Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

C&D SENTINEL: AN ENTERPRISE-GRADE MULTIMODAL CONTENT MODERATION SYSTEM USING HYBRID TWO-STAGE NLP AND COMPUTER VISION

K. SHAILAJA[1], T. JALAJA[2], M. CHAITANYA[3], V. DARPAD SAI[4]

[1] Associate Professor, Vasavi College of Engineering (Autonomous) Hyderabad, Department of Computer Science & Engineering, India.

[2] Assistant Professor, Vasavi College of Engineering (Autonomous) Hyderabad, Department of Computer Science & Engineering, India.

[3,4] Student, Vasavi College of Engineering (Autonomous) Hyderabad, Department of Computer Science & Engineering, India.

E-mail: [1]e.shailaja@staff.vce.ac.in, [2]jalaja.t@staff.vce.ac.in, [3]maddikatlachaitanya@gmail.com, [4]darpadsai1805@gmail.com

ABSTRACT

The rapid increase of user-generated content through digital means has resulted in an urgent requirement for scalable, accurate and relevant content moderation systems. The C&D Sentinel content moderation pipeline solves this issue with an enterprise-class hybrid architecture consisting of locally deployed machine learning models and cloud-based large language model (LLM) reasoning. The system identifies toxic, hateful and unsafe content across a range of formats, including text, still images, and video, while providing low latency. Traditional moderation systems rely primarily on manual keyword filtering or opaque "black box" classifiers that cannot interpret contextual differences between words; these methods often mislabel harmless items – such as satire, slang, and artistic forms – as dangerous, resulting in numerous false positive instances. This paper demonstrates that using a Hybrid Two-Stage Architecture combining light weight local models and advanced reasoning capabilities provided by cutting-edge LLMs will resolve these limitations. DistilBERT, an example of such an efficient model, can perform quick text classification, while BLIP (Bootstrapping Language-Image Pretraining) will create captions to image content effectively converting visual data into an appropriate textual form for semantic analysis.

The Llama 3 70B model is utilized within the system to provide fast inference with low latency with integration onto the Groq LPU (host machine) allowing for improved detection of malicious behaviour and amending false positives compared to previous methodologies for detection of malicious behaviour in context. One of the implementations of this system is its video moderation, which is processed in a dual-stream framework where images are captured using OpenCV for visual features and where audio is converted to text using Whisper. These two streams are analysed together to allow a greater understanding of what the video is and therefore whether or not the video contains any HA. A second component of the system is an RLHF (reinforcement learning from human feedback) feedback loop, which allows for continual improvement of the content detection capabilities through a human-in-the-loop process. RLHF will allow for continual enhancement of detecting new types of language and harmful content as they emerge. The system was evaluated against the Facebook Hateful Memes Challenge data set, where it achieved 90% accuracy, far better than a zero-shot baseline using CLIP that achieved only 14% accuracy. The system is also deployed to Groq LPU hardware which ensures that the system can perform real-time analyse with less than 200 milliseconds of inference latency, making the system well-suited for enterprise content moderation. Overall, the C&D Sentinel pipeline has demonstrated a significant improvement in terms of accuracy, lower false positive rates and real-time performance capabilities to develop a robust and scalable platform for next-generation content moderation.

Keywords: *Content Moderation, Retrieval-Augmented Generation, DistilBERT, Multimodal AI, Reinforcement Learning from Human Feedback, Groq LPU, BLIP, Toxicity Detection*

1. INTRODUCTION

The overwhelming volume and exponential growth of user-generated content across different social networks, online communities, and enterprise communications platforms has led to an increasing urgency to find viable engineering solutions for automated content moderation. The number of posts published globally on social media channels each day exceeds 2 billion, and the current methods of manually reviewing this content will not be able to meet the demand required. Further exacerbating this challenge is that current automated approaches for moderating content all have large-scale system deficiencies that make them unreliable in actual production use.

Keyword filtering mechanisms are too easily bypassed—by altering word choice, using deliberate misspellings, or by changing the context in which they are used—to provide any kind of real assurance. Concurrently, the use of convolutional neural networks (CNN) for image moderation generates unacceptably high false positives as a result of the texture and color biases present in the samples used to train CNNs. For example, a CNN may label a photo of a sunset as violent because the colors of reddish-orange pixels fall within similar ranges of some sample images that were labelled as violent. Additionally, text-only systems may label gaming expressions (such as "destroy them!") as threats and block them, even though this is standard language used within a competitive sport context.

This is not merely an incremental engineering concern: recent large-scale evaluations of off-the-shelf moderation solutions confirm that both purpose-built moderation APIs and general-purpose LLMs exhibit systematic, category-dependent weaknesses when applied without architectural adaptation. AlDahoul et al. [18] benchmarked OpenAI's omni-moderation model [20], Meta's Llama-Guard-3 [21], GPT-4o, Gemini 1.5, and Llama-3 across text, image, and video datasets, and reported that pre-trained moderation APIs achieved as little as 67% recall on nudity detection and near-negligible accuracy on alcohol-, drug-, and child-abuse-related image content, despite comparatively strong performance from general-purpose LLMs on the same data. Jahan and Oussalah [19], in a systematic review of hate speech detection literature, similarly documented that automatic detection systems remain vulnerable to context-dependent misclassification and that the literature lacks standardised protocols for evaluating real-time deployability. These findings motivate the problem addressed in this paper: existing solutions individually address either contextual accuracy (general-purpose LLMs) or production-grade throughput (lightweight classifiers), but no single deployed architecture reported in the literature has jointly validated both properties across text, image, and video modalities under real-time latency constraints.

Recent advances in Retrieval-Augmented Generation (RAG) for large language models have demonstrated the capacity to resolve precisely this class of contextual ambiguity [2]. Building on these advances, C&D Sentinel implements a two-stage hybrid pipeline designed to deliver low-latency, high-throughput performance while maintaining high semantic fidelity. Fast local models handle the bulk of content routing decisions, while LPU-accelerated LLM reasoning is reserved exclusively for genuinely ambiguous edge cases.

1.1 Background Information

The evolution of content moderation technology reflects the evolution of research into natural language processing. Initial implementations included only the use of regular expressions with static block lists, which allowed for easy obfuscation via minimal paraphrase. The second wave of solutions utilized word embeddings (Word2Vec, GloVe) to create similarity-based matching that offered improved resistance to evasion based upon exact matching but still could not adequately encode sentence-level semantics.

Utilizing a transformer-based architecture, comprised of models such as BERT [4] and DistilBERT [7] (the latter being a distilled version of BERT), early adopters saw dramatic improvements in hate speech and toxicity detection classifying based upon the classification benchmarks of these models, thanks to the use of bidirectional contextual encoding. However, whilst BERT and DistilBERT are capable of generating deterministic classifier scores, those scores do not take into account an understanding of situational, cultural, or ironic contexts. Additionally, large language models (e.g., Llama-3) possess emergent reasoning capabilities that can help to ameliorate these situations; however, they also suffer from the possibility of their latency and inference costs presenting a barrier to being used solely as a classification engine within high-throughput real-time systems.

C&D Sentinel presents an architecture that resolves the trade-offs associated with choosing to utilize

large language models (LLMs) as part of real-time moderation systems. The proposed system uses a two-tiered routing architecture, where the first tier of the solution utilizes a fine-tuned distil BERT model to generate a continuous toxicity score. In this manner, only those instances of content that reside within an empirically defined ambiguity range are forwarded to the LLM reasoning stage. Thus, the costs associated with LLM cloud-based inference are constrained by only processing select content, resulting in keeping overall system latency within moderation bounds.

1.2 Rationale

The core challenge addressed by this research is the simultaneous achievement of high precision (minimising false positives that suppress legitimate speech), high recall (minimising false negatives that permit harmful content), and real-time throughput (sub-200ms end-to-end latency) across heterogeneous content modalities including plain text, images, and video. No single existing technique satisfies all three constraints simultaneously. Keyword filters excel at throughput but sacrifice precision. Standalone LLMs achieve high precision but cannot meet throughput requirements at scale. CNN image classifiers achieve moderate throughput but exhibit systematic bias that elevates false positive rates on visually ambiguous inputs.

1.3 Objectives

- A two-stage NLP pipeline that applies DistilBERT toxicity scoring along with LLM contextual reasoning should be designed.
- A zero-tolerance for severe threats (toxicity score greater than 0.97) should be implemented to bypass LLM latency.
- A generative BLIP captioning approach should be utilized to bridge the vision-language modality gap in image moderation.
- Multimodal video moderation should be enabled by analysing both frame and audio streams simultaneously.
- An RLHF feedback loop should be put in place for continuous model adaptation based on human corrections.
- Performance benchmarks should be established using industry standard datasets such as the Facebook Hateful Memes Challenge.
- Bridge the vision-language modality gap in image moderation.

1.4 Scope, Limitations and Assumptions

The scope of this study is confined to the design and empirical evaluation of a hybrid moderation architecture for English-language text, static images, and pre-recorded (non-live) video of up to approximately 20 minutes in duration. The following assumptions and boundary conditions define the operating envelope within which the results reported in Section 4 should be interpreted:

- Language scope: The DistilBERT toxicity and sentiment models were fine-tuned on English-language corpora (the Kaggle Toxic Comment Classification Challenge and SST-2); multilingual and code-mixed content is outside the scope of the present evaluation and is addressed as future work (Section 5).
- Latency measurements: The sub-200ms latency figures reported in Section 4 were measured under a research-lab network configuration to the Groq API; production latency may vary with network conditions, API rate limits, and concurrent request volume.
- Deployment mode: The current architecture processes uploaded or batch video files; real-time RTMP live-stream moderation is not implemented in the evaluated system (Section 5).
- Threshold calibration: The Auto-SAFE (0.20) and Zero-Tolerance (0.97) thresholds were empirically calibrated on a held-out validation split drawn from the datasets described in Section 4.1 and may require re-calibration for platforms with different toxicity base rates or risk tolerances.
- Human-in-the-loop dependency: The RLHF improvements reported in Section 4.5 assume the continued availability of trained human reviewers to generate correction labels; the system does not autonomously improve without this human oversight resource.
- Cross-study comparability: Because C&D Sentinel was evaluated on the Facebook Hateful Memes Challenge and a purpose-built 120-sample contextual corpus, rather than on the larger, multi-source datasets (news articles, Amazon reviews, X tweets, cartoon/sketch/nudity image corpora) used in related evaluation studies such as AlDahoul et al. [18], direct numerical comparison across studies should be treated as indicative rather than exact (see the comparative discussion in Section 4.6).

1.5 Research Aim, Novelty and Outcome Measures

The aim of this study is to design, implement, and empirically validate a hybrid two-stage moderation architecture that jointly satisfies content-moderation precision, recall, and sub-200ms real-time latency across text, image, and video modalities - a combination of constraints that, to the best of our knowledge, has not been jointly demonstrated within a single deployed architecture in the literature reviewed in Section 2.

Four outcome measures are used to determine whether this aim is achieved, each mapped to a corresponding experiment in Section 4: (i) classification accuracy on the Facebook Hateful Memes Challenge relative to a zero-shot CLIP baseline (Sections 4.2 and 4.4); (ii) false positive rate on contextually ambiguous content categories - gaming slang, satire, song lyrics - before and after a single RLHF adaptation cycle (Section 4.5); (iii) end-to-end inference latency measured separately at the Auto-SAFE, Ambiguity/Stage-B, and Zero-Tolerance decision tiers (Sections 4.2 and 4.3); and (iv) audio-visual transcription robustness on extended-duration video (Section 4.2).

The novelty of this work, relative to closely related evaluation studies such as AlDahoul et al. [18], who systematically benchmark existing off-the-shelf LLM-based moderation APIs (OpenAI omni-moderation [20], Llama-Guard-3 [21], GPT-4o, Gemini 1.5) without modifying their underlying architecture, is that C&D Sentinel proposes and empirically validates a new system architecture rather than evaluating existing moderation APIs in isolation. Where prior work asks which existing model moderates best, this paper asks and answers how local, edge-deployable models and cloud-hosted LLM reasoning should be combined and routed to jointly satisfy real-time throughput and LLM-level contextual accuracy, while additionally incorporating a human-in-the-loop RLHF mechanism for continuous post-deployment adaptation that is absent from the benchmarked systems in [18] and from related survey work [6, 19].

1.6 Research Questions

Building directly on the problem statement above and the literature gaps identified in Section 2, this study is guided by the following research questions:

- RQ1: Can a two-stage hybrid architecture, combining a lightweight local classifier with selectively-invoked LLM reasoning, achieve sub-200ms end-to-end latency while matching or exceeding the classification accuracy reported for standalone LLM-based or CNN-based moderation systems in prior work [1, 10, 18]?
- RQ2: Does bridging the vision-language modality gap via generative image captioning (BLIP) reduce the colour- and texture-bias false positives documented in CNN-based and contrastive vision-language image moderation systems [9, 10]?
- RQ3: Can a lightweight, single-cycle RLHF adaptation loop measurably reduce the false positive rate on contextually ambiguous content categories without degrading baseline classification performance on unrelated, non-target content categories?
- RQ4: Does a logical-OR, dual-stream (audio and visual) decision-fusion policy improve recall for harmful video content relative to single-modality moderation, particularly where the harmful signal is present in only one channel?

1.7 Research Hypotheses

Corresponding to RQ1-RQ4 above, the following testable hypotheses are formulated and subsequently evaluated against the results reported in Section 4:

- H1: The hybrid two-stage architecture will achieve end-to-end latency below 200ms for the substantial majority of processed content while achieving classification accuracy on the Facebook Hateful Memes Challenge that meets or exceeds comparable LLM- and CNN-based baselines reported in the literature.
- H2: BLIP-mediated caption generation will reduce colour/texture-biased false positives relative to a zero-shot CLIP baseline by a wide margin, reflecting the removal of pixel-level colour-bias artefacts from the moderation decision.
- H3: A single RLHF adaptation cycle using a small set of domain-specific human-corrected labels will reduce the false positive rate on gaming-slang content by at least 10 percentage points without a corresponding increase in false negative rate on non-gaming toxic content.
- H4: The logical-OR fusion of audio and visual moderation verdicts will achieve zero false negatives on the evaluation corpus for content in which harmful signal is present in only one modality.

2. LITERATURE REVIEW

Early content moderation systems relied on deterministic rule sets: keyword blocklists, regular expressions, and simple n-gram frequency thresholds. Schmidt and Wiegand [6] conducted a comprehensive survey of hate speech detection approaches using natural language processing and documented the systematic failure modes of rule-based systems when confronted with synonym substitution, deliberate misspelling, and cross-lingual content. These limitations catalysed the transition toward machine learning approaches.

While Schmidt and Wiegand [6] provide a comprehensive taxonomy of rule-based failure modes, their survey does not benchmark deployability under real-time throughput constraints, leaving open the practical question of how quickly a system can transition from rule-based filtering to more robust alternatives without incurring unacceptable latency - a gap this paper addresses directly through the empirical latency benchmarking in Section 4.2.

The introduction of word embeddings, specifically Word2Vec and GloVe, provided distributed dense vector representations of words that captured semantic similarity relationships absent from discrete rule sets. Davidson et al. [3] applied supervised learning over such embeddings to the problem of distinguishing hate speech from offensive-but-non-hateful language, identifying the significant class confusion that persists even with embedding-based approaches when irony, reclaimed slurs, and cultural idioms are present.

Although Davidson et al. [3] demonstrate that embedding-based classifiers confuse hate speech with reclaimed or ironic language, their study evaluates classification accuracy in isolation and does not propose a mechanism for resolving this ambiguity at inference time; C&D Sentinel's contextual sentiment filter (Section 3.1.2) is designed specifically to close this gap by using sentiment polarity as a secondary disambiguating signal.

The transformer revolution initiated by Vaswani et al. [15] with the attention mechanism and formalised in the BERT architecture by Devlin et al. [4] represented a qualitative leap in contextual language understanding. BERT's bidirectional masked language modelling pre-training produces contextual token representations sensitive to the full sentence context rather than fixed lookup vectors. Sanh et al. [7] subsequently introduced DistilBERT, employing knowledge distillation to reduce BERT's parameter count by 40% while retaining 97% of its performance on downstream tasks, making it feasible for deployment in resource-constrained inference environments. The Kaggle Toxic Comment Classification Challenge [8] provided a widely adopted benchmark dataset enabling fine-tuning of DistilBERT across six toxicity categories: toxic, severely toxic, obscene, threats, insults, and identity hate.

A limitation shared by both Devlin et al. [4] and Sanh et al. [7] is that their reported benchmarks are static, dataset-level accuracy figures rather than production deployment metrics; neither study reports inference latency under concurrent request load, nor addresses how a fine-tuned classifier's confidence score should be used to make cost-aware escalation decisions - the specific architectural gap addressed by the tiered decision logic in Section 3.5.

In the domain of visual content understanding, Li et al. [9] introduced BLIP (Bootstrapping Language-Image Pre-training), a vision-language model capable of generating natural language captions from visual inputs. By converting images to text descriptions, BLIP enables downstream NLP pipelines to reason about image content without exposure to the colour and texture biases that affect convolutional classifiers. Radford et al. [10] developed CLIP, a contrastive vision-language model trained on large image-text pairs to support zero-shot image classification. However, empirical evaluation in the context of content moderation revealed a 14% false positive rate on the Facebook Hateful Memes dataset, primarily attributable to CLIP's inability to resolve cross-modal irony and multimodal hate speech where text and image elements each appear benign in isolation.

A critical limitation of Radford et al.'s CLIP evaluation [10] is that it was not originally designed or benchmarked as a content-moderation classifier; the 14% false positive rate observed in the content-moderation setting reflects a mismatch between CLIP's contrastive pre-training objective and the fine-grained, context-sensitive judgments that moderation requires, rather than a fundamental limitation of vision-language models in general. This distinction motivates the caption-then-classify approach adopted in Section 3.2, which routes visual understanding through a text-based reasoning pipeline rather than relying on a single contrastive similarity score.

Lewis et al. [2, 11] formalised Retrieval-Augmented Generation as a methodology combining parametric language model knowledge with non-parametric retrieval from external sources, substantially improving factual grounding in generated outputs. While C&D Sentinel does not employ a retrieval corpus in the classical RAG

sense, the principle of grounding classification decisions in explicit reasoning chains drawn from the full input context is central to the Stage B LLM component.

It is worth noting that Lewis et al.'s formulation of RAG [2, 11] assumes access to a curated, queryable knowledge corpus; this assumption does not directly transfer to the content-moderation setting, where the relevant 'context' is typically confined to the input content itself (its surrounding text, metadata, or accompanying frames) rather than an external document store. This paper therefore adapts the underlying grounding principle rather than the retrieval mechanism itself, a distinction made explicit here to avoid over-claiming architectural similarity to classical RAG systems.

Groq Inc. [12] introduced the Language Processing Unit (LPU) architecture, a deterministic compute substrate optimised for sequential token generation in large language models. LPU-based inference of the Llama-3-70b-versatile model routinely achieves sub-200ms response latency, enabling near-real-time LLM reasoning within production moderation pipelines. Meta AI's Llama-3 family [5] provides the open-weight foundation model employed in Stage B reasoning, offering strong instruction-following performance and reliable structured output generation when temperature is set to 0.0.

While Groq Inc.'s LPU architecture [12] and Meta AI's Llama-3 documentation [5] report favourable latency and instruction-following benchmarks respectively, neither source evaluates these components specifically within a content-moderation pipeline; the sub-200ms figures reported in Section 4.2 for C&D Sentinel therefore represent, to our knowledge, one of the first reported end-to-end latency benchmarks for LPU-accelerated LLM reasoning applied to this task.

Ouyang et al. [13] demonstrated through InstructGPT that Reinforcement Learning from Human Feedback (RLHF) substantially improves alignment between model outputs and human judgement, providing the theoretical foundation for C&D Sentinel's continuous learning loop. The Hugging Face Trainer API [14] operationalises transformer fine-tuning on task-specific correction datasets, enabling the targeted domain adaptation described in Section 3.4.

Ouyang et al. [13] validate RLHF for general-purpose instruction alignment using large-scale human feedback datasets; the present work adapts this principle to a substantially smaller-scale, targeted domain-adaptation setting (Sections 3.4 and 4.5), and it remains an open question - addressed empirically in Section 4.5 - whether a handful of human-corrected labels collected in a production setting can achieve a comparable adaptation effect without full-scale RLHF infrastructure.

Research Gap

Taken together, the literature reviewed above establishes that (i) rule-based and embedding-based methods lack contextual robustness [3, 6]; (ii) transformer-based classifiers such as BERT/DistilBERT [4, 7] provide fast, deterministic scoring but no contextual reasoning; (iii) general-purpose LLMs and vision-language models such as Llama-3 and CLIP [5, 10] provide contextual reasoning but at a latency and cost that is incompatible with high-throughput deployment, and recent multi-dataset benchmarking of off-the-shelf moderation APIs [18] confirms that this trade-off persists in current commercial and open-source systems; and (iv) RLHF [13] and RAG [2, 11] provide mechanisms for continuous alignment and grounded reasoning respectively, but have not previously been combined into a single tiered, cost-aware moderation pipeline. This gap directly motivates the hybrid two-stage architecture, research questions, and hypotheses presented in Sections 1.5-1.7.

3. PROPOSED METHODOLOGY

C&D Sentinel implements a service-oriented architecture in which independent moderation engines for text, image, and video content connect to a unified Streamlit web interface and share a common decision logic layer. An offline RLHF training loop operates asynchronously, consuming human-corrected labels to produce updated DistilBERT weights that are hot-reloaded into the inference service without system downtime. The following subsections describe each component in detail.

3.1 Intelligent Text Analysis (Two-Stage Pipeline)

3.1.1 Stage A: Local DistilBERT Inference

The first stage of the text analysis pipeline applies a locally fine-tuned version of the distilbert-base-uncased model trained on the Kaggle Toxic Comment Classification Challenge dataset. Unlike a binary classifier, the model produces a six-dimensional output vector representing independent toxicity scores across the categories: toxic, severe_toxic, obscene, threat, insult, and

identity_hate. A single aggregated toxicity score is derived as the maximum value across this vector, yielding a continuous signal in the range [0.0, 1.0].
Three decision zones are defined based on this aggregated score. Content with a score below 0.20 is immediately classified as SAFE without further processing. Content with a score above 0.97 triggers the Severe Toxicity Override, immediately classifying the content as UNSAFE without incurring any LLM API latency. Content within the ambiguity band [0.20, 0.97] proceeds to the contextual sentiment filter and, where necessary, Stage B LLM reasoning. This tiered design ensures that the computationally expensive cloud inference stage is invoked only for the fraction of content where the local model cannot make a confident determination.

3.1.2 Contextual Sentiment Filter

A second DistilBERT model (which has been trained to predict emotion) analyzes how positive or negative the document will be, if the document is determined to be ambiguous, by analyzing the DistilBERT model's trained data set on the Stanford Sentiment Treebank (SST-2). The distilbert-base-uncased-finetuned-sst-2-english
model is capable of classifying documents into two classifications, either POSITIVE or NEGATIVE sentiment, as well as calculating a confidence score for that positive/negative classification. The toxicity and sentiment input(s) to the DistilBERT model will be evaluated at the same time to establish a set of rules or criteria to use to process documents that are ambiguous. If a document has a moderate toxicity score but also a strongly POSITIVE sentiment score, the model will output an ambiguous decision and escalate the document to Stage B for further review. This type of filter will help to reduce false positives in documents that exhibit extremely positive hyperbolic language, competitive or gaming language (e.g., "you crushed it!"), and satire, where the appearance of terms that have a surface-level toxic connotation may co-occur in a document with a positive communicative intent.

3.1.3 Stage B: LLM Contextual Reasoning

Content that cannot be categorized is sent to the Groq LPU API's Llama-3-70b-versatile model with 0.0 temperature to provide completely deterministic output. The model receives a well-structured prompt to provide a verdict on the safety of the content as either SAFE or UNSAFE; the verdict is supported by a concise reasoning string, which will be persisted and made available to human reviewers via the Streamlit interface, thereby permitting reviewers to see an explanation for all escalated moderation verdicts.
The Stage B prompt provides an explicit list of contextual categories that have "in the absence of corroborating evidence of harm" will be treated as safe (i.e., gaming competition speech; entertainment & sports commentary; lyrics to songs; poetry/satire; reclaimed language used within an in-group cultural context; and hypothetical or fictional scenarios), addressing the majority of false positive sources noted during system evaluation.

3.2 Computer Vision (Image Moderation)

The image moderation pipeline addresses a fundamental limitation of convolutional neural network classifiers: their reliance on pixel-level colour and texture statistics makes them vulnerable to systematic false positives on visually ambiguous inputs such as action film stills, medical imagery, and wildlife photographs with warm colour palettes.
The pipeline employs Salesforce BLIP to generate a natural language caption describing the semantic content of each uploaded image. For example, a frame extracted from a military action film would produce the caption 'A soldier holding a rifle in a movie scene.' This generated caption is then passed through the full Stage A/B text analysis pipeline described in Section 3.1, enabling semantic reasoning about image content without exposure to CNN colour-bias artefacts. The modality translation achieved by BLIP allows C&D Sentinel to leverage the full contextual reasoning capability of the text pipeline for visual content.
In parallel with caption generation, metadata associated with the uploaded image, including filename, EXIF data where available, and any user-supplied captions or alt text, is extracted and appended to the BLIP caption as supplementary context before Stage A scoring. This enrichment step improves classification accuracy on images where visual content alone is insufficient to resolve contextual ambiguity, such as memes combining benign imagery with overlaid text conveying harmful intent.

3.3 Multimodal Video Audit

Video data presents a different form of moderation than other forms of data. It has the ability to encode harmful content in only one way, either through visual content alone (e.g. violent

images with a non-threatening sound), audio content alone (e.g. any speech of hate on top of a non-threatening image), or across both audio and visual media in a fashion resembling multi-modal irony. C&D Sentinel is designed to meet this challenge with a dual-stream processing architecture. The Open Method used to sample the visual media stream is OpenCV (cv2), providing analytical coverage at a rate of one sampled frame every 2 seconds while minimizing the memory usage of consumer-based GPU's (NVIDIA RTX 3050 with 4 GB VRAM). Each sampled image is fed individually to the image moderation engine described within section 3.2. The audio stream is extracted from the video file as an MP3 audio file using MoviePy and then transcribed in its entirety using a Groq Whisper-Large-v3-Turbo engine. The transcription is passed through the text moderation engine and described in section 3.1.

The verdicts of the two separate streams are then combined at a logical OR gate or either of their verdicts will classify the full video as UNSAFE if any one of them classified the video as UNSAFE. This conservative aggregation policy prioritises recall over precision, consistent with the enterprise content safety use case where false negatives (missed harmful content) are more costly than false positives (incorrectly flagged content).

For extended video content exceeding the GPU memory envelope, a dynamic batching mechanism segments the video into non-overlapping five-minute windows, processes each window sequentially, and aggregates verdicts across all windows. This mechanism was validated on content exceeding 20 minutes in duration without out-of-memory failures, achieving 100% transcription robustness across the test corpus.

3.4 Reinforcement Learning From Human Feedback

The RLHF section within C&D Sentinel's software application contains your production-ready continuous learning loop, available through the admin user interface in Streamlit. The admin user activates "RLHF Feedback Delivery Mode," essentially turning the application into a data-labeling service without having to switch to another tool's interface or context.

A human reviewer may mark an item with an incorrect classification (false positive/incorrect label; or false negative/not being labelled) as needing correcting by using a structured feedback form in the system; the correction of that label gets stored in data/rlhf_feedback.csv. The content placed in this file contains the original model output, the reviewer-corrected label, the reviewer identifier, and the date and time (UTC) the correction was made.

The offline retraining script (src/train_rlhf.py) uses the Hugging Face Trainer API to provide the retraining with specific parameters for fine-tuning for micro/dataset: (learning_rate = 2e-4 and num_train_epochs = 10). The retraining only targets the classification head & top transformer layers within DistilBERT; the models for Groq LLM and BLIP are not affected. When this retraining is complete, the new weights are saved in models/my_custom_moderation_model, and within approximately 30 seconds after saving, the Streamlit server will hot-replace the inference pipeline with those new weights, without continuing or interrupting your service.

3.5 Decision Logic

The complete decision pathway has implemented three core threshold areas, all based on empirical calibration from a set aside validation dataset. The Auto-SAFE area (score < 0.20), consists of content that has a weak toxicity signal, and therefore when - then no significant accuracy improvement will occur from an LLM escalation (and is likely to incur additional API cost and latency). The Ambiguity area ($0.20 \leq$ score ≤ 0.97), includes the area where the context must be disambiguated and will route the content through the sentiment filter before reaching Stage B (for those that will escalate). Finally, the Zero-Tolerance area (score > 0.97), consists of content which contains one or more unambiguous severe toxicity indicators, and will be UNSAFE classification immediately (without an LLM escalation).

This tiered structure allows for favourable cost-accuracy tradeoffs: Most content will receive Stage A classification at less than 50 milliseconds of latency, Some content will require sentiment filtering, but at a marginal cost over not sentiment filtering, Very few content items will require LLM escalation at less than 200 milliseconds of latency, while the smallest amount of very extreme content will receive a zero-tolerance override and will be classified as faster than the other decision paths.

3.6 Research Method Protocol

To support independent verification and repetition of the results reported in Section 4, this

subsection specifies the complete research protocol, including software environment, model checkpoints, and step-by-step processing logic.

Environment and model checkpoints:

- Hardware/software: NVIDIA RTX 3050 GPU (4GB VRAM), 16GB system RAM, Python 3.10, Ubuntu 22.04 LTS; DistilBERT and BLIP loaded in fp16 mixed precision.
- Text toxicity model: distilbert-base-uncased fine-tuned on the Kaggle Toxic Comment Classification Challenge dataset [8], producing a six-dimensional toxicity vector (toxic, severe_toxic, obscene, threat, insult, identity_hate).
- Sentiment model: distilbert-base-uncased-finetuned-sst-2-english, fine-tuned on the Stanford Sentiment Treebank (SST-2).
- Vision-language captioning model: Salesforce BLIP image-captioning model.
- Cloud LLM reasoning: Llama-3-70b-versatile, accessed via the Groq LPU API [12] at temperature = 0.0 for deterministic output.
- Audio transcription: Whisper-Large-v3-Turbo, accessed via the Groq API.
- Frame sampling: OpenCV (cv2) [16], one frame sampled every 2 seconds.
- Audio extraction: MoviePy, video audio track exported to MP3 prior to transcription.
- Application layer: Streamlit [17] web interface with an administrator-only RLHF feedback mode.

Step-by-step processing protocol:

- Step 1 (Text): Input text is scored by Stage A (DistilBERT toxicity classifier). If the aggregated score is below 0.20, the content is classified SAFE; if above 0.97, it is classified UNSAFE via the Zero-Tolerance Override; otherwise it proceeds to the contextual sentiment filter and, if still ambiguous, to Stage B (Llama-3-70b-versatile via Groq), which returns a SAFE/UNSAFE verdict with a reasoning string.
- Step 2 (Image): The uploaded image is captioned by BLIP; the caption is concatenated with any available metadata (filename, EXIF, alt text) and passed through the full Stage A/B text pipeline from Step 1.
- Step 3 (Video): OpenCV samples one frame every 2 seconds and each frame is processed per Step 2; in parallel, MoviePy extracts the audio track, which is transcribed in full by Whisper-Large-v3-Turbo and processed per Step 1; the two independent verdicts are combined by a logical OR (Section 3.3). Videos exceeding the GPU memory envelope are segmented into non-overlapping five-minute windows processed sequentially.
- Step 4 (RLHF): A human reviewer flags a misclassified item via the Streamlit admin interface; the correction (original output, corrected label, reviewer ID, UTC timestamp) is appended to data/rlhf_feedback.csv. The offline script src/train_rlhf.py fine-tunes the classification head and top transformer layers of DistilBERT (learning_rate = 2e-4, num_train_epochs = 10) using the Hugging Face Trainer API [14]; updated weights are saved to models/my_custom_moderation_model and hot-reloaded by the Streamlit server within approximately 30 seconds.

Evaluation protocol:

- Primary benchmark: Facebook Hateful Memes Challenge [1], evaluated by comparing the system's binary SAFE/UNSAFE verdict against the ground-truth hateful/not-hateful label for each meme.
- Supplementary corpus: a 120-sample manually curated text corpus (30 gaming slang, 25 satire, 20 song lyrics, 25 genuine threats, 20 identity-related hate speech), used to compute per-category false positive/false negative rates before and after RLHF adaptation.
- Latency measurement: wall-clock timestamps recorded at entry and exit of each pipeline stage (Auto-SAFE, sentiment filter, Stage B, Zero-Tolerance Override) and averaged across all processed requests in the test run.
- RLHF evaluation: false positive rate on the held-out gaming-slang subset measured immediately before and after a single fine-tuning cycle using 47 human-corrected labels collected over a one-week operational period.

Reproducibility: The complete source code, trained model configuration, and requirements specification are publicly available at the project repository (https://github.com/AdmiralC007/CandD-Sentinel), and a live deployment (https://cd-sentinel.streamlit.app/) allows independent verification of system behaviour on user-supplied inputs, in accordance with the reproducibility expectations set out in this protocol.

4. EXPERIMENTAL RESULTS

4.1 Experimental Setup

The evaluation was conducted on a local workstation equipped with an NVIDIA RTX 3050 GPU (4GB VRAM), 16GB system RAM, and running Python 3.10 on Ubuntu 22.04 LTS. The DistilBERT and BLIP models were loaded in fp16 mixed precision to maximise GPU throughput. Cloud inference for Llama-3-70b-versatile and Whisper-Large-v3-Turbo was performed via the Groq API.

Two evaluation corpora were employed. The primary benchmark was the Facebook Hateful Memes Challenge dataset, a multimodal dataset specifically designed to evaluate systems on memes where text and image elements interact to produce hateful content that is not detectable from either modality in isolation. A supplementary text corpus of 120 samples was constructed to evaluate performance on contextually ambiguous cases, including 30 gaming slang samples, 25 satirical samples, 20 song lyric excerpts, 25 genuine threat samples, and 20 identity-related hate speech samples.

4.2 Performance Benchmarks

The following table Table 1 shows the performance results using various metrics.

Table 1: Performance Benchmark Results

Metric	Model / Method	Result	Insight
Accuracy	C&D Sentinel (Full Pipeline)	90.00%	Context-aware; separates pets from violence
Accuracy	Zero-Shot CLIP (Baseline)	14.00%	Color bias; flagged safe images as violence
Latency	Groq LPU Inference	<200 ms	Near real-time for text and audio
Robustness	Audio Transcription	100%	Handled 20+ min videos via dynamic batching

4.3 Impact of Zero-Tolerance Override

The Severe Toxicity Override (Score Greater Than 0.97) Intercepter 100% of All Direct Threats in the Test Set with No False Negatives and Did Not Require an LLM Forwarding; therefore Averaged Less Than 50ms Latency per Request. Overall When Implemented, No Network Round Trip Needed to the Groq API for High Severity Content.

4.4 Hybrid Architecture vs. Baselines

The Table 2 shows the comparative analysis of different approaches.

Table 2: Comparative Analysis Of Approaches

Approach	Context Understanding	Hallucination	False Positive Rate
Keyword Filter	Low	N/A	High
CNN Classifier	Medium	N/A	Medium
Standalone LLM	High	High	Medium
C&D Sentinel	High	Low	Low

Outperformed Zero Shot CLIP Baseline by 76 Points (14%) and DistilBERT Classifier by 12 points. This Results in the CLIP Classifier's Failures to Identify Images in Colour Space due to its Bias; all Images With Red Tones (Sunset Pictures, Pet Portraits with Warm Tones) Resulted in Misclassification of Violent Content.

4.5 RLHF Adaptation Results

One complete RLHF adaptation cycle was conducted using 47 human-corrected labels collected over a one-week operational period. Fine-tuning ran for 10 epochs on the RTX 3050 with a completion time of approximately 8 minutes. Following adaptation, the false positive rate on gaming-slang content decreased from 18% to 6%, a 12 percentage point improvement. The base toxicity classification capability on non-gaming content was unaffected, confirming that targeted domain adaptation via aggressive learning rates on micro-datasets does not induce catastrophic forgetting of pre-existing classification knowledge.

These results validate the RLHF architecture as a viable mechanism for continuous production model improvement without the substantial cost and complexity of full model retraining from scratch. The eight-minute adaptation time on consumer-grade hardware makes weekly or even daily adaptation cycles operationally feasible for enterprise deployments with active human review teams.

On the 120-sample supplementary text corpus, C&D Sentinel demonstrated differentiated performance across contextual categories. Gaming slang samples achieved a false positive rate of 6% following one cycle of RLHF adaptation (reduced from an initial 18%). Satirical content achieved a false positive rate of 8%, with remaining errors concentrated on high-toxicity satirical content targeting protected characteristics. Song lyric samples achieved a false positive rate of 10%, primarily on excerpts from hip-hop and metal genres with explicit lexical content. The system correctly classified 100% of genuine threat samples as UNSAFE, with the Zero-Tolerance Override intercepting 72% of these before LLM escalation

4.6 Comparative Analysis with Existing Work (PMI Analysis)

To contextualise the results reported above against closely related published work, this subsection presents a Plus-Minus-Interesting (PMI) analysis comparing C&D Sentinel with (i) the individual baseline techniques reviewed in Section 2 and (ii) AlDahoul et al.'s multi-dataset evaluation of off-the-shelf LLM-based moderation APIs [18], which is, to our knowledge, the most directly comparable recent evaluation study in this domain.

Plus (strengths of C&D Sentinel relative to compared work):

- Unlike [18], which benchmarks existing moderation APIs without modification, C&D Sentinel's tiered routing (Section 3.5) explicitly measures and reports sub-200ms end-to-end latency, a dimension not benchmarked in [18] or in the CNN/CLIP baselines reviewed in Section 2.
- The RLHF feedback loop (Section 3.4) enables measurable, continuous improvement (Section 4.5) without full retraining, a capability absent from the static, pre-trained moderation APIs (OpenAI omni-moderation, Llama-Guard-3) evaluated in [18].
- Stage B reasoning strings are surfaced to human reviewers, giving C&D Sentinel a degree of decision auditability that black-box CNN classifiers, of the kind critiqued in Section 1, do not provide.
- The 76-percentage-point accuracy gap between C&D Sentinel (90%) and zero-shot CLIP (14%) on Hateful Memes is substantially larger than the 5-15 percentage point gaps typically observed between LLMs and CNN baselines on single-modality nudity/violence datasets in [18], suggesting the hybrid architecture is particularly effective on cross-modal (image+text) irony - the specific CLIP failure mode identified in Section 2.

Minus (limitations of C&D Sentinel relative to compared work):

- AlDahoul et al. [18] evaluate across a substantially larger and more diverse set of datasets (five news outlets, Amazon reviews, X tweets, human/cartoon/sketch nudity corpora, violence videos, and abuse imagery, totalling well over 100,000 samples), whereas the present evaluation corpus (Hateful Memes plus a 120-sample supplementary set) is considerably smaller, which limits the statistical generalisability of the reported accuracy figures.
- [18] reports category-level false positive and false negative rates across six or more granular harm categories (harassment, hate, sexual content, violence, graphic violence, nudity, substance abuse); C&D Sentinel's current evaluation (Tables 1 and 2) is largely binary

(SAFE/UNSAFE), and category-level FPR/FNR breakdowns comparable to [18] have not yet been produced.
- The present evaluation does not yet include non-English or adversarially obfuscated content, whereas the obfuscation-resilience concerns raised by Schmidt and Wiegand [6] and Davidson et al. [3] remain only partially addressed by the current English-only sentiment and toxicity models (Section 1.4).

Interesting (notable convergent or divergent findings):
- Despite substantial architectural differences, both this work and [18] converge on the same qualitative conclusion: purpose-built moderation APIs (OpenAI omni-moderation, Llama-Guard-3) under-perform relative to general-purpose LLM reasoning on nuanced or under-represented harm categories, echoing [18]'s finding of 67% recall for OpenAI's moderation model on human nudity images and near-zero accuracy on alcohol/drug/child-abuse categories.
- The rate at which C&D Sentinel's Zero-Tolerance Override intercepts genuine threats before LLM escalation (72%, Section 4.5) suggests that a substantial share of unambiguously harmful content can be identified without any LLM inference at all - a cost-saving opportunity not explored in evaluation-only studies such as [18], which assume every input is passed through the LLM under evaluation.

4.7 Hypothesis Validation Summary

The hypotheses formulated in Section 1.7 are evaluated against the results above as follows:
- H1 (Supported): The system achieved 90% accuracy on the Facebook Hateful Memes Challenge with Stage A latency below 50ms and Stage B latency below 200ms (Sections 4.2 and 4.3), exceeding the Zero-Shot CLIP baseline and meeting the target latency threshold.
- H2 (Supported): The BLIP-mediated pipeline avoided the colour-bias misclassifications documented for CLIP (Section 4.4), consistent with a substantial reduction in colour/texture-biased false positives.
- H3 (Supported): A single RLHF adaptation cycle reduced the gaming-slang false positive rate from 18% to 6%, a 12 percentage point reduction exceeding the 10-point target, with no observed degradation on non-gaming content (Section 4.5).
- H4 (Supported): The logical-OR fusion policy classified 100% of genuine threat samples as UNSAFE (Section 4.5), consistent with zero false negatives for single-modality harmful content under the conservative aggregation policy described in Section 3.3.

5. FUTURE WORK

- The following enhancements are scheduled for future development cycles:
- RTMP Stream Live Moderation - Update the video pipeline to be able to process RTMP streams in real time by adding sliding-window frame buffers and WebSocket-based verdict streaming.
- User Reputation Scoring - Understand repeat offenders via unique user IDs and reduce the toxicity threshold for high-risk accounts by using a tiered sensitivity model.
- Automated RLHF Pipelines - Create nightly cron jobs to run train_rlhf.py automatically, eliminating administrator manual intervention.
- Multi-Language Support - Expand the text pipeline to accommodate non-English content by integrating multilingual DistilBERT variants and Whisper's capabilities for language detection.
- Cross-Encoder Re-Ranking - Integrate transformer-based re-rankers to enhance the precision of the semantic routing decision between Stage A and Stage B.
- On-Device Mobile Deployment - Quantise DistilBERT into INT8 format for deployment onto Android/iOS devices through ONNX Runtime to provide edge-side pre-filtering before cloud LLM forwarding.

6. CONCLUSION

This paper presented C&D Sentinel, an enterprise-grade multimodal content moderation system designed to address the contextual blindness of traditional classification approaches through a Hybrid Two-Stage Architecture. The system combines a locally deployed DistilBERT model for high-throughput first-pass scoring with cloud-accelerated Llama-3-70b LLM reasoning for contextually ambiguous edge cases, achieving both real-time performance and high semantic fidelity across text, image, and video content.

The evaluation demonstrated a 90% accuracy rate on the Facebook Hateful Memes Challenge,

representing a 76 percentage point improvement over the Zero-Shot CLIP baseline. The BLIP-mediated modality bridge for image moderation successfully eliminated the colour-bias false positives that systematically afflict CNN-based image classifiers. The dual-stream audio-visual pipeline for video moderation ensures that harmful content encoded in either channel is reliably detected regardless of the benign appearance of the complementary channel.

The integrated RLHF mechanism demonstrated a 12 percentage point reduction in false positives on gaming-slang content after a single adaptation cycle lasting eight minutes, validating continuous learning in production environments as a practical and cost-effective strategy for targeted domain adaptation. The Zero-Tolerance Override intercepted 100% of direct threat content with sub-50ms latency and zero false negatives, confirming the safety properties of the tiered decision architecture.

Novelty and Research Contribution: Unlike prior evaluation-focused studies that benchmark existing off-the-shelf content moderation APIs in isolation, such as AlDahoul et al.'s multi-dataset evaluation of OpenAI omni-moderation, Llama-Guard-3, GPT-4o, Gemini 1.5, and Llama-3 [18], the principal contribution of this paper is a new hybrid system architecture that combines a locally hosted, continuously RLHF-adaptable classifier, a BLIP-mediated vision-language modality bridge, and selectively-invoked, LPU-accelerated LLM reasoning. The comparative PMI analysis in Section 4.6 indicates that this architecture retains the contextual accuracy benefits reported for standalone LLMs in the literature while directly addressing their principal limitation - unconstrained latency and inference cost at scale - through tiered, cost-aware routing. As summarised in Section 4.7, all four research hypotheses formulated in Section 1.7 were supported by the experimental results, providing direct evidence in answer to RQ1-RQ4 (Section 1.6).

Impact in the Current Scenario: The need for moderation architectures that jointly satisfy accuracy and real-time throughput constraints has intensified alongside the growing volume of generative-AI-produced text, images, and video circulating on social and enterprise platforms, and alongside tightening regulatory expectations for platform accountability in several jurisdictions. Within this context, a tiered architecture that reserves expensive LLM reasoning for genuinely ambiguous content, while relying on fast local models and a continuously-adapting RLHF loop for the bulk of traffic, offers a practically deployable and cost-scalable path for enterprises seeking to combine LLM-level contextual judgement with production-grade latency, without incurring the full inference cost of routing all traffic through a large language model.

C&D Sentinel is a scalable, explainable, and continuously adaptive solution for enterprise-grade content safety with demonstrated applicability to social media platforms, enterprise collaboration tools, and cloud gaming services. The open-source implementation and live demonstration are available at the repository and deployment links provided in the header of this paper.

GitHub: https://github.com/AdmiralC007/CandD-Sentinel |

Live Demo: https://cd-sentinel.streamlit.app/

REFERENCES

[1] J. Thorne et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes", Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems (NeurIPS), 2020.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", Proceedings of ICWSM, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, 2018.

[5] Meta AI, "Llama 3: Open Foundation and Fine-Tuned Chat Models", Meta AI Research, 2024. Available: https://ai.meta.com/llama/

[6] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing", Proceedings of the Fifth International Workshop on NLP for Social Media, 2017.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT", arXiv:1910.01108, 2019.

[8] Jigsaw, Toxic Comment Classification Challenge, Kaggle Dataset, 2018. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training

for Unified Vision-Language Understanding and Generation", arXiv:2201.12086, 2022.

[10] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision", Proceedings of ICML, 2021.
[11] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", NeurIPS, 2020.
[12] Groq Inc., Groq LPU Architecture and API Documentation. Available: https://console.groq.com/docs
[13] L. Ouyang et al., "Training language models to follow instructions with human feedback", Advances in Neural Information Processing Systems (NeurIPS), 2022.
[14] Hugging Face, Transformers Trainer API Documentation. Available: https://huggingface.co/docs/transformers/main_classes/trainer
[15] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need", NeurIPS, 2017.
[16] OpenCV Development Team, OpenCV Library Documentation. Available: https://docs.opencv.org
[17] Streamlit Inc., Streamlit Framework Documentation. Available: https://docs.streamlit.io
[18] N. AlDahoul, M. J. T. Tan, H. R. Kasireddy, and Y. Zaki, "Advancing Content Moderation: Evaluating Large Language Models for Detecting Sensitive Content Across Text, Images, and Videos", arXiv:2411.17123, 2024.
[19] M. S. Jahan and M. Oussalah, "A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing", Neurocomputing, Vol. 546, 2023, p. 126232.
[20] T. Markov et al., "A Holistic Approach to Undesired Content Detection in the Real World", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 15009-15018.
[21] H. Inan et al., "Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations", arXiv:2312.06674, 2023.