AntiHateAgent: A Knowledge-Augmented Evidence-Based Reasoning Agent for Implicit Hate Speech Detection

Anonymous ACL submission

Abstract

Hate speech poses a growing challenge to online platforms, particularly as it becomes increasingly implicit and subtle. While recent advances in machine learning have improved automated detection, they largely rely on static internal knowledge or explicit semantics, leading to a critical knowledge gap and a lack of comprehensive reasoning. To address these limitations, we propose AntiHateAgent, an agent-based framework for hate speech detection that enables structured reasoning, contextual knowledge integration, and transparent decision-making, thereby improving robustness and reliability in real-world content moderation scenarios. Experimental results show that AntiHateAgent significantly improves the performance of implicit hate speech detection. On three datasets, it achieves up to a 22.3% increase in overall F1 score and up to a 43.6% improvement in recall for hate samples. The framework excels particularly in detecting newly emerging implicit hate that relies on cultural context, reaching 89.7% recall on the latest 4chan dataset implicit hate subset. Its evidence-driven reasoning process also ensures explainability and transparency in decision-making.

Warning: This paper contains contents that may be offensive or upsetting.

1 Introduction

Hate speech entails public prejudice, hostility, or offensive expression targeting specific individuals or groups based on identity characteristics such as race, ethnicity, or gender (Waldron, 2012). The spread of hate speech through online platforms has become a critical social concern as it reinforces discrimination against targeted communities (Howard, 2019; Matamoros-Fernández and Farkas, 2021). In response, recent years have witnessed extensive progress in automated hate speech detection, driven by both the advances in model architectures, from feature-engineering-based machine learning (Waseem and Hovy, 2016; Davidson et al., 2017) to deep learning (Caselli et al., 2021) and LLMs (Albladi et al., 2025), and the construction of large and diverse annotated datasets (Fortuna and Nunes, 2018; Poletto et al., 2021; Alkomah and Ma, 2022).

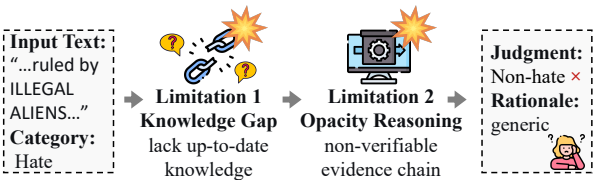

Figure 1: Limitations of existing zero-shot LLM-based detection of implicit hate speech.

Despite significant progress, detecting implicit hate speech still faces critical challenges (Ocampo et al., 2023b; Huang et al., 2023). Unlike **explicit** hate speech, which often contains overtly offensive terms such as racial slurs (Schmidt and Wiegand, 2017), **implicit** content requires knowledge of cultural codes and the ability to infer hidden meanings (Warner and Hirschberg, 2012). For example, *"how is Mexico doing these days? people come here because you couldn't build it"* constitutes implicit hate speech as it implicitly conveys the stereotype that *"Mexicans are incompetent"* (ElSherief et al., 2021). These characteristics make detection of implicit hate speech challenging from both data and model perspectives (Caselli et al., 2020). From the data perspective, human annotation of implicit hate speech is inherently difficult due to reliance on background knowledge, leading to low inter-annotator agreement and label noise (Talat et al., 2017). From the model perspective, representing implicit hate speech necessitates reasoning over context and cultural references, in comparison to the simple lexical cues in explicit forms (Zhang and Luo, 2019; Corazza et al., 2020).

Given their strong semantic understanding and reasoning capabilities, LLMs have recently been explored for implicit hate speech detection (Zhu et al., 2023; Li et al., 2024; Zeng et al., 2025; Giorgi et al., 2025). Several studies evaluate their zero-shot performance, showing that LLMs exhibit stronger reasoning and understanding capability (Roy et al., 2023; Das et al., 2024) compared to traditional classifiers. However, existing work also reports two key limitations, as shown in Figure 1. First, current approaches employ LLMs in a passive classification setting and rely on static internal knowledge, which limits their ability to adapt and incorporate external evidence when interpreting up-to-date culturally coded expressions (Huang et al., 2023; Roy et al., 2023; Zhang et al., 2024). Second, such methods often suffer from opacity in reasoning and produce unverifiable evidence chains (Das et al., 2024).

To overcome these limitations, we propose AntiHateAgent, a novel LLM-driven agent framework for implicit hate speech detection. Unlike static classification approaches, AntiHateAgent operates through an active, three-phase reasoning chain: (1) **Search Query Generation**, where the input text is dynamically decomposed into precise queries targeting key entities and claims; (2) **External Knowledge Retrieval**, an iterative process that actively searches the web, filters for relevant evidence, and integrates findings until sufficient information is gathered for a reliable judgment; and (3) **Final Judgment**, where a verdict is produced alongside a verifiable rationale explicitly linked to the curated evidence. This framework achieves dynamic knowledge enhancement by retrieving up-to-date contextual information from the web, while its chain-of-thought architecture ensures decision-making transparency and auditability through evidence-based reasoning.

Our main contributions are:

- We propose AntiHateAgent, a novel agent framework that integrates dynamic external knowledge retrieval with chain-of-thought reasoning. By actively gathering and verifying contextual evidence, AntiHateAgent addresses the reliance on cultural nuance and background knowledge in implicit hate speech detection. The framework demonstrates superior performance and robustness across multiple LLMs compared to standard zero-shot prompting.
- We conduct comprehensive experiments on three implicit hate speech datasets. Results demonstrate that AntiHateAgent significantly enhances the recall of hateful samples by up to 43.6% and improves the overall F1 score by up to 22.3%. The approach proves especially effective for detecting newly implicit hate speech, achieving 89.7% recall, and obtains an 8.5% F1 gain in identifying emerging hate expressions.
- We provide an interpretable, evidence-based detection process. AntiHateAgent outputs not only a classification but also a transparent reasoning chain and verifiable external evidence, enhancing the accountability and trustworthiness of automated moderation systems.

2 Related Work

2.1 Definition of Hate Speech

We follow the common definition of implicit hate speech (ElSherief et al., 2021) and the UN's definition of hate speech followed by recent hate speech studies (Shen et al., 2025; Antypas and Camacho-Collados, 2023; Markov et al., 2023; Mathew et al., 2021; Toraman et al., 2022; Vidgen et al., 2021) as shown in Appendix A.

2.2 Explicit Hate Speech Detection

The evolution of hate speech detection has progressed from traditional machine learning (Burnap and Williams, 2015; Watanabe et al., 2018; Ombui et al., 2019) and deep learning (Gambäck and Sikdar, 2017; Pitsilis et al., 2018; Setyadi et al., 2018; Agarwal and Chowdary, 2021) to pre-trained language models (Caselli et al., 2021; Wang and Ding, 2019), with recent work exploring LLMs for both generation and detection of hate speech (Jin et al., 2024; Păiș, 2024; Shen et al., 2025; Zeng et al., 2024).

2.3 Implicit Hate Speech Detection

Detecting implicit hate speech is generally more challenging than identifying explicit hate speech (Ocampo et al., 2023b). To enable deeper exploration of this issue, prior research has developed fine-grained benchmark datasets (Sap et al., 2020; ElSherief et al., 2021; Kennedy et al., 2022; Ocampo et al., 2023b; Hartvigsen et al., 2022).

Some studies fine-tune pre-trained language models via contrastive learning (Kim et al., 2022; Ahn et al., 2024; Kim et al., 2024). In addition,

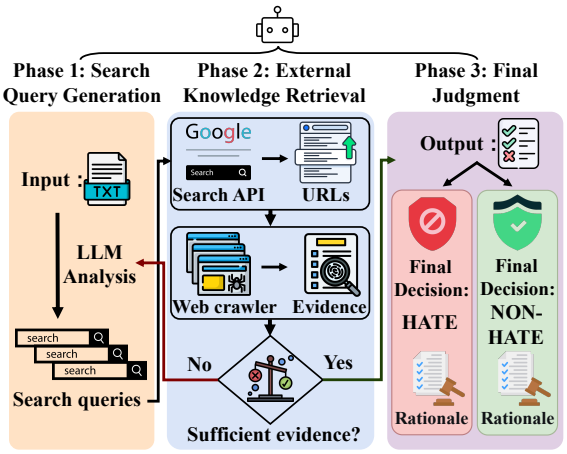

Figure 2: Overview of our AntiHateAgent framework.

other research focuses on building specialized foundation models Kim et al. (2023), architectural innovation (Ghosh et al., 2023; Yadav and Singh, 2024), and leveraging adversarial samples to improve robustness (Ocampo et al., 2023a).

Recent studies explore the zero-shot detection abilities of LLMs. Li et al. (2024) show that ChatGPT's alignment with human judgments depends heavily on prompt design and concept type. Huang et al. (2023) and Roy et al. (2023) find LLMs tend to be oversensitive in implicit hate detection, frequently flagging non-hateful content. Zhang et al. (2024) further analyze this issue, noting miscalibration between confidence scores and accuracy. To leverage both traditional models and LLMs, Damo et al. (2025) introduce an approach using BERT-based detectors to supply labels, probabilities, statements, and key tokens for LLM guidance. Nonetheless, in these approaches, the LLM still functions as a passive classifier, lacking the capacity to proactively explore and validate external knowledge, or to generate a verifiable chain of reasoning supported by external evidence.

3 Methodology

To address the reliance on cultural and background knowledge in implicit hate speech detection, we propose AntiHateAgent, an LLM-based agent framework. Its core design follows a chain-of-thought reasoning architecture, which achieves dynamic knowledge enhancement through external knowledge retrieval and evidence-based reasoning, while ensuring decision-making transparency by outputting verifiable rationales. As illustrated in Figure 2, the framework operates through three sequential phases: (1) Search Query Generation, (2) External Knowledge Retrieval, and (3) Final Judgment. The technical implementations of AntiHateAgent are provided in Appendix C and the actual prompts are in Appendix E.

Phase 1: Search Query Generation. It transforms the input text into a set of precise search queries to effectively retrieve relevant external knowledge in support of judgment. Directly using the original text often leads to irrelevant results due to its narrative form and implicit context. To address this, we instruct an LLM to identify key entities and claims requiring verification, and synthesize these elements into a structured list of concise search queries for the subsequent knowledge retrieval phase. For example, for the text *"These people are spreading the virus intentionally"* the LLM generates search queries like "intentional virus spread" and "COVID-19 blame rhetoric."

Phase 2: External Knowledge Retrieval. In this phase, the search queries are transformed into reliable decision-making evidence. It addresses two key limitations of retrieval-augmented methods for hate speech detection: the prevalence of noisy and irrelevant content in web search results, and the lack of a dynamic mechanism to determine when sufficient evidence has been gathered for a reliable judgment. Concretely, it relies on an active retrieval paradigm that interleaves evidence collection with continuous assessment, enabling iterative query refinement and transparent evidence curation.

(a) Relevance Filtering. Upon receiving search results (URLs and snippets) from a search API (Google), the LLM evaluates their relevance to the input text. Only webpages deemed highly relevant are crawled, thereby filtering out noise at the source and improving the quality of the evidence.

(b) Incremental Integration. For each crawled page, AntiHateAgent extracts text segments that support or refute the hateful nature of the input. It then integrates this new evidence into an accumulating judgment, updating its stance and logging its reasoning steps. This process builds a transparent, auditable reasoning chain.

(c) Closed-Loop Control. As shown in Figure 2, the framework centers on a critical decision point ("Sufficient evidence?") where AntiHateAgent itself assesses whether the accumulated evidence is sufficient for a reliable final judgment. This assessment is based on the relevance and strength of the evidence gathered in the incremental integration step. If the agent cannot extract sufficient evidence from the current webpages, the process reverts to

Dataset	Year	Source	Size	Subset in Exp.	Key Annotations
SBIC	Pre-2020	Compilation	44,671	1,000 (balanced)	Offensiveness (3-class), Group targeting (binary), Implied statement
GHC	2018	Gab	27,665	504 (balanced)	Human Degradation (binary), Calls for Violence (binary), Framing (implicit/explicit)
4chan	2025	4chan	2,000	1,408 (balanced)	Hate Identification (hate/nonhate), Hate Type Classification (implicit/explicit)

Table 1: Overview of hate speech datasets used in our experiments.

Phase 1 to generate more precise queries. This loop continues until evidence is deemed sufficient or a predefined limit on query reformulations is reached, ensuring robust termination even in ambiguous cases. This design enables AntiHateAgent to autonomously determine the stopping point of the retrieval process, making it highly adaptive to the varying complexity and online information availability of different input cases, rather than relying on a fixed retrieval depth.

This active retrieval phase offers two principal benefits. First, it autonomously optimizes the search strategy in real-time based on the quality and relevance of online information. Second, it ensures transparency and verifiability by maintaining a complete log of all retrieval, assessment, and reasoning steps.

Phase 3: Final Judgment. Upon confirming that sufficient evidence has been collected, AntiHateAgent consolidates the curated evidence to produce a final, evidence-grounded decision (HATE or NON-HATE) accompanied by a verifiable rationale. This rationale explicitly links the decision to the key supporting or refuting evidence extracted during the active retrieval process, ensuring transparency and verifiability for the final output.

4 Experiments

4.1 Datasets

We evaluate AntiHateAgent on three datasets (see Table 1): SBIC (Sap et al., 2020), GHC (Kennedy et al., 2022) and a newly constructed 4chan dataset.

SBIC. SBIC is a large-scale dataset containing 150,000 structured annotations of social media posts, which facilitates the modeling and evaluation of social bias implications. We classify the "offensiveness" category in SBIC as the "hate" class following prior work (Yang et al., 2023). To concentrate on social group biases, we filter the dataset to include only the group-targeted samples. This is done by selecting those marked as "0.0" in the `whoTarget` variable. From this subset, we randomly sample 500 hate and 500 non-hate instances, creating a balanced dataset of 1,000 samples.

GHC. GHC is a benchmark dataset expertly annotated for hate-based rhetoric, providing a theoretically grounded resource for hate speech detection and analysis. We employ the "hate" class (Kennedy et al., 2022), which is defined for a post if its majority-vote aggregated label is either Human Degradation or Calls for Violence in GHC. We use all 252 implicitly framed hate samples, and randomly sample 252 non-hate instances to maintain balance, resulting in a total of 504 samples.

4chan Dataset. To evaluate the performance of AntiHateAgent on emerging and evolving hate speech, we construct a new dataset called 4chan. The texts are collected from the /pol/ board of 4chan (4chan) in November 2025, a forum known for hosting a substantial volume of extremist content, making it a suitable source for capturing recently surfaced hate expressions (Hine et al., 2017; Papasavva et al., 2020). After preprocessing such as removing irrelevant metadata and duplicate content, we compile a total of 2,000 text samples. The annotation is performed by four annotators specialized in hate speech research. Each text is first independently annotated by two annotators to classify it as "hate" or "non-hate." Samples labeled as hate are further categorized as either "explicit hate" or "implicit hate," with implicit hate referring to expressions that convey hate through implication, irony, or insinuation. Disagreements between annotators are resolved through discussion to reach a consensus. Inter-annotator agreement is substantial, with an average Cohen's Kappa coefficient (Cohen, 1960) of 0.756. The final dataset consists of 704 hate samples and 1,296 non-hate samples. Among the hate samples, 78 are unanimously annotated as implicit hate by both annotators, with the remainder classified as explicit hate. We construct a balanced subset for the 4chan dataset by including all 704 hate samples and pairing them with 704 randomly chosen non-hate samples, leading to a final set of

Dataset	LLM	Method	Hate Samples			Non-Hate Samples			Overall			
			P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	ΔF1(%)
SBIC	DeepSeek	Baseline	98.0	39.0	55.8	61.9	99.2	76.2	80.0	69.1	66.0	-
		AntiHateAgent	84.3	61.0	70.8	69.4	88.6	77.9	76.8	74.8	74.3	(+8.3)
	Claude	Baseline	95.5	42.2	58.5	63.0	98.0	76.7	79.2	70.1	67.6	-
		AntiHateAgent	89.5	46.2	60.9	63.8	94.6	76.2	76.7	70.4	68.6	(+1.0)
	Gemini	Baseline	91.0	24.2	38.3	56.4	97.6	71.4	73.6	61.0	54.9	-
		AntiHateAgent	83.9	67.8	75.0	73.0	87.0	79.4	78.4	77.4	**77.2**	(+22.3)
	GPT	Baseline	97.0	45.0	61.5	64.2	98.6	77.8	80.6	71.8	69.6	-
		AntiHateAgent	92.5	39.6	55.5	61.6	96.8	75.3	77.1	68.2	65.4	(-4.2)
GHC	DeepSeek	Baseline	94.2	51.6	66.7	66.7	96.8	79.0	80.4	74.2	72.8	-
		AntiHateAgent	79.2	80.2	79.7	79.9	79.0	79.4	79.6	79.6	**79.6**	(+6.8)
	Claude	Baseline	93.6	40.5	56.5	62.0	97.2	75.7	77.8	68.8	66.1	-
		AntiHateAgent	82.7	58.7	68.7	68.0	87.7	76.6	75.3	73.2	72.6	(+6.5)
	Gemini	Baseline	92.8	56.3	70.1	68.7	95.6	79.9	80.7	76.0	75.0	-
		AntiHateAgent	74.6	84.1	79.1	81.8	71.4	76.3	78.2	77.8	77.7	(+2.7)
	GPT	Baseline	94.5	54.4	69.0	68.0	96.8	79.9	81.2	75.6	74.4	-
		AntiHateAgent	86.5	61.1	71.6	69.9	90.5	78.9	78.2	75.8	75.3	(+0.9)
4chan	DeepSeek	Baseline	89.7	74.1	81.2	78.0	91.5	84.2	83.8	82.8	82.7	-
		AntiHateAgent	86.3	98.0	91.8	97.7	84.4	90.5	92.0	91.2	**91.2**	(+8.5)
	Claude	Baseline	88.5	72.4	79.7	76.7	90.6	83.1	82.6	81.5	81.4	-
	Gemini	Baseline	85.8	87.9	86.9	87.6	85.5	86.6	86.7	86.7	86.7	-
	GPT	Baseline	85.3	76.4	80.6	78.6	86.8	82.5	81.9	81.6	81.6	-

Table 2: Performance comparison of AntiHateAgent and Baseline across different LLM backbones on SBIC, GHC, and 4chan datasets. P: Precision, R: Recall.

1,408 samples for experiments.

4.2 Experimental Settings

Baseline. To establish a comparative baseline against AntiHateAgent, we develop a zero-shot hate speech detection method based on direct prompting of LLMs. Following prior work (Sachdeva et al., 2022), the primary prompt is: "Does this comment contain hate speech?", accompanied by instructions that require the model to adhere to a specific output format and provide reasoning for its decision. Technical implementation details, including the complete prompt template, are provided in Appendix D.

LLM and Metrics. We select the most popular LLMs for AntiHateAgent and the baseline, specifically the endpoints "deepseek-chat," "claude-3-haiku-20240307," "gemini-2.5-flash," and "gpt-4.1-nano." We employ the default temperature settings for each LLM. All results are from a single evaluation run due to API cost constraints. We follow the common practice to use three key metrics: precision, recall, and F1 score (Schütze et al., 2008).

4.3 Main Results

As shown in Table 2, AntiHateAgent yields positive **Δ**F1 in 8 of 9 cases, indicating its general superiority in detecting implicit and emerging hate speech. For instance, on SBIC, the F1 score for the Gemini model improves by 22.3%. The framework consistently enhances performance across DeepSeek, Claude, and Gemini, demonstrating the robustness across backbone LLMs. In contrast, the baseline using GPT surpasses AntiHateAgent on SBIC. Table 2 shows that the baseline using GPT performs strongly on SBIC compared to baselines using other LLMs, reflecting its robust holistic reasoning and effective recognizing of implicit hate contexts. When external knowledge is less relevant in some cases, AntiHateAgent's step-by-step reasoning may interfere with this process, causing performance drops. Thus, while the framework benefits most LLMs, it can be detrimental for those already strong in holistic comprehension. Nonetheless, AntiHateAgent remains effective and demonstrates competitive performance overall.

The main performance advantage of AntiHateAgent derives from its substantially higher recall for hate samples compared to baselines. As shown in Table 2, AntiHateAgent improves recall across eight of nine combinations, with a notable increase from 24.2% to 67.8% in the SBIC/Gemini. This enhancement enables AntiHateAgent to effectively reduce the false negative rate in hate speech detection and constitutes the primary factor contributing to the improvement in overall F1 score.

LLM	Method	Recall(%)
DeepSeek	AntiHateAgent	**89.7**
	Baseline	48.7
Claude	Baseline	38.5
Gemini	Baseline	70.5
GPT	Baseline	51.3

Table 3: Comparison of AntiHateAgent using DeepSeek and Baseline on 78 implicit hate samples of 4chan.

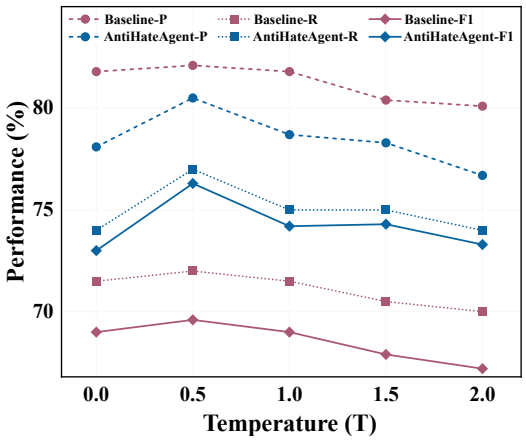

Figure 3: Performance comparison of AntiHateAgent and Baseline under different temperature settings of DeepSeek.

AntiHateAgent's performance gains are not symmetrically distributed across classes. While the F1 score on hate samples improves substantially due to higher recall, gains on non-hate samples are limited and slightly negative in some cases (e.g., SBIC/Claude). This asymmetry reflects a trade-off, where heightened sensitivity to hate-related cues effectively reduces false negatives for hate samples but also increases false positives for borderline non-hate cases. In addition, AntiHateAgent demonstrates superiority in detecting newly emerging hate expressions, evidenced by its high F1 score on our constructed up-to-date 4chan dataset. Specifically, it outperforms the baseline by 8.5% F1 when using DeepSeek. In particular, AntiHateAgent effectively detects emerging implicit hate speech. For example, *"We got Luigi Mangione fine ahh and now we have the French chad duo. Are women right about fucking criminals because they are chads? Will I become a chad if I do crime?"* contains implicit hate that requires knowledge of an event involving Luigi Mangione in December 2024. Baselines using pre-2025 LLMs lack the context for accurate classification, while AntiHateAgent retrieves relevant information to classify correctly. As shown in Table 3, on a subset of 78 implicit hate samples from the 4chan dataset, AntiHateAgent with DeepSeek achieves 89.7% recall, substantially surpassing baselines using DeepSeek and the other LLMs. These results underscore the framework's capability to capture subtle hate expressions.

4.4 Ablation Studies

To evaluate the effectiveness of core components of AntiHateAgent, we conduct ablation studies on a balanced test subset of 200 samples from SBIC and GHC datasets. We adopt DeepSeek as the backbone considering its superior performance.

LLM Temperature. To investigate the impact of the temperature parameter of the backbone LLM, we compare the performance of AntiHateAgent and the baseline on the DeepSeek across different temperature settings ($T \in 0, 0.5, 1.0, 1.5, 2.0$), as shown in Figure 3.

The main findings are as follows. First, across all settings, AntiHateAgent consistently and significantly outperforms the baseline in terms of recall and F1 score, though with slightly lower precision. This aligns with the analysis in Section 4.3, indicating that the framework prioritizes the detection of subtle hate at the cost of precision. Second, performance exhibits moderate sensitivity to the temperature: AntiHateAgent achieves its peak F1 score at $T = 0.5$, while showing a decline under completely deterministic or highly stochastic conditions. This suggests that a moderate randomness is beneficial for the detection. Notably, AntiHateAgent demonstrates strong robustness to temperature variation. Specifically, in the lower temperature range ($T < 0.5$), the F1 score of AntiHateAgent rises more sharply than the baseline, indicating its ability to better leverage moderate randomness for performance gains. Beyond the optimal temperature, the performance degradation of AntiHateAgent is also considerably slower. These show that AntiHateAgent not only achieves superior performance but also exhibits stronger robustness to variations in the temperature hyperparameter of the backbone LLM.

External Knowledge Retrieval. To evaluate the impact of external knowledge retrieval in AntiHateAgent, we compare three configurations: AntiHateAgent, AntiHateAgent with external web retrieval completely removed (AntiHateAgent w/o knowledge), and AntiHateAgent where the Google Search API for obtaining URLs is replaced by URLs generated by the backbone LLM (Anti-

Dataset	Method	Hate Samples			Non-hate Samples			Overall		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
SBIC	Baseline	100.0	43.0	60.1	63.7	100.0	77.8	81.8	71.5	69.0
	AntiHateAgent w/o knowledge	95.7	44.0	60.3	63.6	98.0	77.2	79.6	71.0	68.7
	AntiHateAgent w/o Google Search API	87.7	50.0	63.7	65.0	93.0	76.5	76.4	71.5	70.1
	AntiHateAgent w/o instructional details	84.5	60.0	70.2	69.0	89.0	77.7	76.7	74.5	74.0
	AntiHateAgent	89.1	57.0	69.5	68.4	93.0	78.8	78.7	75.0	**74.2**
GHC	Baseline	93.2	55.0	69.2	68.1	96.0	79.7	80.7	75.5	74.4
	AntiHateAgent w/o knowledge	85.5	71.0	77.6	75.2	88.0	81.1	80.4	79.5	79.4
	AntiHateAgent w/o Google Search API	85.4	70.0	76.9	74.6	88.0	80.7	80.0	79.0	78.8
	AntiHateAgent w/o instructional details	79.2	80.0	79.6	79.8	79.0	79.4	79.5	79.5	79.5
	AntiHateAgent	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	**80.0**

Table 4: Ablation study of AntiHateAgent components: the full framework versus variants without external knowledge retrieval (w/o knowledge), without the Google Search API which using backbone LLM-generated URLs instead (w/o Google Search API), and without instructional details for non-core processes (w/o instructional details). DeepSeek is used as the backbone LLM for all experiments.

HateAgent w/o Google Search API). As shown in Table 4, experimental results indicate that external knowledge retrieval significantly affects model performance.

On SBIC and GHC, AntiHateAgent achieves the best overall performance across all settings, demonstrating the effectiveness of external knowledge retrieval in implicit hate speech detection. Specifically, if external knowledge retrieval is entirely removed, AntiHateAgent w/o knowledge suffers a drop in F1 score from 74.2% to 68.7% on SBIC, indicating that external knowledge retrieval is crucial. When the external knowledge retrieval is retained, but the URLs obtained from a search API are replaced with URLs generated by the backbone LLM, AntiHateAgent w/o Google Search API still declines to 70.1%, mainly due to LLM hallucination. LLM tends to generate non-existent URLs, which compromises retrieval quality and leads to inferior performance. Notably, AntiHateAgent's superiority is evidenced by its high recall on hateful samples. As shown in Table 4, it achieves the highest hate recall on both SBIC and GHC compared with the other two ablated settings. This suggests that high-quality external knowledge is essential for recognizing implicit hate, as it provides the cultural and contextual cues needed to interpret subtle expressions often missed without it.

In sum, external knowledge retrieval in AntiHateAgent strengthens contextual awareness and sensitivity to implicit hate, and demonstrates superior hate detection performance across different datasets.

Prompt Engineering. The impact of prompt engineering is evaluated by comparing the full prompt (AntiHateAgent) with a simplified version (AntiHateAgent w/o instructional details), which eliminates instructional details for non-core processes, such as "Always anchor analysis in the original text's full context." As shown in Table 4, AntiHateAgent w/o instructional details achieves F1 scores nearly identical to that of AntiHateAgent on SBIC and GHC. This indicates that the effectiveness of the framework derives from its core processes and remains robust to modifications in auxiliary reasoning instructions.

5 Discussion

5.1 Case Study

In the following, we provide a case study to demonstrate how AntiHateAgent proactively retrieves and integrates external evidence to enhance the credibility and persuasiveness of its judgment, as shown in Figure 4. This case text, sourced from the GHC dataset and labeled as "hate," is shown below:

> "*California, if you keep voting Democrats in office, you will be ruled by ILLEGAL ALIENS soon.*"

The baseline method using DeepSeek misclassifies the text as "non-hate," arguing that "illegal aliens" though politically charged, targeted a political party rather than a protected group. This interpretation ignores the term's dehumanizing implications, leading to the mistake.

In contrast, AntiHateAgent relies not only on the internal knowledge of its backbone LLM but also actively constructs queries around key expressions (e.g., "illegal aliens") to retrieve external informa-

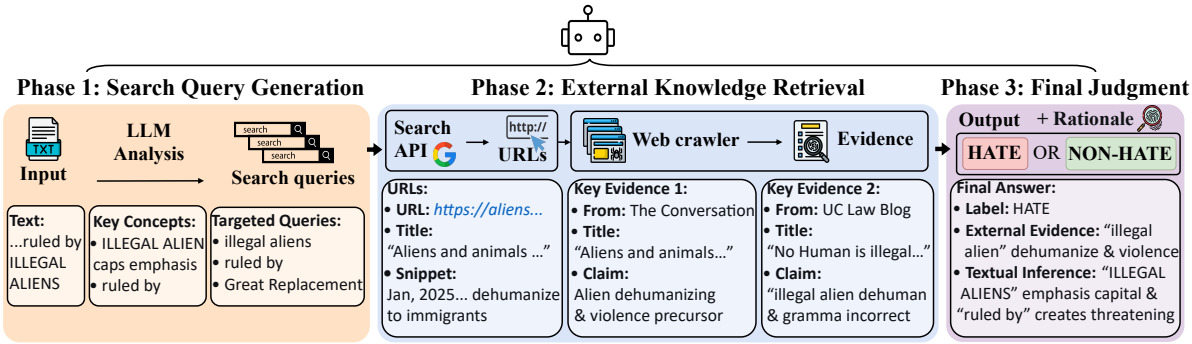

Figure 4: A case study demonstrating how AntiHateAgent proactively retrieves and integrates external evidence.

tion from the web. The retrieved results indicate that multiple authoritative sources identify the term "illegal aliens" as a common dehumanizing rhetoric associated with hate speech and incitement to violence. By integrating this external evidence with intrinsic textual cues, such as capitalization for emphasis and threat-based narratives like "ruled by," AntiHateAgent concludes that the text is not a neutral political statement but rather a dehumanizing portrayal of undocumented immigrants aligned with hate speech.

Compared to the baseline, AntiHateAgent demonstrates a clear advantage through its transparent reasoning chain of internal analysis, external verification, and evidence integration. This process not only improves detection performance but also enhances the interpretability and verifiability of the judgment.

5.2 Failure Analysis

We examine the primary sources and underlying causes of misclassification in false negative and false positive cases. For clarity, positive denotes hate, whereas negative denotes non-hate.

False Negative Analysis. False negatives occur mainly due to semantic-retrieval misalignment and external knowledge scarcity.

(a) Semantic Retrieval Misalignment. AntiHateAgent can detect implicit hateful intent but fails to translate it into effective search queries, leading to superficial fact-checking that misses the core hateful meaning. For instance, *"Even atheist like Richard Dawkins think Islam is bad"* (hate), searches of this text focus on verifying Dawkins' views instead of the rhetorical use of authority to promote bias, causing misclassification. The baseline also misclassifies such cases due to its inability to recognize implicit hate.

(b) External Knowledge Scarcity. Hate speech from fringe communities often lacks public documentation. Web searches with limited results may miss relevant sources, creating knowledge gaps and causing misclassification. The baseline makes similar errors, primarily due to a lack of up-to-date knowledge.

False Positive Analysis. False positives derive from AntiHateAgent's over-sensitivity to controversial keywords, amplified by external knowledge.

(a) Keyword Sensitivity. AntiHateAgent is overly sensitive to terms common in hateful contexts (e.g., "Gab," "Hitler") and lacks contextual discrimination. Retrieved external information reinforces these associations. For example, *"Happy New Year #GabFam"* (non-hate) might be misclassified simply because it references "Gab," a platform often associated with hateful content. The baseline typically avoids such errors as it lacks the external knowledge that amplifies this sensitivity.

6 Conclusion

We propose AntiHateAgent, an LLM-based agent framework to detect implicit hate speech via dynamic knowledge enhancement and transparent decision-making through chain-of-thought reasoning. Experiments on three datasets, including our newly constructed 4chan emerging hate speech dataset, show that AntiHateAgent significantly improves recall on hate instances and overall F1 score. Ablation studies confirm the critical role of external knowledge retrieval and the robustness of the framework. AntiHateAgent also provides transparent justification of the decision process through proactive and verifiable reasoning, thereby contributing to more reliable content moderation systems.

Limitations

Despite its promising performance, AntiHateAgent has limitations. First, its multi-round retrieval and reasoning framework introduces increased computational overhead and latency, which limits its applicability in real-time scenarios. Additionally, although the current experimental evaluation covers multiple datasets, it focuses on English text, leaving the framework's generalizability to other languages unexplored.

Ethical Considerations

We acknowledge the ethical considerations in developing and evaluating automated hate speech detection systems like AntiHateAgent, and have taken steps to mitigate potential risks. While AntiHateAgent advances the detection of implicit hate speech, it may also increase misclassifications of borderline content, which could result in unfair moderation decisions. Additionally, our framework's dependence on web-retrieved knowledge carries the risk of amplifying social biases present in online data. Therefore, real-world deployment should be accompanied by rigorous fairness audits across diverse communities and maintained under continuous human oversight. We use established hate speech detection benchmark datasets (SBIC, GHC), which are publicly released under research licenses and have been widely used in prior studies. SBIC and GHC datasets are both licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). We rely on the original creators' compliance in obtaining informed consent from data subjects. The data is solely used for academic research to address hate speech more effectively. This study utilizes a newly constructed dataset sourced from 4chan. The entire dataset is preprocessed to remove all personally identifiable information, such as names or unique identifiers, only the textual content of the posts is retained and processed in strict compliance with the platforms' terms of use. From this anonymized corpus, samples are manually annotated as hate speech by the trained research authors, thereby preventing exposure of external annotators to harmful material. Due to the offensive nature of the content, access to the dataset is restricted to prevent misuse. We will manually review all applicant details before granting access.

References

4chan. Politically incorrect (/pol/). https://boards.4chan.org/pol/. Accessed: November 2025.

Shivang Agarwal and C Ravindranath Chowdary. 2021. Combating hate speech using an adaptive ensemble learning model with a case study on covid-19. *Expert Systems with Applications*, 185:115632.

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10444–10455.

Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. 2025. Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13:273.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6193–6202.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20:1–22.

Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2025. "detectors lead, llms follow": Integrating llms and traditional models on implicit hate speech detection to generate faithful and plausible explanations. *Data & Knowledge Engineering*, page 102535.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2024. Evaluating chatgpt against functionality tests for hate speech detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6370–6380.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, pages 512–515.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51:1–30.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. CoSyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6159–6173.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 653–670.

Google. Custom search json api.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 92–101.

Jeffrey W Howard. 2019. Free speech and hate speech. *Annual Review of Political Science*, 22:93–109.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Yiping Jin, Leo Wanner, and Alexander Shvets. 2024. GPT-HateCheck: Can LLMs write better functional tests for hate speech detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7867–7885.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and 1 others. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16177–16188.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679.

Youngwook Kim, Shinwoo Park, Youngsoo Namgoong, and Yo-Sub Han. 2023. Conprompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10964–10980.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18:1–36.

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

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media*, 22:205–224.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023a. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In *ACL 2023-61st Annual Meeting of the Association for Computational Linguistics*, pages 2758–2772. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023b. An in-depth analysis of implicit and subtle hate speech messages. In *EACL 2023-17th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2023, pages 1997–2013. Association for Computational Linguistics.

Edward Ombui, Lawrence Muchemi, and Peter Wagacha. 2019. Hate speech detection in code-switched text messages. In *2019 3rd international symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pages 1–6. IEEE.

Vasile Păiș. 2024. Racai at climateactivism 2024: Improving detection of hate speech by extending llm predictions with handcrafted features. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 67–72.

Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 885–894.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Nabiila Adani Setyadi, Muhammad Nasrun, and Casi Setianingsih. 2018. Text analysis for hate speech detection using backpropagation neural network. In *2018 international conference on control, electronics, renewable energy and communications (ICCEREC)*, pages 159–165. IEEE.

Xinyue Shen, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2025. HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns. In *USENIX Security Symposium (USENIX Security)*. USENIX.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online*, pages 78–84.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.

Jeremy Waldron. 2012. *The harm in hate speech*. Harvard university press.

Bin Wang and Haiyan Ding. 2019. Ynu nlp at semeval-2019 task 5: Attention and capsule ensemble for identifying hate speech. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 529–534.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Ashok Yadav and Vrijendra Singh. 2024. Hatefusion: Harnessing attention-based techniques for enhanced filtering and detection of implicit hate speech. *IEEE Transactions on Computational Social Systems*.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505.

Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. 2025. Sheep's skin, wolf's deeds: Are llms ready for metaphorical implicit hate speech? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16657–16677.

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

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10:925–945.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Hate Speech Definition

Definition of Implicit Hate Speech: "Implicit hate speech is a subclass of hate speech defined by the use of coded or indirect language such as sarcasm, metaphor and circumlocution to disparage a protected group or individual, or to convey prejudicial and harmful views about them."
Definition of Hate Speech: "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor."

B Dataset Annotation

We provide the annotators with definitions of hate speech and implicit hate speech in in Appendix A, and require them to first determine whether each text is hate speech. If so, they are then asked to judge whether it constitute implicit hate speech. All annotators are researchers involved in this study. Specifically, they are four women between the ages of 20 and 30 who specialize in hate speech detection.

C AntiHateAgent Technical Implementation

The implementation of AntiHateAgent relies on a carefully designed interaction between structured prompts and programmatic components to achieve its three-stage analysis pipeline. The initial prompt establishes strict formatting requirements through the `format_prompt_dict` specification, which enforces a consistent thought-action-observation cycle throughout the analysis process. This cyclic structure is implemented in the program, where each iteration follows the prescribed format of generating thoughts, executing actions, and processing observations.

The retrieval stage's iterative evidence evaluation is implemented through the observation processing loop in program. After each URL crawl (executed via the SearchWeb action), the system stores raw content and triggers the `summarize_observation()` method to extract relevant evidentiary segments. This method employs a sliding window approach to handle large documents, ensuring comprehensive analysis while maintaining context. The prompt's requirement for incremental judgment updates is implemented through the history tracking mechanism, where each step's evidence and analysis are preserved in `history_steps`.

The framework's final classification decision follows a rigorous set of determination principles enforced through carefully designed prompt constraints on the LLM. First, the model must conduct a comprehensive analysis of the original text's complete contextual framework, including lexical choices, semantic framing, tonal qualities, and underlying intent. Second, the LLM is required to systematically integrate and evaluate all collected external evidence, assessing its collective consistency, potential contradictions, and degree of relevance to the original text's content. Upon meeting all specified determination criteria, AntiHateAgent outputs its classification result (hate or non-hate), accompanied by the complete evidence chain and a detailed analytical justification explaining how the decision was derived based on the original text context and the synthesis of all evidence.

The implementation includes several safeguards: the `reconfirm_text` mechanism ensures the original input remains unchanged throughout analysis, while the evidence chain preservation is implemented through the `history_steps` tracking and research log system. The program's action validation system (via `parse_action_input`) prevents deviations from the prescribed analytical process, ensuring all classification decisions follow the framework's methodological requirements.

Our code is primarily based on the PyTorch (>=2.0.0) framework, relies on NumPy (>=1.21.1) for fundamental computations, and employs toolkits such as scikit-learn, Pandas, OpenAI, and Anthropic to conduct the experiments. The complete dependency list will be provided later along with the code repository.

D Baseline Technical Implementation

The implementation of the baseline method consists of three key components:

- **Prompt Design**: In the prompt design, the model is first asked to determine whether the input text constitutes hate speech, followed by the target text to be evaluated. The model is required to respond strictly in one of two structured formats: "hate speech: [reason]" or "not hate speech: [reason]." This enforced output structure ensures consistency and facilitates reliable parsing of the model's responses.

- **Model Configuration**: We employ the default temperature settings for each LLM and set the maximum output tokens to 200, in order to balance creativity and consistency in the model's responses. All classification decisions are made based solely on the model's parametric knowledge, without access to any external data sources.

- **Error Handling**: The implementation includes automatic retry mechanisms (up to 3 attempts) for failed API calls and strict output validation to ensure all responses conform to the required format.

The complete prompt template used for the baseline LLM zero-shot classification method is as follows:

```
Does this comment contain hate speech?

Your response must follow this exact format:

hate speech: [reason], if the comment
    contains hate speech.

not hate speech: [reason], if the comment
    does not contain hate speech.

Replace [reason] with a brief explanation
    for your choice.

Comment:
{text}
```

E AntiHateAgent Prompt

```
   You are a helpful research assistant
   specialized in detecting hate speech.
   You have access to TOOL PROMPT. Your
   task is to analyze a text from the
   dataset.txt and determine if it
   contains hate speech.

1. For the text, design search queries to
    find relevant information:
   - Each query MUST:
     * Use advanced search operators
     * Be relevant to the original text
     * Be specific and targeted
     * Be helpful for gathering evidence to
     support or refute the presence of hate
     speech
   - Generate at least 3 distinct queries

2. Execute Google searches for each query:
   - Carefully review the returned URLs (
     title and snippet)
   - Select most relevant URLs for crawling
   - For each selected URL:
     * Crawl the content
     * Extract key evidence relevant to the
     original text
     * Update your judgment incrementally
     * Preserve all crawled evidence in
     observation

3. After each URL crawl:
   - Reconfirm the original text content
   - Document current judgment and
     supporting evidence
   - Explain how new evidence affects your
     assessment

4. Make final decision:
   - HATE or NON-HATE only
   - Output the confidence level for your
     final judgment
   - Based on:
     * Full original text context
     * All accumulated evidence from crawled
      content

ANALYSIS METHODOLOGY & REQUIREMENTS
Required format for each step:
{format_prompt_dict}

Reference Usage:
   - Crawled content serves ONLY for
     evidence

Critical Requirements:
- Minimum one successful crawl with
    preserved evidence
- Maintain complete evidence chain across
    all steps
- Before final judgment, MUST reconfirm text
     from the dataset.txt
- Final judgment must:
  * MUST output either "HATE" or "NON-HATE"
    as final judgment
  * MUST respond in the exact format
    specified in "Required format for each
    step"

WARNING:
- Never use crawled content as direct
    classification basis
- Always anchor analysis in original text's
    full context
- Never make any changes to the original
    text or the dataset.txt file. The
    original text must remain completely
    identical to the content read from
    dataset.txt.

format_prompt_dict:
  - Thought: Current analysis and next steps
  - Action: Next action to take
  - Action Input: Input to action as valid
    JSON

TOOL PROMPT: Read File, Google Search tool,
    Crawl web tool (Simplified)
```

F Computational Resources

All experiments in this study were deployed and conducted on the Alibaba Cloud platform. The hardware environment utilized an Alibaba Cloud Elastic Compute Service (ECS) instance with the

specification ecs.e-c1m2.2xlarge, which provides 8 virtual central processing units (vCPUs) and 16 GiB of system memory.

The experimental design employs a multi-process parallel execution strategy to reduce the overall experimental duration. In the main experiment, multiple AntiHateAgent processes invoke LLM APIs in parallel to process tasks, with an average processing time of approximately 2.5 minutes per sample. Subsequent extension and ablation experiments are then scheduled and executed sequentially. Under ideal failure-free conditions, the cumulative single-process equivalent computing time amounts to approximately 320 hours. In practice, due to task failures and retries during execution, the actual computing time increases accordingly. Through multi-process parallel execution, the entire experimental cycle is typically completed within about two weeks.

Since both LLM APIs and Google Search APIs are commercial services, the costs are relatively high. The total expenditure for the entire process, from preliminary experiments to finalizing the results, amounts to approximately 1000 $.

G AI Assistant Usage Statement

We used AI assistants in this research: ChatGPT for polishing writing and assisting with simple data processing code. All core research ideas, experimental design, result analysis, and primary text were independently developed by the authors. AI tools served only as auxiliary aids, and the authors take full responsibility for all content.