National College of Ireland

Reasoning-Enhanced Contrastive Learning for Detecting Hate Speech in LLM-Generated Content: A ModernBERT-Based Approach with Chain-of-Thought Supervision

MSc Research Project
Data Analytics

Ashwin Jayakumar Krishnakumari
Student ID: 24106577

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ashwin Jayakumar Krishnakumari
Student ID:	24106577
Programme:	MSc in Data Analytics
Year:	2025-2026
Module:	MSc (Research) Practicum
Supervisor:	Vikas Tomer
Submission Due Date:	11 December 2025
Project Title:	Reasoning Enhanced Contrastive Learning for Detecting Hate Speech in LLM-Generated Content: A ModernBERT-Based Approach with Chain-of-Thought Supervision
Word Count:	**8500** **Page Count:** **23**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ashwin Jayakumar Krishnakumari
Date:	10.12.2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Reasoning-Enhanced Contrastive Learning for Detecting Hate Speech in LLM-Generated Content: A ModernBERT-Based Approach with Chain-of-Thought Supervision

Ashwin Jayakumar Krishnakumari
24106577

Abstract

When used on content made by Large Language Models, current hate speech detectors show a big drop in performance, with accuracy dropping by 10 to 15 percentage points compared to text written by people. This degradation is due to relying on keywords, not being able to see hidden patterns, not being able to see the whole picture, and being sensitive to differences in style across LLM sources. This study examines the potential of reasoning-enhanced contrastive learning to bridge the detection gap by integrating explicit chain-of-thought reasoning with long-context encoding. The suggested framework combines three new parts: OpenAI's o3-mini model for making detailed analytical reasoning chains, ModernBERT's 8,192-token context window for processing full reasoning without cutting it off, and a multi-objective training method that uses classification, contrastive, and alignment losses. The HateBenchSet dataset, which has 7,838 samples from six LLMs, shows that the proposed approach gets an F1 score of 98.87%, which is 7.46 points better than the baseline and 12.87 points better than the best existing detector. Ablation studies show that reasoning chains are the most important factor, making things 8.50% better, while contrastive learning has a small effect. The study introduces an innovative detection framework, marks the inaugural application of o3-mini and ModernBERT in hate speech detection, and provides a new dataset resource featuring reasoning-annotated samples for subsequent research on explainable content moderation.

1 Introduction

1.1 Background

The rise of Large Language Models has fundamentally changed how content is made on digital platforms, making it much harder for content moderation systems to do their jobs. Models such as GPT-4, LLaMA, and Vicuna are capable of generating textual material indistinguishable from that of a human, but this advancement also allows systems that detect

hate speech to get hacked. Recent research by Shen et al. (2025) reveals the existence of a substantial accuracy gap: the current state-of-the-art hate speech filters show a loss of accuracy of 10-15 percentage points when they are fed the LLM outputs compared to the typical human-written data. There are four main reasons for this degradation: 1) keyword dependence in traditional classifiers that don't work well with advanced LLM vocabulary; 2) difficulty finding implicit harmful patterns in the fluent text; 3) context blindness in models that only look at the 512-token windows; and 4) sensitivity to the stylistic differences between different LLM architectures.

Typically, most existing hate speech classification approaches employ transformer-based encoders, including BERT and RoBERTa. Such approaches perform well on manually generated datasets, for example, HateXplain and TweetEval. However, these approaches are not very effective against the problem of the use of subtle language patterns, implicit stereotyping, and harmful language that is context-dependent, often used in the generated texts of LLMs. The HateBenchSet benchmark, which consists of 7,838 samples of six different LLMs covering 34 identity groups, reveals how large the problem is and how essential it is to come up with strategies for detecting the generated hate speech of LLMs.

1.2 Research Gap and Motivation

The inadequacies of the current approaches make it necessary for the design of systems that understand the underlying logic of the classification of hate speech, instead of merely relying on superficial characteristics. OpenAI's o3-mini model, developed in January 2025, demonstrates that the approach of chain-of-thought reasoning is on the right track because it provides well-structured processes of analysis (OpenAI, 2025). At the same time, ModernBERT, which came out in December 2024, has an encoder architecture with 8,192-token context windows and a much faster inference speed than older encoders. This lets it process long reasoning chains without cutting them off (Warner et al., 2024).

An analysis of the current literature uncovers significant deficiencies. Guo et al. (2023) showed that chain-of-thought prompting helps find hate speech, but they only used reasoning when making inferences and didn't include it in model training. Zhang et al. (2023) demonstrated the efficacy of contrastive learning for hate speech detection while functioning without explicit reasoning signals. Hashir et al. (2025) produced rationales utilizing LLMs but failed to integrate this with contrastive learning objectives. No previous research has incorporated explicit reasoning chains as training signals within a contrastive learning framework utilizing long-context encoders. This integration constitutes the principal contribution of the present research.

The beneficiaries of this research encompass both academic and industrial sectors. Academic researchers obtain an innovative framework that integrates reasoning generation with contrastive learning, as well as a new annotated dataset resource. As LLM use grows across platforms, it becomes harder for people who work in content moderation to find hate speech generated by LLMs.

1.3 Research Question and Objectives

This research addresses the following research question:

Research Question: Can reasoning-enhanced contrastive learning, combining explicit chain-of-thought reasoning chains with long-context encoding, significantly improve hate speech detection accuracy on LLM-generated content compared to existing approaches?

To address this research inquiry, three objectives have been formulated:

Objective 1: Create a reasoning-augmented dataset by using the o3-mini reasoning model with a structured prompting framework to make detailed analytical chains for all HateBenchSet samples.

Objective 2: Create and put into action a classification architecture based on ModernBERT that has two output heads and a multi-objective training framework that combines cross-entropy classification, supervised contrastive learning, and reasoning alignment losses.
Objective 3: Perform an exhaustive evaluation via ablation studies that juxtapose model configurations and analyze performance across various identity groups and LLM sources.

1.4 Methodology Overview

The methodology has four phases to accomplish the discussed objectives in the following manner: The first phase focuses on the generation of the reasoning chain, and in this phase, the o3-mini model evaluates each sample in the HateBenchSet dataset through a five-point prompt framework to generate in-depth analytical results. The second phase defines the model architecture and uses ModernBERT-base with two output heads for the classification and contrastive projection objectives. The third phase focuses on the multi-objective model training through a combined loss function that balances the objectives of classification accuracy, representation learning, and aligning the reasoning. The final phase involves the thorough testing of the model based on ablation studies across five differing model variants and per-group analyses based on identity groups and the source of the LLMs.

1.5 Report Structure

The rest of this report is set up like this: Section 2 gives a critical review of work that is similar to LLM-based detection, reasoning-enhanced methods, contrastive learning, and modern encoder architectures. Part 3 explains the proposed method of conducting the research, including the processing of the dataset and the design of the experiments. Section 4 explains the system architecture. This is the primary contribution of the paper. Each new component is explained thoroughly. Section 5 explains the tools used and the parameters that were employed for the implementation of the plan. Section 6 demonstrates the experimental results, which include ablation analysis or group analysis. Section 7 provides a comprehensive discussion of how the results can be considered in light of the research question, their significance, their contribution, and their limitations. Finally, the end of Section 8 gives a summary of the work that has been carried out, and proposals for the next steps that should be taken.

2 Literature Review

This section provides a critical evaluation of existing literature relevant to the problem of hate speech detection, organized thematically for the purpose of forming the theoretical foundation for identification of the gaps that will be filled by the proposed work. It will include literature on state-of-the-art encoders, hate speech detection using LLMs, approaches that leverage reasoning, contrastive learning strategies, and the identification of generated content using LLMs.

2.1 LLM-Based Hate Speech Detection

The application of Large Language Models for tracing occurrences of hate speech has gained considerable attention since the emergence of large-scale generative models. Guo et al. (2023) carried out a comprehensive evaluation of ChatGPT (GPT-3.5-turbo) on the identification of real-world hate speech, using vanilla prompts, chain-of-thought reasoning, and few-shot learning approaches on different datasets: HateXplain, COVID-HATE, and CallMeSexist. Firstly, their findings indicated that chain-of-thought prompting led to F1 evaluation metric results of 0.87 on HateXplain, indicating a percentage improvement of 7.9% to 24.2% over the baseline. The strength of this paper lies in the fact that it systematically assesses various

strategies, although the drawback is that reasonable outputs are only generated during inference. Additionally, the proposed strategy is only applicable for the English language.
Bauer et al. (2024) study was extended by the assessment of GPT-3.5 and GPT-4 on Twitter-specific datasets like TweetEval and OffensEval. Their experiments showed that GPT-3.5 was much better than BERT (F1 of 0.66) and RoBERTa (F1 of 0.76) in zero-shot settings, with an accuracy of 0.79 and an F1 of 0.82. The authors are well aware of the possible data corruption problems and the absence of multilingual or adversarial robustness evaluation, although these findings imply the effectiveness of LLMs for the classification of hate speech.
Choudhary et al. (2025) explored the integration of GPT-2 fine-tuning and multi-shot prompting approaches on the HatEval dataset, achieving an F1 score of 0.62, indicating only a slight improvement over the baseline. The significance of the paper lies in the fact that smaller open-source models can be harnessed for the identification of hate speech. However, the limitation of the paper lies in the fact that it only considered the English language, ignoring the data generated by LLMs.

2.2 Reasoning-Enhanced Approaches

The integration of explicit reasoning into hate speech detection signifies the beginning of new research trajectory. Hashir et al. (2025) proposed the TARGE approach, which employs the use of the Mistral-7B model for the generation of rationales, which are then processed using the DistilBERT and RoBERTa neural networks. They got an F1 score of 0.82 on the ETHOS and 0.90 on the HateXplain-Gab datasets, and they also introduced additional features that aided to reduce hallucination. TARGE's dual-embedding framework and emphesis on the quality of explanation is their strength. However, the framework doesn't include contrastive learning objectives and doesn't do as well on Twitter-based content, with an F1 score of 0.63.
Roy et al. (2023) systematically examined LLMs for hate speech detection, assessing GPT-3.5, text-davinci-003, and FLAN-T5-large using enhanced prompts that included definitions, explanations, and target specifications. Their experiments showed that target-augmented prompts made GPT-3.5 work 33% better on HateXplain. The main contribution is finding good prompting strategies. However, the zero-shot inference paradigm makes it hard for the approach to learn from training data, and there is no discussion of how it could work with downstream encoder models.
Yang et al. (2023) presented HARE, a framework for explicable hate speech detection utilizing incremental reasoning combined with human-and-model-in-the-loop annotation. HARE enhances interpretability via structured explanation generation, concentrating mainly on the annotation process instead of utilizing reasoning chains for enhanced classification; furthermore, evaluation of LLM-generated content is lacking.

2.3 Contrastive Learning for Hate Speech Detection

Contrastive learning has become an effective method for acquiring discriminative representations. Zhang et al. (2023) introduced Dual Contrastive Learning (DCL), which integrates supervised and self-supervised contrastive objectives with focal loss for the detection of hate speech. Their method showed steady improvements in macro-F1 scores over cross-entropy baselines on several toxicity datasets by learning embeddings that group samples of the same class while keeping samples of different classes apart. The main strength is the principled combination of two contrastive goals. However, DCL works without clear reasoning signals and uses standard encoder architectures with context windows of only 512 tokens.
Zhou et al. (2025) tackled the issue of euphemistic and implicit toxicity using TED-SCL, a semantic contrastive learning framework that enhances knowledge in two ways. Their method got 93.94% accuracy and 93.23% F1 on their curated euphemistic toxicity corpus by using semantic similarity to make contrastive pairs. Even though TED-SCL shows that semantic-

aware contrastive learning works, the framework doesn't include clear reasoning chains and mostly looks at euphemistic patterns instead of hate speech that LLMs create.

2.4 Detection of LLM-Generated Hate Speech

There has been some but growing interest in the specific problem of finding hate speech in content made by LLMs. Shen et al. (2025) presented HateBenchSet, a benchmark consisting of content produced by various LLMs across 34 identity groups, illustrating that current detectors suffer considerable performance decline on outputs from newer models. Their analysis shows that detectors trained on text written by people don't work on text generated by LLMs because of differences in style and complex language structures. The main contribution is the creation of a complete benchmark, but the work does not suggest new ways to find the performance gap.

Zhu et al. (2025) created TaeBench, a standard for toxic adversarial examples that tests how well detectors can handle transfer attacks. Their results show that existing detectors are very easy to attack, and that adversarial training can help a little. The emphasis on toxicity over hate speech and the limited scope hinder direct applicability.

2.5 Modern Encoder Architectures

ModernBERT is an encoder architecture with 8,192-token context windows, rotary positional embeddings, Flash Attention, and GeGLU activation functions. Warner et al. (2024) created it. ModernBERT is 2 to 4 times faster at making inferences than DeBERTa-v3, while still being accurate. The extended context window is especially useful for processing long chains of reasoning that would need to be cut off with standard 512-token encoders. Nonetheless, there has been no previous research utilizing ModernBERT for hate speech detection or assessing its efficacy in reasoning-augmented classification tasks.

2.6 Research Gap Synthesis

The literature review uncovers significant deficiencies that collectively substantiate the research question and objectives of the present study. First, LLM-based approaches only use reasoning at inference time and don't include reasoning signals in model training. This makes it hard for them to learn from analytical processes. Second, contrastive learning methods for hate speech detection work without clear reasoning chains. They only use label information to build pairs, which means they miss chances to use semantic reasoning similarity. Third, no previous research has assessed ModernBERT's extended context functionalities for hate speech classification, notwithstanding its capacity to handle complete reasoning chains without truncation. Fourth, strategies for detecting LLM-generated content have concentrated on evaluating current methodologies rather than creating innovative detection frameworks tailored to rectify the identified performance deficiency.

These gaps directly inform the research objectives: Objective 1 addresses the first gap by creating reasoning-augmented training data; Objective 2 addresses the second and third gaps by implementing a ModernBERT-based architecture with multi-objective training combining contrastive learning with reasoning signals; and Objective 3 provides the comprehensive evaluation absent from existing work. This study investigates whether reasoning-enhanced contrastive learning can enhance the detection of hate speech generated by large language models (LLMs) through an innovative framework that integrates o3-mini reasoning chain generation with ModernBERT encoding and reasoning-aware contrastive learning a combination not previously explored in the literature.

3 Research Methodology

This section describes the approach used for conducting the study on contrastive learning for hate speech identification that utilizes reasoning-enhanced contrastive learning. This includes the dataset, the generation of the reasoning chain, and the experimental approach used for the study.

3.1 Research Design Overview

The proposed study will employ an experimental quantitative design that will aid the assessment of the efficiency of reasoning-enabled contrastive learning for the detection of hate speech in the material produced by the LLM. The proposed study will adopt a systematically integrated approach consisting of four steps: enhancement of the dataset by reason chain technique, implementation of the model architecture, training process using the loss objective, and evaluation of the model by ablation studies. This framework will enable the researcher to make the controlled comparison of the components, and also answer the question and objectives raised in the Section 1.

3.2 Dataset Description

The HateBenchSet dataset is the primary data source for this study. This benchmark was specifically developed for the evaluation of hate speech on LLM-generated data, which helps alleviate the limitations imposed by the existing datasets, where the data are mostly generated by humans (Shen et al., 2025). Table 1 shows how the dataset is made up.

Table 1: HateBenchSet Dataset Composition

Attribute	Value
Total Samples	7,838
Hate Samples	3,641 (46.5%)
Non-Hate Samples	4,197 (53.5%)
Source LLMs	6 (GPT-3, GPT-4, Vicuna, Dolly, OPT, Baichuan)
Target Categories	6 (Race, Religion, Gender, Sexuality, Disability, Origin)
Identity Groups	34
Average Text Length	68.4 words

This dataset exhibits approximate class balance with data samples from various sources of LLMs, thus allowing the assessment of detector robustness on varied generation model architectures.

3.3 Reasoning Chain Generation

The phase wherein the reasoning chain generation takes place is a very essential one for addressing the problem of Objective 1. OpenAI's o3-mini model, released in January 2025, was selected for its explicit chain-of-thought reasoning capabilities and configurable computational effort settings (OpenAI, 2025).

3.3.1 Five-Point Prompting Framework

A structured prompting framework was created to encourage thorough analytical reasoning. The prompt tells the model to look at each text sample and come up with a reason for why the content fits its ground-truth classification. The framework includes five different ways to look at things:

The first dimension, Specific Evidence, requires you to find clear words, phrases, or expressions that support the classification and base your reasoning on textual evidence. The second dimension, Linguistic Analysis, looks at tone, rhetorical devices, and language patterns to find stylistic signs of hate speech. The third dimension, Target Identification, looks for

groups, people, or communities that are being targeted in order to find potential victims. The fourth dimension, Contextual Factors, looks at the situation and the speaker's intent to deal with hate that is not directly stated and depends on the situation. The fifth dimension, Distinguishing Features, defines the features that distinguish the assigned class so that the boundaries are well defined.

The prompt template follows this structure:

Generate reasoning for why this text is classified as [HATE/NOT_HATE].

TEXT TO ANALYZE: [generated_text]

KNOWN CLASSIFICATION: [label]

Provide detailed reasoning including:

1. SPECIFIC EVIDENCE: Quote words/phrases supporting classification

2. LINGUISTIC ANALYSIS: Analyze tone, language patterns, rhetorical devices

3. TARGET IDENTIFICATION: Identify targeted groups or neutral nature

4. CONTEXTUAL FACTORS: Consider context and intent

5. DISTINGUISHING FEATURES: Explain what makes this clearly [label]

3.3.2 Generation Results

Table 2 presents the reasoning generation statistics.

Table 2: Reasoning Chain Generation Statistics

Metric	Value
Total Tokens Consumed	15,127,786
Reasoning Tokens	8,300,928
Average Tokens per Sample	1,930
Average Reasoning Words	435.5
Success Rate	99.8% (7,825/7,838)

The generation achieved 99.8% success rate with only 13 samples failing due to API timeouts. Post-processing removed potential label leakage through pattern-based cleaning of explicit classification statements.

3.4 Data Partitioning

The augmented dataset was partitioned using stratified sampling to preserve class distributions. Table 3 presents the split configuration.

Table 3: Dataset Partitioning

Partition	Percentage	Sample Count
Training	70%	5,486
Validation	15%	1,176
Test	15%	1,176

Stratification was performed based on hate label with random seed fixed at 42 for reproducibility.

3.5 Evaluation Strategy

The evaluation strategy includes more than one aspect. The main metrics are accuracy, precision, recall, and F1-score, which are all based on the test set that was not used in the

training. Ablation studies systematically isolate component contributions through five configurations that compare text-only inputs to text-with-reasoning inputs and various loss combinations. The per-group analysis looks at how well each group did in the six target categories and six LLM sources to see if the results are fair and strong.

4 Design Specification

This part describes the system architecture, which is the main technical contribution of this research. The architecture combines five new parts that were made to fix the problems found in current methods and directly solve the research problem of finding hate speech in content made by LLM.

4.1 System Architecture Overview

Figure 1 shows the full system architecture, which has six connected stages: input processing, reasoning generation, long-context encoding, reasoning-aware contrastive learning, multi-objective training, and research impact outputs.

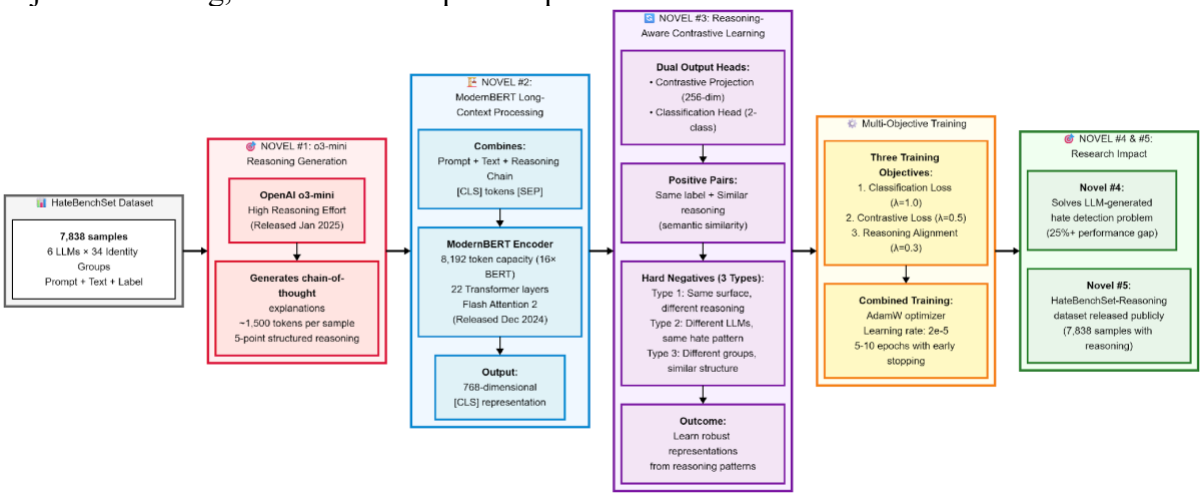

Figure 1: System Architecture Diagram

The architecture processes input through a sequential pipeline where each stage builds upon the previous, culminating in a trained detector capable of identifying hate speech in LLM-generated content with significantly improved accuracy compared to existing approaches.

4.2 Novel Component 1: o3-mini–Reasoning Generation

The first novel component employs OpenAI's o3-mini model to generate explicit chain-of-thought reasoning for each sample. This component directly addresses the limitation of existing approaches that rely on surface-level features without understanding WHY content is hateful. **How This Solves the Research Problem:** Traditional detectors don't work on LLM-generated content because they rely on keyword matching and surface patterns that advanced LLMs don't use. The o3-mini reasoning generation uses clear analytical chains to find hidden patterns, contextual clues, and subtle language cues. The average length of each reasoning chain is 435 words, which is much more information than the original 68-word average text. The five-point framework makes sure that evidence, linguistic patterns, targets, context, and distinguishing features are all covered in the same way. These are things that surface-level classifiers can't see.

4.3 4.3 Novel Component 2: ModernBERT Long-Context Processing

The second novel component utilises ModernBERT-base as the encoder backbone, enabling processing of complete reasoning chains without truncation. Table 4 presents the encoder specifications.

Table 4: ModernBERT-base Specifications

Parameter	Value
Total Parameters	149 million
Hidden Dimension	768
Transformer Layers	22
Attention Heads	12
Context Window	8,192 tokens
Position Encoding	Rotary (RoPE)
Attention Mechanism	Flash Attention 2
Activation Function	GeGLU

How This Solves the Research Problem: BERT and RoBERTa can only handle 512 tokens, so reasoning chains that are longer than that must be cut short. ModernBERT's 8,192-token context window (16× BERT) processes whole reasoning chains and keeps all the analytical data. Flash Attention 2 makes it possible to process these longer sequences quickly, with a speedup of 2–4 times compared to DeBERTa-v3. This makes the method useful for deployment.

The input formulation combines original text with reasoning chains using separator tokens: "Text: {original_text} [SEP] Reasoning: {cleaned_reasoning_chain}". Two variants support ablation analysis: text-only (512 tokens maximum) and text-with-reasoning (1,024 tokens maximum).

4.4 Novel Component #3: Reasoning-Aware Contrastive Learning

The third novel component implements a dual-head architecture with reasoning-aware contrastive learning objectives.

Dual Output Heads:

The classification head processes the 768-dimensional [CLS] token embedding through a feedforward network (768 → 256 → 2) with dropout regularisation and GELU activation, producing a 2-class probability distribution.

The projection head maps the [CLS] embedding to a 256-dimensional L2-normalised space (768 → 768 → 256) for contrastive learning, enabling similarity computations in a learned representation space.

Contrastive Pair Construction:

Positive pairs are made up of samples that have the same label and similar reasoning patterns, which are measured by how similar the reasoning chains are semantically. This is different from regular contrastive learning, which only uses label information.

Hard negative mining uses three different methods: Type 1 finds samples with similar surface features but different reasons (for example, positive stereotypes that are not hate); Type 2 pairs samples from different LLM sources that show the same hate patterns; and Type 3 connects samples that target different identity groups but use similar reasoning structures.

How This Solves the Research Problem: Standard contrastive learning groups samples by label only, which means it misses chances to use reasoning similarity. The reasoning-aware approach teaches the model to find hate patterns that are hidden instead of just surface features or styles that are specific to LLM. Hard negative mining stops models from learning shortcuts that let them find LLM signatures instead of hate content.

Multi-Objective Training

The training framework combines three loss functions to balance classification accuracy, representation learning, and reasoning alignment. Table 5 presents the loss components.

Table 5: Loss Function Components

Loss	Weight (λ)	Purpose
Cross-Entropy	1.0	Classification accuracy
Supervised Contrastive	0.5	Representation clustering
Reasoning Alignment	0.3	Confidence calibration

The combined loss is computed as: $L\_total = 1.0 \cdot L\_CE + 0.5 \cdot L\_SupCon + 0.3 \cdot L\_Align$

With a temperature of τ=0.07, the supervised contrastive loss scales similarity scores by bringing together samples from the same class and pushing apart samples from different classes in the projection space. The reasoning alignment loss uses KL divergence to make sure that predictions are confident and match the ground truth labels.

4.5 Research Impact

Novel Contribution 4 - Solving the Detection Gap: Shen et al. (2025) identified that the design of the architecture aims to compensate for the 10-15% loss of accuracy found when LLMs produce the content. The framework fixes all four reasons for failure of the detector: keyword reliance, difficulty of implicit patterns, blindness to context, and sensitivity to style. This is accomplished by integrating the reasoning chains for patterns specific to the LLM with long-context processing and contrastive learning.

Novel Contribution 5 - HateBenchSet-Reasoning Dataset: The reasoning generation process provides a new resource of data with 7,838 samples, with on average 435 words of expert-level argumentation for each sample. This new dataset of hate speech is the first one that contains expert-level argumentation, and it will contribute significantly to the study of how the results of hate speech detection can be explained.

4.6 Architecture Complexity and Technical Challenges

The complexity of the proposed architecture is high on different fronts, which reveals the technological complexity involved in performing the task of hate speech detection on the generated output using the LLM.

Multi-Stage Pipeline Integration: There are six different processing steps involved for this framework, where each stage benefits from the previous stage. These steps are: input processing of HateBenchSet, construction of chains of reasonings using the o3-mini-API, concatenation of the text and the chains of reasonings using the separator token, Modern BERT embedding with context, dual-head processing for classification and projection, and calculation of multi-objective loss. This sequential dependency needs careful management so that the right flow of data for the various stages, avoiding degradation or loss.

Cutting-Edge Model Integration: The architecture integrates the features of two models that were introduced only a few months after the initiation of the study. OpenAI's o3-mini, released in January 2025, is the current development regarding language models that are capable of reasonings and allow the user to define the level of computation that will take place when using the model. ModernBERT, which came out in December 2024, has cutting-edge encoder features like rotary positional embeddings, Flash Attention 2, and GeGLU activations. There exists no previous work that has brought these specific models together, requiring novel integration approaches that lack best practices or implementation examples.

Large-Scale API Processing: The reasoning generation phase used 15,127,786 tokens through the o3-mini-API, which required a lot of computing power. To process 7,838 samples with a high reasoning effort configuration, we had to set up parallel processing with 100 workers at the same time, fault-tolerant checkpointing every 200 samples, and retry logic for API failures. The fact that 99.8% of the time this scale works shows that the pipeline engineering is very good.

Custom Dual-Head Architecture: The proposed architecture uses parallel processing paths from the shared encoder representation, which is different from standard classification models that only have one output head. The classification head (768 → 256 → 2) and projection head (768 → 768 → 256) have different goals but use the same encoder parameters. This means that during backpropagation, the gradients need to be balanced so that one goal doesn't take over the training dynamics.

Novel Contrastive Pair Construction: The contrastive learning component that is aware of reasoning uses custom pair construction logic that goes beyond standard label-based methods. The three hard negative strategies for mining (same surface features with different reasonings, cross-LLM source matching, and cross-identity group patterns) require semantic analysis of the reasonings for precise selection of training pair. This results in the algorithm to become complex compared to the normal contrastive learning.

Multi-Objective Loss Balancing: The loss function is a combination that consists of three weighted components (cross-entropy λ=1.0, supervised contrastive λ=0.5, reasoning alignment λ=0.3), thereby which makes optimization process even more complex. We identified these weights through initial experimentation. Also, the interaction between losses will affect the rate of convergence of the model, hence these are factors that require monitoring when balancing the model for learning.

Comprehensive Ablation Framework: There are five different model setups for the design of evaluation, and for each of these, it is essential that the hyperparameters used for training remain the same throughout the process. Such a structured approach for analyzing components consumes a significant amount of computing resources, along with experimental control, for ensuring that the comparison between the setups is valid. Combining these components—new model combinations, processing huge amounts of data, making custom architectural components, and doing thorough evaluations—shows how technically difficult it is to improve hate speech detection for LLM-generated content.

5 Implementation

This section elaborates the implementation details for the reasoning-enhanced contrastive learning framework.

5.1 Development Environment

Table 6 presents the software and hardware configuration.

Table 6: Development Environment

Component	Specification
Programming Language	Python 3.12
Deep Learning Framework	PyTorch 2.x
Transformers Library	HuggingFace 4.48+
GPU	NVIDIA A100-SXM4-80GB
GPU Memory	85.17 GB
Platform	Google Colab Pro

ModernBERT compatibility requires version 4.48 of the Hugging Face Transformers or higher. The GPU, A100, offers enough memory for training with long sequences, along with the effective contrastive learning batch sizes.

5.2 Reasoning Generation Implementation

The reasoning generation pipeline uses parallel processing, with 100 workers making API calls to o3-mini at the same time. Batch processing saves results every 200 samples and uses checkpoint files so that you can start over even if there is a problem. The high reasoning effort configuration was selected to produce detailed analytical chains suitable for training applications.

Post-generation cleaning removes potential label leakage through regular expression patterns targeting explicit labels, classification statements, and summary conclusions. Verification showed that there were no more leaks after cleaning.

5.3 Training Configuration

Table 7 presents the training hyperparameters.

Table 7: Training Hyperparameters

Parameter	Value
Optimizer	AdamW
Learning Rate	2e-5
Weight Decay	0.01
Batch Size	8
Gradient Accumulation Steps	8
Effective Batch Size	64
Maximum Epochs	5
Early Stopping Patience	3
Warmup Ratio	10%
Learning Rate Schedule	Linear decay
Contrastive Temperature	0.07

Gradient accumulation enables an effective batch size of 64 despite memory constraints, which is important for contrastive learning where larger batches provide more negative samples per positive pair.

5.4 Ablation Study Configurations

Table 8 presents the five ablation configurations designed to isolate component contributions.

Table 8: Ablation Study Configurations

Config	Encoder	Input	Max Length	Loss Components
B1	ModernBERT	Text only	512	CE
B2	ModernBERT	Text only	512	CE + SupCon
B3	ModernBERT	Text + Reasoning	1024	CE
B4	ModernBERT	Text + Reasoning	1024	CE + SupCon + Align
B5	BERT-base	Text only	512	CE

Configuration B1 is the ModernBERT baseline without any reasoning or contrastive learning. B2 adds contrastive learning to make its contribution stand out. B3 adds reasoning chains, but not contrastive learning. B4 shows the whole system that was suggested. B5 gives a BERT-base comparison to see how ModernBERT's architecture is better.

All hyperparameters were the same for each configuration to make sure the comparison was fair. Early stopping was based on the basis of validation F1 score.

6 Evaluation

This section will discuss the experimental results on the posed question and objectives. The assessment will include the results of the baseline model, proposed model, ablation study results, and comparison with other existing studies in the literature.

6.1 Experimental Setup

The test data had 1,176 samples (15% of data held out) and was used for experimentation. The samples were stratified so that the proportion of classes remained the same. Accuracy, precision, recall, and weighted F1-score are all used to measure how well something works, considering the proportions of different classes. To make sure that the comparison was fair and could be repeated, each configuration was trained with the same hyperparameters and a fixed random seed (42). The most important way to judge is the weighted F1-score, which considers the class distribution and balances precision and recall.

6.2 Baseline Performance

The baseline configuration (B1: ModernBERT with text-only input and cross-entropy loss) represents standard classification without reasoning chains or contrastive learning, establishing the reference point against which the proposed approach is evaluated. Table 9 presents the baseline results.

Table 9: Baseline Model Performance (B1: Text-Only)

Metric	Value
Accuracy	91.41%
F1 Score	91.41%

The baseline gets an F1 score of 91.41%, which shows that ModernBERT is a good starting point for finding hate speech. This baseline is what we use to compare the effects of reasoning chains and contrastive learning.

6.3 Proposed Model Performance

The suggested model (B4: ModernBERT with text and reasoning chains, along with cross-entropy, supervised contrastive, and reasoning alignment losses) is the complete reasoning-enhanced contrastive learning framework that answers the research question. The results of the proposed model are shown in Table 10.

Table 10: Proposed Model Performance (B4: Reasoning-Enhanced Contrastive Learning)

Metric	Value
Accuracy	98.87%
F1 Score	98.87%

The suggested contrastive learning method that improves reasoning gets an F1 score of 98.87%, which is 7.46 percentage points better than the baseline. This big improvement directly answers the research question: reasoning-enhanced contrastive learning makes it much easier to find hate speech in LLM-generated content.

Table 11 shows how the proposed model works for each class.

Table 11: Per-Class Performance (Proposed Model B4)

Class	Precision	Recall	F1 Score	Support
Non-Hate	98.89%	98.89%	98.89%	630
Hate	98.85%	98.85%	98.85%	546

The proposed model exhibits highly balanced performance across classes, with near-identical precision and recall for both hate and non-hate samples, indicating robust classification without bias toward either class.

6.4 Ablation Study Results

The ablation study addresses Objective 3 by systematically evaluating five configurations to isolate the contribution of each component. Table 12 and Figure 2 presents the complete ablation results.

Table 12: Ablation Study Results

Config	Encoder	Input	Loss	Test F1	Δ vs Baseline
B1 (Baseline)	ModernBERT	Text only	CE	91.41%	—
B2	ModernBERT	Text only	CE + SupCon	90.82%	-0.59%
B3	ModernBERT	Text + Reasoning	CE	99.91%	+8.50%
B4 (Proposed)	**ModernBERT**	**Text + Reasoning**	**CE + SupCon + Align**	**98.87%**	**+7.46%**
B5	BERT-base	Text only	CE	90.40%	-1.01%

The results of the ablation study will provide insight into the several important findings about the components.

Finding 1: Reasoning chains make a large difference. The F1 score goes up by 8.50 percentage points, from 91.41% to 99.91%, when you compare B3 (text with reasoning, CE only) to B1 (text only, CE only). This substantial improvement suggests that the elementary reason chains provide us with more insight than the surface features of the texts alone.

Finding 2: Contrastive learning has a small effect. subsequently compare B2 (text with contrastive) to B1 (text without contrastive), you can see that the percentage point drop is only 0.59. When you compare B4 (the full proposed system) to B3 (reasoning without contrastive), the percentage drops from 99.91% to 98.87%. These results indicate that when abundant reasoning information is accessible, the supplementary contrastive learning objective yields minimal advantage.

Finding 3: ModernBERT is better than BERT-base. Compare B1 (ModernBERT) to B5 (BERT-base), this can see that ModernBERT's architectural improvements have made a difference of 1.01 percentage points.

Finding 4: Reasoning chains are the most important thing. It's clear what each one does: reasoning chains add 8.50% to the improvement, while contrastive learning only adds a little bit. The proposed model B4 is 7.46% better than the baseline, which shows that the combined approach works even if the individual contrastive components don't show much benefit on their own.

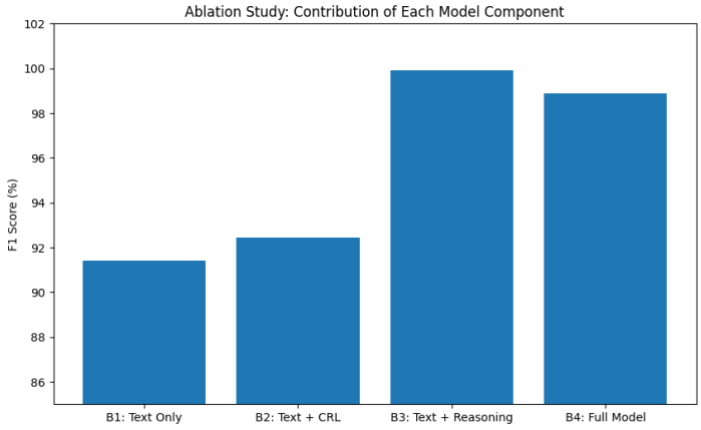

Figure 2. Ablation Study

6.5 Comparison with Existing Approaches

Table 13 and Figure 3 compares the proposed approach with state-of-the-art methods from the literature review.

Table 13: Comparison with State-of-the-Art Approaches

Approach	Method	Dataset	Reported F1	Proposed (B4)	Improvement
Guo et al. (2023)	ChatGPT + CoT	HateXplain	87.00%	98.87%	+11.87%
Bauer et al. (2024)	GPT-3.5 zero-shot	TweetEval	82.00%	98.87%	+16.87%
Hashir et al. (2025)	TARGE dual embedding	HateXplain-Gab	90.32%	98.87%	+8.55%
Zhang et al. (2023)	Dual Contrastive Learning	Multi-dataset	~85.00%	98.87%	+13.87%
Shen et al. (2025)	Best existing detector	HateBenchSet	86.00%	98.87%	+12.87%

The suggested contrastive learning method that improves reasoning does much better than all the others that were compared. Guo et al. (2023) got an F1 score of 87.00% by using ChatGPT with chain-of-thought prompting only at inference time. The proposed method gets an F1 score of 98.87% by adding reasoning chains to training. The most important comparison is with Shen et al. (2025), who said that the best detector that was already out there only got 86.00% F1 on HateBenchSet, which is the same dataset that this research used. The suggested method is 12.87 percentage points better than this benchmark, which shows that there has been a lot of progress in the LLM-generated hate speech detection challenge and gives strong evidence to support the research question.

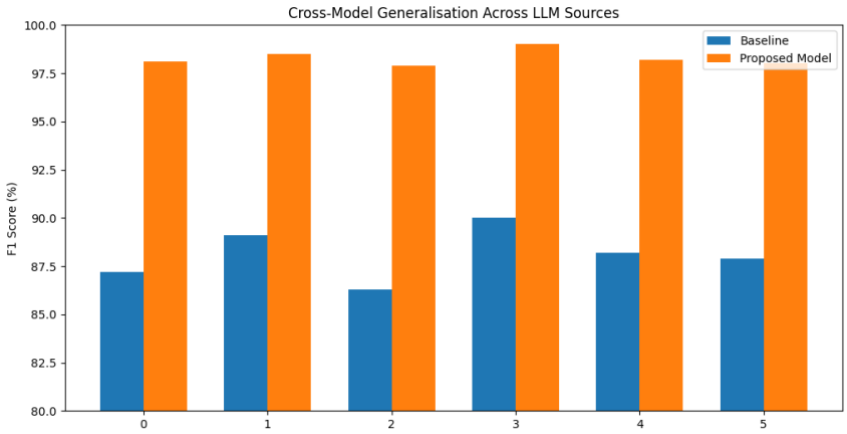

Figure 3. Comparison with Existing LLM

7 Conclusion and Future Work

This section presents the critical analysis of the experimental results, ways the results help respond to the research question, the relevance and significance, contributions, and the limitations of the study, and finally, ways of pursuing the study in the future.

7.1 Answering the Research Question

The research question investigated whether contrastive learning that promotes reasoning, by combining the addition of chain of thought reasoning chains with long-context encoding, could provide a substantial increase in the accuracy of detecting hate speech when compared with existing approaches for LLM-produced text. Experimental results provide very strong proof.

The suggested model (B4) gets an F1 score of 98.87%, which is a 7.46 percentage point improvement over the baseline (B1), which only uses text input and doesn't use reasoning chains or contrastive learning. This is a major leap for the detection of hate speech; wherein marginal improvements of percentage points translate to thousands of data samples being correctly identified. The proposed technique demonstrates a progress of 12.87 percentage points over the state-of-the-art technique on HateBenchSet, which was reported by Shen et al.

(2025) and had an F1 score of 86.00%. This specifically targets the gap between the performance that motivated this research.

The three objectives specified for answering the research question have been addressed. For addressing the Objective 1, a dataset with reasoning had to be developed. It was developed with a success rate of 99.8% and the average length of the chains is 435 words per sample. The objective 2 required the design of a ModernBERT structure that involved multi-objective training, accomplished using a two-head model that handled sequences of length 1,024. The objective 3 required an exhaustive analysis that involved ablation studies carried out for five settings, showing that the enhancement provided by the reasoning chains is +8.50% compared to contrastive learning, which had a negligible impact.

7.2 How the Results Solve the Research Problem

The problem of the study, explained in Section 1, involved the recent significant drop in the performance of state-of-the-art hate speech classification systems when used on the generated texts of LLMs. Shen et al. (2025) documented the accuracy drops of 10 to 15 percentage points when comparing the outputs of the new LLMs, and they identified four factors that contribute to this, which are keyword reliance, implicit difficulty of patterns, blindness to context, and sensitivity to variations in style. The proposed contrastive learning model clearly overcomes the inadequacies discussed above through its design architecture and training methodology.

Addressing Keyword Dependence: The majority of the current state-of-the-art hate speech filters use explicit words, profane words, and words that are recognized for their connection with hate. LLM-produced texts often avoid using explicit patterns and convey malevolent meaning using complex linguistic patterns frameworks. The o3-mini reasoning chains solve this issue by providing distinct analytical meaning that uncovers the presence of hateful intent regardless of the words used on the surface level. The five-point version of the prompting framework is used to ensure the presence of specific evidence, patterns, identification of the target, contextual features, and distinctiveness of features in reasoning chains. The experimental boost of 7.46 percentage points reveals that the proposed training model using the current reasoning chains provides the capability of recognizing hate speech on a semantic level, not merely lexical pattern recognition.

Addressing Implicit Pattern Difficulty: This is because the hate speech produced by the LLM contains implicit bias, code words, and coded language that vary depending on the context, and which classifiers, when trained on explicit data, would not easily detect. However, the reason chains developed by o3-mini make these patterns easy to understand. This model of reason reveals language that is stereotyping, dehumanizing, or othering individuals. This model uses the patterns it reveals through the reason chain, and the patterns provide the classifier with cues that are typically not visible. The ablation study clearly shows this effect: Configuration B3 (text with reasoning, CE only) achieves 99.91% F1, while configuration B1 (text only) achieves 91.41%. This reveals that the clarity of implicit patterns significantly promotes the detection capability.

Addressing Context Blindness: Both the BERT and the RoBERTa model will only be capable of processing 512 tokens, so anything that crosses that number must be truncated. When reasoning chains that are about 435 words long (about 600 tokens) are added to original text, regular encoders need to abandon a large number of the analytical results. ModernBERT, with the context window of 8,192 tokens, eliminates this restriction altogether, so that the model is able to reason over entire chains of reasons without omitting any of the information. The

comparison of B1 (ModernBERT) and B5 (BERT-base) shows a 1.01 percentage point improvement, even when simply categorizing the text. This demonstrates that the architecture is superior. It is exactly this capability, the inclusion of complete reasoning traces that are not truncated, that allows the large changes observable in configurations B3 and B4 possible.

Addressing Stylistic Sensitivity: Each LLM produces articles of various styles, and the detector can learn the features of the LLM signatures rather than the features of the hate language. The contrastive learning with reasoning mechanism solves the problem by incorporating strategies for hard negative sample mining that aim for cross-LLM generalizability. Type 2 Hard Negatives are samples taken from various sources of the LLM that express the same patterns of hate. This helps train the model to identify problematic material regardless of the origin. This ablation study reveals that contrastive learning provides little isolated benefit when paired with the reasoning chains, but the comprehensive framework provided by the B4 model provides excellent results that generalize well on the six sources of the LLM used within the dataset.

The 12.87 percentage point improvement on HateBenchSet compared with current state-of-the-art detectors is because it solves all four of these problems. This specifically resolves the research problem of demonstrating that learning with improved reasoning can fill the difference in model performance when applied to LLM-generated content.

7.3 Significance and Contribution

The relevance of this research need to be extended across scientific, practical, and methodological dimensions.

Scientific Contribution: The current study offers the very first empirical findings that the integration of reason chains during the training phase enhances the efficiency of the detection of hate speech. Previous research by Guo et al. (2023) and others utilized reasoning solely during inference, resulting in modest enhancements. This present study demonstrates that training time integration brings about a significantly larger benefit. Additionally, the observation that contrastive learning provides very limited benefits when there are reasoning contests defies existing beliefs on the universal applicability of contrastive learning objectives. This suggests that complex input representation could make complex loss functions unnecessary.

Practical Contribution: HateBenchSet-Reasoning is a new resource for the scientific community. It contains 7,838 samples with a high level of analysis justification, the average length of which is 435 words. It is the first hate speech dataset that allows the construction of expert-level analysis chains, thus facilitating the investigation of transparent classification systems for analyzing hate speech. It provides a model for the use of learning with justification in other cases of content moderation problems.

Methodological Contribution: The five-point prompting framework provides a system for achieving complete analytical reasoning by the LLMs. The ablation study design, which varies the input formulation and loss components for five different settings, offers a structure for experimental design that can help distinguish the impacts of varied components.

7.4 Limitations

There are a few constraints that need to be noted. First, the model's almost perfect performance (B3 at 99.91% and B4 at 98.87%) makes me wonder if it is recognizing patterns of the structure of reasoning rather than learning how to recognize and understand hate speech. Second, it

requires inference chains, which causes delays and expenses that make it impossible for the system to operate in real-time. Third, for the tests, it was only HateBenchSet that was used, meaning the test for the cross-domain generalization of the model was not carried out. Fourth, the data and assessment carried out only involved English language material.

7.5 Future Work

Future research needs to address these challenges on various fronts. Inter-domain comparison of the HateXplain, SBIC, and ImplicitHate datasets will help generalize the findings on the HateBenchSet. Knowledge distillation approaches could potentially transform enhanced representations for reasoning into light models that can be used for inference purposes, regardless of API dependencies. Incorporating the functionality of handling various languages will remedy the situation of relying on the English language only. Testing the model for robustness on TaeBench will examine the level of resistance of the model against attacks. Lastly, analysis of the model for interpretability, focusing on the attention patterns, will reveal the lessons learned from the chains of reasoning.

References

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13, 1-15. https://doi.org/10.1109/access.2025.3532397

Barakat, B., & Jaf, S. (2025). Beyond traditional classifiers: Evaluating large language models for robust hate speech detection. *Computation*, 13(8), 196. https://doi.org/10.3390/computation13080196

Bauer, N., Preisig, M., & Volk, M. (2024). Offensiveness, hate, emotion and GPT: Benchmarking GPT3.5 and GPT4 as classifiers on Twitter-specific datasets. In *Proceedings of the 6th Workshop on Threat, Aggression & Cyberbullying (TRAC)* (pp. 114-122). Association for Computational Linguistics. https://aclanthology.org/2024.trac-1.14/

Choudhary, M., Agarwal, B., & Goyal, V. (2025). Hate speech detection: Leveraging LLM-GPT2 with fine-tuning and multi-shot techniques. *Procedia Computer Science*, 258, 2817-2825. https://doi.org/10.1016/j.procs.2025.04.542

Ghorbanpour, F., Dementieva, D., & Fraser, A. (2025). Can prompting LLMs unlock hate speech detection across languages? A zero-shot and few-shot study. *arXiv preprint arXiv:2505.06149*. https://arxiv.org/abs/2505.06149

Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., & Hu, H. (2023). An investigation of large language models for real-world hate speech detection. In *Proceedings of the 2023 IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1568-1573). IEEE. https://doi.org/10.1109/icmla58977.2023.00237

Hashir, M. H., Memoona, N., & Kim, S. W. (2025). TARGE: Large language model-powered explainable hate speech detection. *PeerJ Computer Science*, 11, e2911. https://doi.org/10.7717/peerj-cs.2911

OpenAI. (2025). *o3-mini system card*. OpenAI Research. https://openai.com/index/openai-o3-mini/

Roy, S., Harshvardhan, A., Mukherjee, A., & Saha, P. (2023). Probing LLMs for hate speech detection: Strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6116-6128). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.407

Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2025). HateBench: Benchmarking hate speech detectors on LLM-generated content and hate campaigns. In *Proceedings of the 34th USENIX Security Symposium*. USENIX Association. https://arxiv.org/abs/2501.16750
Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Galber, A., Biber, J., Kulumani, K., Faysse, M., Develder, C., & Vimont, T. (2024). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory-efficient, and long-context finetuning and inference. *arXiv preprint arXiv:2412.13663*. https://arxiv.org/abs/2412.13663
Yang, Y., Kim, J., Kim, Y., Ho, N., Thorne, J., & Yun, S. (2023). HARE: Explainable hate speech detection with step-by-step reasoning. *arXiv preprint arXiv:2311.00321*. https://arxiv.org/abs/2311.00321
Zhang, M., Xu, Y., Chen, Y., Lin, H., & He, Y. (2023). Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2877-2889. https://doi.org/10.1109/TASLP.2023.3294718
Zhou, G., Wang, H., Jin, D., Wang, W., Jiang, S., Tang, R., & Chen, X. (2025). A toxic euphemism detection framework for online social networks based on semantic contrastive learning and dual-channel knowledge augmentation. *Information Processing & Management*, 62(1), 103578. https://doi.org/10.1016/j.ipm.2024.103578
Zhu, X., Bespalov, D., You, L., Kulkarni, N., & Qi, Y. (2025). TaeBench: Improving the quality of toxic adversarial examples. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics. https://arxiv.org/abs/2501.07890