline. On the left, augmented samples are generated and validated to preserve the original distribution. On the right, a committee of diverse retrievers ensures retrieval coverage across multiple facets of potentially harmful content.

3.2. Dataset and Realistic Splits (Unseen-Group Evaluation)

A central limitation of prior experiments is their reliance on small or partially synthetic corpora, which fail to reflect the novelty structure observed in deployed moderation—specifically regarding new topics, new sources, and distributional drift. We address this by adopting **group-based** splits on Jigsaw that prevent near-duplicate leakage and enforce novelty.

Let $\mathcal{D}$ be the full Jigsaw dataset with toxicity score `target` $\in [0, 1]$. We binarize labels as

$$y(x) = \mathbb{I}[\texttt{target}(x) \geq 0.5] \in \{\texttt{FLAGGED}, \texttt{NOT FLAGGED}\}.$$

We construct train/test splits using **StratifiedGroupKFold** over the group key `article_id`, selecting the fold whose test fraction is closest to 10%. This produces *unseen-article* evaluation, i.e., no `article_id` appears in both train and test:

$$\mathcal{G}_{\text{train}} \cap \mathcal{G}_{\text{test}} = \emptyset, \quad \mathcal{G} = \{\texttt{article\_id}\}.$$

We then cap the split sizes by stratified subsampling to enable tractable end-to-end runs while preserving label priors (approximately 8% `FLAGGED`).

3.3. Retrieval-Augmented Classification Protocol

For each test query q with text x_q, a retriever returns top-k neighbors $\{n_1, \ldots, n_k\}$ from the training index. A fixed, policy-first prompt injects the query and retrieved examples. The model outputs a single label $\hat{y}(q) \in \{\texttt{FLAGGED}, \texttt{NOT FLAGGED}\}$ at temperature 0.

Pipeline differences:

- **Pipeline A (Commercial)** indexes examples using API embeddings and performs hosted inference (deterministic).
- **Pipeline B (Open-source)** indexes examples using a Hugging Face sentence embedder with FAISS retrieval, and performs inference locally using `Power-LLaMA-3-7B-Instruct` (quantized GGUF) via `llama.cpp`.

Crucially, both pipelines share the *same* split, label mapping, evaluation code, prompt structure, and deterministic settings, isolating the comparison to a realistic "stack choice" rather than experimental noise.

3.4. **Operational Metrics (Recall-First Evaluation)**

Our primary safety metric is sensitive-class recall:

$$\text{Recall}_{\texttt{FLAGGED}} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FNR} = 1 - \text{Recall}_{\texttt{FLAGGED}}.$$

To make the moderation trade-off explicit, we also report:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{ReviewRate} = \frac{\text{TP} + \text{FP}}{N},$$

We include accuracy and macro/weighted F1 strictly as secondary context, since high accuracy can coexist with unacceptable miss rates in imbalanced moderation.

3.5. **Uncertainty, Significance, and Subgroup Analysis**

We quantify uncertainty using nonparametric bootstrap confidence intervals (95%) over the test set for key metrics (accuracy, macro-F1, $\text{Recall}_{\texttt{FLAGGED}}$, and FPR). We test the difference between pipelines using McNemar's test on paired predictions.

To address fairness and realism concerns, we compute subgroup metrics using Jigsaw identity columns when present. For each identity attribute a, we define subgroup membership as $\mathbb{I}[a \geq 0.5]$ and report subgroup $\text{Recall}_{\texttt{FLAGGED}}$, FPR, and review rate for all groups with $n \geq 50$.

4. **Experiments and Results**

4.1. **Experimental Setup: Jigsaw Unseen-Article Split**

We construct a leakage-resistant split using `article_id` as the grouping key (Section 3.2). From the full Jigsaw corpus (1,804,874 rows), we create an unseen-article test set of approximately 10% and then cap to:

$$|\mathcal{D}_{\text{train}}| = 80{,}000, \quad |\mathcal{D}_{\text{test}}| = 8{,}000.$$

Label priors remain stable across splits (both $\approx 8\%$ `FLAGGED`). In the capped test set, `FLAGGED` support is 637 and `NOT FLAGGED` support is 7363.

4.2. **Main Result: Recall–Review Trade-off on Unseen Articles**

Table 1 summarizes the operational performance of both pipelines on the unseen-article test set. The key finding is a clear *Pareto trade-off*: Pipeline B dramatically reduces missed unsafe content (high $\text{Recall}_{\texttt{FLAGGED}}$), but at the cost of substantially higher false positives and review burden; Pipeline A induces far lower review load and achieves higher overall accuracy, but misses more unsafe items.

Confusion-level interpretation. Pipeline A yields TP = 402, FN = 235, FP = 1677, TN = 5686. Pipeline B yields TP = 575, FN = 62, FP = 4331, TN = 3032. Thus, relative to Pipeline A, Pipeline B recovers +173 additional true `FLAGGED` items, but incurs +2654 additional false positives and increases the review queue by +2827 items.

Equivalently, switching from A→B costs approximately $\frac{2654}{173} \approx 15.3$ additional false positives per additional true positive.

Table 1. Operational metrics on Jigsaw unseen-article test set ($n = 8000$). Systems-level comparison showing operating-point trade-off.

Metric	A (Commercial)	B (Open-source)
Accuracy	0.7610	0.4509
$F1_{macro}$	0.5760	0.3937
Safety:		
$Recall_{\texttt{FLAGGED}}$	0.6311	0.9027
FNR	0.3689	0.0973
Cost:		
FPR	0.2278	0.5882
ReviewRate	0.2599	0.6132

Additionally, it costs approximately $\frac{2827}{173} \approx 16.3$ additional reviews per additional true positive. This marginal-cost view makes explicit that high recall can be achieved by conservative flagging, but the resulting review burden may be operationally prohibitive without calibrated thresholds or secondary filters.

4.3. **Baseline Comparisons**

To contextualize our RAG pipelines against similar state-of-the-art approaches, we evaluate several baselines on the same unseen-article test set. Table 2 reports results for: (i) *threshold-based* classifiers using the raw Jigsaw toxicity score at various operating points, which serve as oracle upper bounds; (ii) *kNN* (k=5) and *logistic regression* classifiers trained on MiniLM sentence embeddings; (iii) trivial baselines (majority class, random, always-flag); and (iv) *Class-RAG*, a retrieval-augmented baseline that retrieves balanced examples (2 safe + 2 unsafe) from separate FAISS indices and augments the LLM prompt with explanations for each retrieved example, as described in [2]

The threshold@0.50 baseline achieves perfect performance by construction, since labels are derived from the same score—it represents the ceiling if ground-truth toxicity scores were available at inference. Among text-only methods, Pipeline B achieves the highest recall (0.903), substantially outperforming Class-RAG (0.767), logistic regression (0.694), and kNN (0.124). However, this recall comes at considerable cost: Pipeline B's FPR (0.588) far exceeds that of Class-RAG (0.285) and logistic regression (0.199).

Class-RAG occupies an intermediate position: it achieves higher recall than both Pipeline A and logistic regression (0.767 vs. 0.631 and 0.694), but with moderately elevated FPR (0.285). This suggests that the balanced retrieval strategy—retrieving equal numbers of safe and unsafe examples—provides some recall benefit over unbalanced top-k retrieval, though it does not match the aggressive flagging behavior of Pipeline B. Notably, logistic regression achieves competitive recall (0.694) with the lowest FPR among non-oracle methods (0.199), suggesting that simpler supervised classifiers remain strong baselines for this task.

These comparisons highlight that Pipeline B occupies a distinct Pareto-efficient operating point—useful when maximizing recall justifies high review burden—while Class-RAG offers a middle-ground trade-off between recall and review cost. The results also reveal that retrieval-augmented LLM pipelines do not uniformly dominate traditional classifiers across all metrics.

4.4. **Uncertainty and Statistical Significance**

Bootstrap 95% confidence intervals (CI) and McNemar's test confirm that these differences are not attributable to noise:

Table 2. Baseline comparison on Jigsaw unseen-article test set (n=8000). Threshold-based methods use the raw toxicity score (oracle access). kNN and LogReg use MiniLM embeddings trained on 10k samples. Class-RAG uses DRAGON-RoBERTa embeddings with balanced retrieval (2 safe + 2 unsafe examples) and local LLM inference. Methods sorted by $\text{Recall}_{\texttt{FLAGGED}}$.

Method	**Acc**	**Recall**$_F$	**FPR**	**Review**	**F1**$_{\text{macro}}$
Th@0.50	1.00	1.00	0.00	0.08	1.00
Th@0.40	0.96	1.00	0.03	0.11	0.90
Pipeline B	0.45	**0.90**	0.58	0.61	0.39
Class-RAG	0.72	0.76	0.28	0.32	0.45
LogReg	0.79	0.69	0.19	0.23	0.61
Pipeline A	0.76	0.63	0.22	0.26	0.57
kNN (k=5)	0.91	0.12	0.01	0.0	0.57
Majority	0.92	0.00	0.00	0.0	0.47

Table 3. Subgroup metrics on unseen-article test set (identity score ≥ 0.5, $n \geq 50$). We report `FLAGGED` recall and FPR.

Group	n	**Rec**$_{\texttt{F}}$ **(A)**	**FPR (A)**	**Rec**$_{\texttt{F}}$ **(B)**	**FPR (B)**
male	196	0.857	0.453	0.971	0.677
female	247	0.892	0.495	0.946	0.681
christian	188	0.867	0.370	1.000	0.601
muslim	107	0.889	0.713	0.889	0.825
black	56	0.875	0.675	0.875	0.725
white	116	1.000	0.718	0.936	0.894

- Pipeline A: Acc $\in (0.7519, 0.7705)$, $\text{F1}_{\text{macro}} \in (0.5634, 0.5885)$, $\text{Recall}_{\texttt{FLAGGED}} \in (0.5934, 0.6672)$, FPR $\in (0.2184, 0.2371)$.
- Pipeline B: Acc $\in (0.4408, 0.4614)$, $\text{F1}_{\text{macro}} \in (0.3839, 0.4036)$, $\text{Recall}_{\texttt{FLAGGED}} \in (0.8788, 0.9253)$, FPR $\in (0.5774, 0.5997)$.

McNemar's test on paired predictions yields $b_{01} = 2912$ and $b_{10} = 431$ with $p \approx 0$, indicating a decisive difference in error patterns between stacks.

4.5. **Subgroup Analysis: Identity-Associated Review Burden**

A major advantage of Jigsaw is the availability of identity annotations. Table 3 reports subgroup metrics for identity attributes with $n \geq 50$ (membership defined as attribute score ≥ 0.5). Both pipelines exhibit elevated false positive rates for identity-associated text, consistent with known unintended-bias behavior in toxicity systems; Pipeline B, while recall-strong, amplifies this effect via higher overall flagging.

These subgroup patterns clarify an operational reality: aggressively increasing recall via broad flagging can disproportionately route identity-associated benign text to review. This motivates budget-aware calibration (e.g., fixing a maximum review rate and optimizing recall subject to that constraint), which we treat as a key direction for follow-up experiments.

4.6. **Retrieval Diversity Controls the Recall–Review Operating Point**

A central claim of recall-first moderation is that *operating points matter*: in deployment, teams select a safety–cost trade-off (missed unsafe content vs. human review burden), rather than optimizing a single scalar metric. To make this claim concrete within our RAG protocol,

Table 4. Jigsaw medium sweep (B+ only; 20k/2k capped unseen-article protocol). Retrieval strategy moves the system along the recall–review frontier: graph-based retrieval substantially improves flagged recall but increases review burden and FPR. Metrics are stable across thresholds and committee sizes in this regime.

Retriever	k	t	Acc ↑	Rec_F ↑	Review ↓
MMR	5	0.40–0.70	0.714	0.660	0.312
Graph	5	0.40–0.70	0.504	0.855	0.553
MMR	3	0.40–0.60	0.714	0.660	0.312
MMR	3	0.70	0.715	0.654	0.310
Graph	3	0.40–0.70	0.503–0.505	0.855	0.552–0.553

we ran a focused sweep that isolates the effect of **retrieval diversity strategy** (holding the overall pipeline structure fixed).

We used the same `group_article_id_stratified` split construction to prevent article overlap between train and test, preserving the class prior (about 8% FLAGGED). We capped the split for tractable end-to-end evaluation (20k train / 2k test), then indexed the full 20k training set into a graph store (20,000 nodes; 153,845 edges). We evaluated Pipeline B+ under two retrieval strategies: (i) **MMR** diversification and (ii) **graph-based** selection, with committee sizes $k \in \{3, 5\}$ and decision thresholds $t \in \{0.40, 0.50, 0.60, 0.70\}$.

Across all thresholds and both committee sizes, the retrieval strategy dominates the operating point. Graph-based retrieval achieves substantially higher sensitive-class recall than MMR:

$$\text{Recall}_{\text{FLAGGED}}(\text{graph}) = 0.855 \quad \text{vs.} \quad \text{Recall}_{\text{FLAGGED}}(\text{MMR}) = 0.660$$

a $+$**19.5 point** absolute improvement. This gain comes with markedly higher burden:

$$\text{ReviewRate}(\text{graph}) = 0.553 \quad \text{vs.} \quad \text{ReviewRate}(\text{MMR}) = 0.312,$$

and higher false positives:

$$\text{FPR}(\text{graph}) = 0.527 \quad \text{vs.} \quad \text{FPR}(\text{MMR}) = 0.281.$$

As expected under heavy conservative flagging, overall accuracy drops (0.504 vs. 0.714). Importantly, this is not a failure mode of the protocol; it is precisely the *operational trade-off* recall-first moderation must surface and quantify.

Within the tested range, changing the decision threshold t does not meaningfully shift metrics, and committee size $k \in \{3, 5\}$ produces negligible differences. This suggests that, for this configuration, **retrieval selection is the primary control knob** for recall–review behavior, more so than small threshold adjustments. Practically, it implies that if we need to meet a fixed review budget, we should (i) introduce explicit budget-aware calibration and/or (ii) add a secondary filter stage, rather than relying on minor threshold tuning.

4.7. **Component Ablations Study**

In order to evaluate the contributions of each design element within the Methods section, we conduct a structured ablation study across its three core modules: (i) distribution-preserving contrastive augmentation (**A**), (ii) committee-based retrieval (**R**), and (iii) recall-oriented decision policy (**P**).

This analysis holds constant the dataset split, backbone architecture, retrieval budget, prompt formulation, decoding strategy, embedding model, and retrieval index. Systematic modifications are applied exclusively to the specified component(s) in each experimental condition, thereby isolating their individual and combined effects on overall system performance.

Table 5. **Component ablation.** Six configurations at N=2000 and N=16000. **P** (rows 3, 5) maximizes recall (≈0.94) with scale-invariant gains. **R** (rows 2, 4) reduces FPR and review burden, with effects emerging at scale. **A**+**R** (row 4) achieves the lowest review burden (0.385) and highest precision (0.163) at N=16000.

				$N = 2000$					$N = 16000$				
Row	A	R	P	Rec_F	FPR	FN	Prec_F	Review	Rec_F	FPR	FN	Prec_F	Review
0	55	55	55	0.874	0.590	20	0.113	0.613	0.897	0.592	131	0.116	0.616
1	51	55	55	0.874	0.538	20	0.123	0.565	0.881	0.558	152	0.120	0.584
2	55	51	55	0.874	0.570	20	0.117	0.595	0.844	0.416	199	0.149	0.450
3	55	55	51	0.931	0.682	11	0.105	0.702	0.945	0.668	70	0.109	0.690
4	51	51	55	0.874	0.609	20	0.110	0.630	0.789	0.350	269	0.163	0.385
5	51	55	51	0.931	0.685	11	0.105	0.705	0.944	0.674	71	0.108	0.695

Threshold calibration (P). For rows enabling **P**, decision thresholds are calibrated on a held-out dev split only (never on test). In the current ablation runs, we select a threshold from a fixed grid and choose the operating point that maximizes $\text{Recall}_\text{Flagged}$ on dev under the same procedure for all **P**-enabled rows. (We report the chosen threshold in the logs; for all three sample sizes shown here, the selected threshold is $t = 0.05$.)

Extended results across $N \in \{2000, 4000, 8000, 16000, 32000\}$ for baseline, R-only, and P-only configurations are provided in Appendix B, confirming that P effects are stable across all scales while R effects grow monotonically with training set size.

Observed effects (partial grid). Across three sample sizes, enabling **P** (Row 3) consistently increases $\text{Recall}_\text{Flagged}$ (e.g., ≈ 0.94 at N=4000 and N=8000) while substantially increasing FPR and review rate, reflecting the intended recall-first behavior when thresholds are tuned to maximize recall. Enabling **R** alone (Row 2) exhibits sample-dependent trade-offs: at N=8000, it reduces FPR (0.5280 vs. 0.5939) and lowers review rate (0.5555 vs. 0.6178), improving precision (0.1251 vs. 0.1151), but can reduce recall relative to baseline in exchange for fewer false positives.

Individual vs combined component ablation (experiments with augmentation). To isolate individual and combined component effects, we conduct an ablation over augmentation (**A**), committee retrieval (**R**), and recall-oriented policy (**P**) at two scales: N=2000 and N=16000.

Scale-dependent component effects. Analysis across both scales reveals three consistent patterns.

(i) ***P*** *is scale-invariant.* Policy calibration (rows 3, 5) achieves $\text{Recall}_\text{F} \approx 0.93$–$0.94$ at both N=2000 and N=16000, reducing false negatives by 45–47% relative to baseline (e.g., FN: 131 → 70 at N=16000). The associated cost-elevated FPR (≈0.67–0.68) and ReviewRate (≈0.69–0.70) is also consistent across scales. This indicates that the recall benefit of aggressive threshold calibration does not depend on training set size.

(ii) ***R*** *effects emerge with scale.* At N=2000, committee retrieval produces negligible FPR reduction (row 2: 0.590 → 0.570). At N=16000, the same configuration yields substantial gains: FPR drops from 0.592 to 0.416, and ReviewRate drops from 0.616 to 0.450, while precision increases from 0.116 to 0.149. This suggests that diverse retrieval requires a sufficiently large index to surface informative neighbors.

(iii) ***A*** *amplifies* ***R*** *at scale.* Augmentation alone (row 1) provides modest, consistent FPR reduction (≈0.03–0.05 points) at both scales without affecting recall. Combined with retrieval (row 4), augmentation amplifies selectivity gains at scale: at N=16000, A+R achieves the lowest FPR (0.350), lowest ReviewRate (0.385), and highest precision (0.163) among all configurations. At N=2000, where R alone has minimal effect, A+R similarly

shows no benefit—confirming that augmentation enhances retrieval diversity rather than operating independently.

Practical implications. These results support a *configuration selection* view of recall-first moderation:

- **Maximize recall (safety-critical):** Enable **P** (rows 3 or 5). Recall ≈0.94 is achieved immediately, regardless of training set size.
- **Minimize review burden (resource-constrained):** Enable **R** or **A**+**R** (rows 2 or 4) with a sufficiently large training index. At N=16000, A+R reduces ReviewRate to 0.385 while maintaining precision at 0.163.

The components control orthogonal axes of the recall–precision–cost frontier: **P** targets missed harms (FN), while **R**/**A**+**R** targets unnecessary escalations (FPR, ReviewRate). This orthogonality enables deployment-specific tuning without requiring a single "best" configuration.

5. Conclusion

We studied *recall-first* content moderation, where the dominant safety risk is a false negative (unsafe content predicted safe) and where systems are permitted to escalate uncertain cases for human review. To improve scale and generalizability realism, we conducted our headline evaluation on the Jigsaw Unintended Bias in Toxicity Classification corpus using a leakage-resistant *unseen-article* split. This protocol better reflects deployed novelty by ensuring that no `article_id` appears in both train and test, while preserving label priors in a large-scale setting.

Across two deployment-realistic retrieval-augmented classification stacks, we observe a clear operational trade-off between safety recall and review burden. Pipeline A (commercial stack) achieves higher overall utility (Accuracy 0.761) at a substantially lower review rate (0.260), but with lower sensitive-class recall ($\text{Recall}_{\texttt{FLAGGED}} = 0.631$). Pipeline B (open-source stack) reaches substantially higher $\text{Recall}_{\texttt{FLAGGED}}$ (0.903) and a much lower miss rate (FNR 0.097), but does so by flagging far more content (ReviewRate 0.613) and increasing false positives (FPR 0.588). Bootstrap confidence intervals and McNemar's test confirm that these differences are statistically robust under paired evaluation.

Baseline comparisons provide critical context for these results. Among text-based methods operating without access to ground-truth toxicity scores, Pipeline B achieves the highest recall (0.903), outperforming state-of-the-art RAG systems (0.767), logistic regression (0.694) and kNN (0.124) by substantial margins. However, simpler supervised classifiers are not uniformly dominated: logistic regression achieves higher recall than Pipeline A (0.694 vs. 0.631) with lower FPR (0.199 vs. 0.228), indicating that RAG-based LLM inference does not guarantee improvements over well-tuned traditional classifiers. Pipeline B thus occupies a distinct Pareto-efficient operating point, valuable when maximizing recall justifies high review burden, while Pipeline A's value proposition requires further investigation. A focused retrieval-diversity sweep isolates the mechanism underlying this trade-off. Holding the pipeline structure fixed, graph-based retrieval achieves $\text{Recall}_{\texttt{FLAGGED}} = 0.855$ versus 0.660 for MMR diversification—a +19.5 point gain—while increasing ReviewRate from 0.312 to 0.553. Notably, decision threshold and committee size produce negligible variation in this regime, indicating that *retrieval selection is the primary control knob* for recall–review behavior. This mechanistic finding directly motivates budget-aware calibration: practitioners can select retrieval diversity to optimize recall subject to a maximum tolerable review rate.

A component ablation over the three core components—distribution-preserving augmentation (**A**), committee retrieval (**R**), and recall-oriented policy (**P**)—reveals that these components control orthogonal axes of the recall–precision–cost frontier, with scale-dependent effects. Policy calibration (**P**) provides immediate, scale-invariant recall gains: $\text{Recall}_{\text{F}} \approx 0.94$

at both N=2000 and N=16000, reducing false negatives by 45–47%. Committee retrieval (**R**) reduces review burden, but only at sufficient scale: FPR drops from 0.592 to 0.416 at N=16000, while showing negligible effect at N=2000. Augmentation (**A**) amplifies retrieval selectivity at scale, with A+R achieving the lowest review burden (0.385) and highest precision (0.163). These findings support a *configuration selection* view: practitioners choose **P** when minimizing missed harms is paramount, or **R**/**A**+**R** when review capacity is the binding constraint. A key advantage of the Jigsaw setting is that it enables subgroup analysis using identity annotations. We find that both stacks can incur elevated false positive rates for identity-associated text, and that aggressively increasing recall via broad flagging can amplify review burden in these subgroups. This highlights that recall-first moderation should be evaluated jointly with operational constraints (review capacity) and fairness-sensitive metrics, not accuracy alone. Future work should investigate the interaction between retrieval and policy calibration at scale, operationalize budget-aware threshold selection (optimize $\text{Recall}_{\texttt{FLAGGED}}$ subject to a maximum review rate), and extend evaluation to multilingual and cross-platform corpora.

Limitations. Our headline moderation evaluation is centered on a single English-language benchmark (Jigsaw), so generalization to multilingual, cross-platform, or domain-shifted settings remains unverified. The capped splits ($|\mathcal{D}\text{train}|$=80k, $|\mathcal{D}\text{test}|$=8k) make tractable end-to-end LLM evaluation possible, but may reduce retrieval-index diversity and limit the strength of scaling conclusions. The A/B comparison is systems-level: observed differences reflect full-stack interactions rather than isolated component effects.

Acknowledgements

A.SM. acknowledges the financial support from Secretaría de Ciencia Humanidades Tecnología e Innovación (SECIHTI, Mexico)for funding his Ph.D. program.

References

[1] T. Huang. "Content moderation by LLM: from accuracy to legitimacy". In: *Artificial Intelligence Review* 58 (2025), p. 320. DOI: 10.1007/s10462-025-11328-1. URL: https://doi.org/10.1007/s10462-025-11328-1.

[2] J. Chen, E. Shen, T. Bavalatti, X. Lin, Y. Wang, S. Hu, H. Subramanyam, K. S. Vepuri, M. Jiang, J. Qi, L. Chen, N. Jiang, and A. Jain. *Class-RAG: Content Moderation with Retrieval Augmented Generation*. 2024. arXiv: 2410.14881 [cs.CL]. URL: https://arxiv.org/abs/2410.14881.

[3] OpenAI. "GPT-4 for Content Moderation". In: *OpenAI Blog* (2023). URL: https://openai.com/index/using-gpt-4-for-content-moderation/.

[4] I. Nguyen, H. Suresh, and E. Shieh. *Representational Harms in LLM-Generated Narratives Against Nationalities Located in the Global South*. HEAL Workshop, CHI 2025. Received 24 February 2025; Accepted 27 March 2025. 2025. URL: https://heal-workshop.github.io/chi2025_papers/50_Representational_Harms_in_L.pdf.

[5] T. Huang. *Content Moderation by LLM: From Accuracy to Legitimacy*. arXiv:2409.03219v2 (submitted 5 September 2024; version 2, 1 June 2025). 2024. arXiv: 2409.03219 [cs.CY]. URL: https://arxiv.org/abs/2409.03219.

[6] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, and S. Joty. *Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges*. Submitted 5 March 2024. 2024. arXiv: 2403.02990v1 [cs.CL]. URL: https://arxiv.org/html/2403.02990v1.

[7] H. He and E. A. Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.

[8] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen. "A survey on imbalanced learning: latest research, applications and future directions". In: *Artificial Intelligence Review* 57 (2024), p. 137. DOI: 10.1007/s10462-024-10759-6. URL: https://link.springer.com/article/10.1007/s10462-024-10759-6.

[9] Z. Papakipos and J. Bitton. *AugLy: Data Augmentations for Robustness*. 2022. arXiv: 2201.06494 [cs.AI]. URL: https://arxiv.org/abs/2201.06494.

[10] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. *A Survey of Data Augmentation Approaches for NLP*. Accepted to ACL 2021 Findings. 2021. arXiv: 2105.03075 [cs.CL]. URL: https://arxiv.org/abs/2105.03075.

[11] D. Kaushik, E. Hovy, and Z. C. Lipton. *Explaining The Efficacy of Counterfactually-Augmented Data*. Published at ICLR 2021. 2020. arXiv: 2010.02114 [cs.CL]. URL: https://arxiv.org/abs/2010.02114.

[12] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Long Papers*. Published as a long paper at ACL 2022. 2022. arXiv: 2203.09509 [cs.CL]. URL: https://arxiv.org/abs/2203.09509.

[13] F. Zhang, X. Jin, J. Cheng, H. Yu, and H. Xu. "Rethinking the Role of LLMs for Document-level Relation Extraction: a Refiner with Task Distribution and Probability Fusion". In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-Long Papers), Volume 1: Long Papers*. USA: Association for Computational Linguistics, 2025, pp. 6293–6312. URL: https://aclanthology.org/2025.naacl-long.319.pdf.

[14] F. Chernogorskii, S. Averkiev, L. Kudraleeva, Z. Martirosian, M. Tikhonova, V. Malykh, and A. Fenogenova. *DRAGON: Dynamic RAG Benchmark On News*. 2025. arXiv: 2507.05713 [cs.CL]. URL: https://arxiv.org/abs/2507.05713.

[15] M. R. Rezaei and A. B. Dieng. *Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs*. 2025. arXiv: 2502.11228 [cs.CL]. URL: https://arxiv.org/abs/2502.11228.

[16] M. Kolla, S. Salunkhe, E. Chandrasekharan, and K. Saha. "LLM-Mod: Can Large Language Models Assist Content Moderation?" In: *CHI 2024 - Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI EA '24. Honolulu, HI, USA: Association for Computing Machinery, May 2024. DOI: 10.1145/3613905.3650828. URL: https://doi.org/10.1145/3613905.3650828.

[17] M. Franco, O. Gaggi, and C. E. Palazzi. "Integrating Content Moderation Systems with Large Language Models". In: *ACM Transactions on the Web* 19.2 (2025). DOI: 10.1145/3700789. URL: https://doi.org/10.1145/3700789.

[18] G. Izacard and E. Grave. *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering*. 2020. arXiv: 2007.01282 [cs.CL]. URL: https://arxiv.org/abs/2007.01282.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[20] J. G. Carbonell, J. Goldstein, and F. n. i. a. Stewart. "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries". In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. 1998, pp. 335–336. DOI: 10.1145/290941.291025. URL: https://doi.org/10.1145/290941.291025.

[21] J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

Appendix A. Distribution-Preserving Augmentation Validation

We aim to generate label-preserving paraphrases that remain *statistically indistinguishable* from the original texts. To achieve this, indistinguishability is enforced at three complementary levels: global surface form, measured via word and character histograms; global semantic mixture proportions, captured in a compact embedding space; and local per-cluster consistency, where form is conditioned on latent semantics. Full algorithmic details are given in Appendix Section A. The overall augmentation and validation workflow is illustrated in Figure 2, which shows how embedding, clustering, constrained paraphrasing, and multi-level validation are combined to ensure distribution-preserving generation.

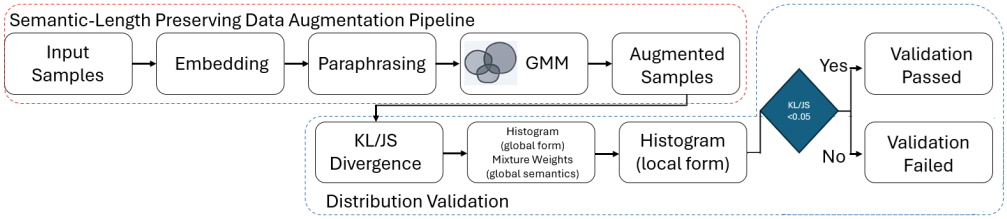

Figure 2. **Overview of the data augmentation and validation workflow**: paraphrased samples are generated under semantic- and structure-preserving constraints and validated through multi-level KL/JS divergence checks to ensure distributional fidelity.

The augmentation process begins by embedding texts, reducing dimensionality, and clustering them with a Gaussian Mixture Model (GMM). Augmented samples are then produced to satisfy integer quotas defined over the joint *class*×*component* distribution, ensuring that both label priors and latent structure are preserved. To validate that our augmentation procedure generates samples statistically indistinguishable from the original distribution, we evaluate on two auxiliary corpora: therapeutic chatbot responses ($n = 80$) and resumes ($n = 125$). Table 6 reports JS divergence across three complementary views: global word/character histograms, GMM mixture weights in embedding space, and per-component histograms.

Table 6. JS divergence validation for distribution-preserving augmentation. Global metrics remain well below the $\tau_{\text{JS}} = 0.05$ threshold. Per-component deviations in low-support clusters reflect small-sample instability rather than semantic drift.

Dataset	$\mathbf{JS}(H_{\mathbf{words}})$	$\mathbf{JS}(H_{\mathbf{chars}})$	$\mathbf{JS}(w)$	$\mathbf{max}_c\ \mathbf{JS}^{(c)}_{\mathbf{words}}$	$\mathbf{max}_c\ \mathbf{JS}^{(c)}_{\mathbf{chars}}$
Therapy	7.5×10^{-3}	4.0×10^{-3}	0.000	9.37×10^{-2}	9.98×10^{-2}
Resumes	6.0×10^{-4}	4.5×10^{-3}	5.3×10^{-4}	1.15×10^{-2}	6.20×10^{-1}

Global divergences (columns 2–4) remain well below $\tau_{\text{JS}} = 0.05$ for both corpora, confirming that augmented distributions are nearly identical to their originals at the corpus level. The elevated per-component character JS for resumes (0.62) occurs in clusters with very low support ($n < 5$), attributable to small-sample instability rather than substantive distributional drift. These results validate that our augmentation procedure successfully generates paraphrases that preserve the statistical structure of the original data.

Appendix B. Scaling Behavior Across Sample Sizes

Table 7 reports baseline (Row 0), retrieval-only (Row 2), and policy-only (Row 3) across five sample sizes, complementing the component ablation in Table 5.

Row	A	R	P	N	$\text{Recall}_{\text{Flagged}} \uparrow$	FPR ↓	FN ↓	$\text{Precision}_{\text{Flagged}} \uparrow$	ReviewRate ↓	Calls / 1k ↓
0	55	55	55	2000	0.8742	0.5904	20	0.1134	0.6130	1000
2	55	51	55	2000	0.8742	0.5703	20	0.1169	0.5945	5000
3	55	55	51	2000	0.9308	0.6822	11	0.1054	0.7020	1000
0	55	55	55	4000	0.8589	0.5849	45	0.1129	0.6068	1000
2	55	51	55	4000	0.9154	0.6058	27	0.1158	0.6305	5000
3	55	55	51	4000	0.9404	0.6724	19	0.1081	0.6938	1000
0	55	55	55	8000	0.8933	0.5939	68	0.1151	0.6178	1000
2	55	51	55	8000	0.8728	0.5280	81	0.1251	0.5555	5000
3	55	55	51	8000	0.9403	0.6696	38	0.1083	0.6911	1000
0	55	55	55	16000	0.897174	0.591742	131	0.115958	0.616062	1000
2	55	51	55	16000	0.843799	0.415659	199	0.149389	0.449750	5000
3	55	55	51	16000	0.945055	0.667731	70	0.109088	0.689813	1000
0	55	55	55	32000	0.894819	0.589909	268	0.116007	0.614187	1000
2	55	51	55	32000	0.791994	0.352098	530	0.162900	0.387125	5000
3	55	55	51	32000	0.936077	0.657698	140	0.114128	0.676844	1000

Table 7. **Component ablations:** partial factorial ablations over augmentation (**A**), committee retrieval (**R**), and recall-oriented policy (**P**). Rows shown for $N \in \{2000, 4000, 8000, 16000, 32000\}$ correspond to completed run logs: baseline (Row 0), retrieval-only (Row 2), and policy-only (Row 3). We report safety-critical metrics ($\text{Recall}_{\text{Flagged}}$, FN, FPR) plus operational burden (review rate) and cost proxy (calls per 1k items).