DCAdapt: A Dual-Cognition Driven Framework for Adaptive Hate Speech Detection

Xi Zhong[1,2], Cheng Ouyang[1*], Yan Guo[1], Yuanhai Xue[1], Xinran Liu[1]
[1]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{zhongxi,ouyangcheng,guoy,xueyuanhai,liuxinran}@ict.ac.cn

***Abstract*—Hate speech detection faces significant challenges due to inconsistent criteria across datasets, rooted in inherent subjectivity and annotation bias. This paper argues that the core of this task lies in deep adaptation to specific annotator stances within particular scenarios rather than merely fitting data distributions across dataset. Inspired by the Dual-Process Theory in human cognitive science, we propose DCAdapt, a dual-cognition driven framework for adaptive hate speech detection that achieves high adaptability by simulating the synergy between intuitive association and logical calibration. The framework decomposes the detection task into two collaborative stage: decoupled dual-knowledge base construction and dual-pathway cognitive reasoning. In the construction stage, a mutual-indexing mechanism decouples objective knowledge from subjective precedents to achieve synergistic modeling of general knowledge and annotation criteria. In the reasoning stage, the framework emulates the decision sequence from affective association to rational judgment: an implicit pathway leverages a general knowledge base to uncover latent contextual clues, while an explicit pathway employs a specialized knowledge base to enforce semantic constraints and stance alignment based on precedents for informed adjudication. This dual-pathway coordination enables DCAdapt to guide Large Language Models (LLMs) to precisely identify complex decision boundaries and align with target-scenario annotation stances. Experimental results demonstrate that DCAdapt exhibits superior effectiveness, architectural flexibility, and cross-scenario transferability, particularly in detecting complex and implicit hate speech.**

***Index Terms*—Hate speech detection, Content moderation, Knowledge enhancement, Dual-process theory.**

I. Introduction

Warning: This paper contains examples of content that is offensive and may be upsetting.

With the rapid development of social media, online networks have become the primary battleground for public opinion in today's digital era. Hate speech, leveraging the strong connectivity of social platforms, spreads rapidly and exhibits significant viral propagation patterns [1], [2]. Consequently, investigating effective detection methods is crucial for purifying the online environment and maintaining cyberspace order. However, there currently exists no unified annotation criteria for hate speech, with classification criteria often drifting based on annotators' subjective cognitive factors such as cultural background, value orientation, and life experiences. A typical example is that the same statement may exhibit vastly different levels of offensiveness toward different demographic groups due to variations in societal sensitivity. Researchers have identified substantial discrepancies in annotation criteria across various hate speech detection datasets [3]–[5], with these differences rooted in annotators' cultural backgrounds, annotation objectives, annotation methodologies, label granularity, and data collection approaches. This leads to identical text being categorized into entirely different classes under varying benchmarks. Fasching et al. [6] evaluated the performance of mainstream LLMs in hate speech detection, revealing significant inconsistencies in classification results for identical content across models. Notably, these disparities are particularly pronounced when addressing content targeting different demographic groups. This inherent subjectivity and potential annotator bias highlight the task's high sensitivity to detection contexts, significantly increasing its complexity [3], [4], [7]–[11].

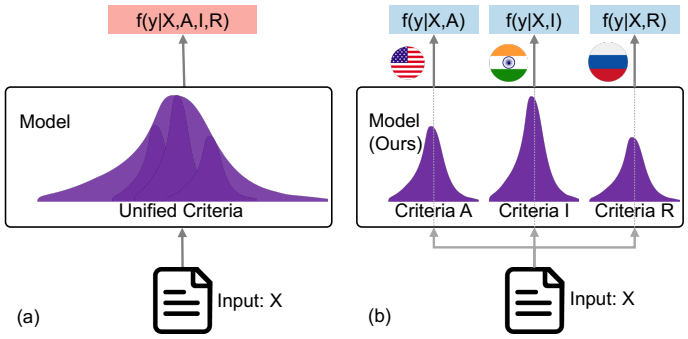

Fig. 1. Comparison between universal generalization and adaptive alignment paradigms. (a) Existing methods often focus on fitting a joint distribution to establish universal criteria. (b) Our proposed DCAdapt explicitly acknowledges subjectivity and achieves self-adaptation to diverse dataset criteria.

To address the aforementioned challenges, existing research primarily focuses on cross-dataset model generalization as an optimization objective to mitigate discrepancies in toxicity annotation standards across datasets. Methods based on pre-trained language models learn latent toxicity boundaries through fine-tuning [11]–[16] and utilize emotional signals or prompt engineering to enhance representation capabilities. The emergence of LLMs has driven the detection paradigm

*Cheng Ouyang is the corresponding author.
Code is available at https://github.com/xizhong/DCAdapt.

toward generative reasoning, with works like [17], [18] demonstrating their potential in hate speech detection. Furthermore, knowledge-enhanced approaches incorporate annotated data or fine-grained knowledge as prompts [19]–[22] to provide decision-making references for model judgment. Recent studies [4], [23] have highlighted the over-sensitivity issue in LLMs for hate speech detection, prompting the research community to explore solutions. PREDICT [3] propose multi-agent fitting of different annotation perspectives with ensemble learning voting to enhance generalization, while Yao et al. [24] improve decision accuracy in complex scenarios through personalized rules and high-quality example retrieval. However, these approaches still fail to fundamentally address the scenario adaptation challenges for diverse annotation criteria. However, these methods still fail to fundamentally solve the problem of scenario adaptation for diverse annotation standards. As illustrated in Fig 1, existing methods typically aim to design a unified model with cross-dataset generalization by fitting the joint distribution $f(y|X, A, I, R)$ to establish universal criteria. While this approach enhances out-of-distribution generalization performance, it implicitly balances different dataset standards and overlooks inherent differences in annotation contexts, making it difficult to precisely identify decision boundaries under specific cultural backgrounds or annotation objectives. Given that subjectivity and bias are inherent attributes of this task, this paper explicitly acknowledges these phenomena and proposes a detection framework capable of self-adapting to diverse dataset annotation standards.

To achieve precise adaptation to diverse annotation criteria, we propose DCAdapt, a dual-cognition driven framework for adaptive hate speech detection, which transforms human cognitive processes of implicit association and explicit reasoning into computable knowledge-enhanced pathways [25], [26]. The framework decomposes the task into two collaborative stages: decoupled dual-knowledge base construction and dual-pathway cognitive reasoning. The framework comprises two decoupled knowledge engines: a general knowledge base, which facilitates broad general knowledge activation for fast thinking through a dual-layered semantic network of entities and abstract groups, and a specific knowledge base, which integrates domain-specific discriminative criteria via a collection of annotated exemplars. Furthermore, this decoupled design not only enables the extraction of cross-scenario general knowledge but also establishes dynamic correlations between general knowledge and specific precedents through mutual-indexing mechanisms. During the reasoning stage, DCAdapt synergizes dual cognitive pathways by mimicking expert-level reasoning. It first activates the implicit cognition pathway to trigger multi-hop associations via semantic graphs, enabling intuitive semantic expansion of text clues. Subsequently, it switches to the explicit cognition pathway, utilizing the specific knowledge base as a stance anchor point to perform rigorous logical screening of associated clues through semantic consistency verification. This mechanism ensures that the information delivered to LLMs is both contextually enriched and ideologically aligned with the target scenario. Consequently, DCAdapt guides the model toward robust judgment at complex decision boundaries while maintaining superior cross-dataset adaptability.

Compared to existing retrieval-augmented generation (RAG) methods [19], [22] or knowledge reasoning (KGQA) methods [21], this paper successfully transforms originally implicit and fragmented contextual knowledge into explicit knowledge through dual-layer graph-based multi-hop association mining. This completion of background information bridges the cognitive gap between surface-level statements and underlying hate logic, making decision boundaries intuitive and identifiable with rich evidence support. When the framework can trace back to strongly explanatory evidence chains of evidence, the LLM can escape from context-based speculative judgments, fundamentally mitigating misjudgments and over-sensitivity caused by contextual ambiguity. Moreover, this decoupled design grants the framework strong cross-scenario adaptability: the general commonsense network can be continuously accumulated, while adapting to a new detection task only requires dynamically updating the specialized knowledge base, providing a new path for flexible and precise adaptive hate speech detection.

The contributions of this paper are as follows:

1) We propose DCAdapt, a dual-cognition driven framework for adaptive hate speech detection. This framework discards the conventional approach of fitting cross-data distributions, instead leveraging the scenario-specific variations in annotation criteria. By providing LLMs with stance-explicit and semantically complete context information, DCAdapt achieves precise adaptation to new dataset standards without retraining, effectively addressing the core challenge of model adaptation under inconsistent criteria.

2) We construct a decoupled dual-knowledge base architecture based on the mutual-indexing mechanism. The general knowledge base accumulates cross-dataset general semantic knowledge graphs, while the specialized knowledge base maintains toxic judgment precedents for the target domain. This design enables flexible switching of discriminative standards and independent evolution of general knowledge base.

3) We introduce dual-cognitive mechanisms to guide the framework's dynamic reasoning process. The fast thinking mode mode corresponds to an implicit cognitive pathway that performs rapid association mining within the general knowledge base to capture latent correlation clues. The slow thinking mode executes an explicit cognitive pathway that performs logical filtering based on the specialized knowledge base to guide deep reasoning. This mechanism emulates the decision sequence from intuitive association to rational judgment, ensuring consistent detection stances.

4) Extensive experiments across two representative hate speech datasets. Results validate DCAdapt's effectiveness, flexibility, and cross-scenario adaptability in handling complex standard drift scenarios.

II. Related Work

Current research on fine-tuning models for hate speech detection primarily focuses on the construction of fine-grained datasets and the optimization of training strategies. In the data dimension, researchers have developed specialized corpora targeting youth internet slang [16], child protection [27] and Chinese linguistic features [28], while exploring pseudo-label generation using high-performance models to assist manual annotation [29]. Regarding training strategies, HateBERT [12] and multi-source feature fusion models [30] have enhanced hate semantic capture through domain-adaptive pre-training, whereas SHAREDCON [14] introduced contrastive learning for implicit hate feature identification. Despite these advancements, task-adaptive frameworks [31] and reasoning-enhanced fine-tuning strategies [32] have significantly improved model generalization and robustness. However, the traditional fine-tuning paradigm possesses inherent limitations; due to its heavy reliance on static subjective annotations and prolonged model update cycles, it struggles to meet the demands for rapid standard iteration and agile response to dynamic safety guidelines in content moderation [20].

To overcome the temporal latency of fine-tuning methods, Retrieval-Augmented Generation has achieved a transition from parametric classification to a retrieval-and-compare judgmental paradigm by incorporating externally updated knowledge bases [19]. Researchers have utilized vector databases containing safety exemplars [20], multilingual corpora [22], or specific compliance policy documents [19] as augmented contexts, enhancing deployment flexibility and decision interpretability in zero-shot scenarios. Additionally, frameworks like ED2D [33] have introduced multi-agent debate mechanisms to strengthen judgmental evidence persuasiveness. Despite the flexibility advantages, vanilla retrieval models rely excessively on query-document semantic similarity. When faced with semantically ambiguous or implicit toxic content, retrieval lacking structural guidance often introduces irrelevant documents, leading to semantic drift and consequent misjudgment [21].

In contrast to retrieval-augmented approaches, graph-enhanced retrieval and Knowledge Graph Question Answering provide new pathways for handling complex hate reasoning by injecting structured semantic relationships into LLMs. While advanced frameworks like HippoRAG [34] and ToG [35] have emerged in general domains to suppress hallucinations and optimize retrieval efficiency, their application in hate speech detection remains in early stages. Zhao et al. [21] conducted representative exploration by constructing the first Meta-Toxic knowledge graph to explicitly represent toxic concept hierarchies through structured triplets. However, compressing complex hate judgments into discrete triplets inevitably leads to contextual association loss and subtle semantic detail degradation, thereby inhibiting the model's ability to recognize implicit hate and cross-cultural expressions. HA-GCEN [36] and StereoGraph [37] have attempted hypergraph structures and domain ontologies for implicit bias capture, but balancing graph structural advantages with textual context richness remains a critical challenge facing the academic community.

III. DCAdapt Framework

We propose DCAdapt, a dual-cognition driven framework for adaptive hate speech detection (Fig 2). The DCAdapt framework comprises five core modules that span the entire pipeline from decoupled dual-knowledge base construction to dual-process cognitive reasoning. During the construct stage, the framework includes the general knowledge base construction, specialized knowledge base construction, mutual-indexing construction. The reasoning stage consists of two sequential pathways: implicit cognitive pathway reasoning and explicit cognitive pathway reasoning.

A. General Knowledge Base Construction

The fine-grained knowledge graph layer of the general knowledge base is composed of triplets consisting of head entities, relations, and tail entities. We leverage LLMs to extract these triplets from raw texts, thereby constructing a fine-grained general knowledge base. Following conventional knowledge graph practices, we adopt entity types including People, Location, Organization, Event, and appropriately annotate each entity with its corresponding type. To enhance semantic coherence and resolve ambiguities, we generate detailed descriptive information for each entity based on context clues, creating unique identifiers through joint representations of entity names and semantic descriptions. Relations denote factual connections between head and tail entities, while relation descriptions provide in-depth interpretations of these connections. To address knowledge redundancy and achieve standardization, we perform semantic matching across global entities and relations using semantic similarity metrics, merging semantically equivalent or closely related entities and relations.

Due to the potential issues of knowledge fragmentation, obsolescence, or incompleteness in fine-grained knowledge bases constructed from limited samples, mismatches may arise during the reasoning stage. Consequently, we introduce the concept of target groups from the toxicity detection domain, which represents the groups or individuals targeted by hate speech [38]. In this section, following the definitions provided by [39], we extend the group concept to non-toxic data. This definition outlines 13 social groups, such as Asians, Black people, Muslims, and Women. Nevertheless, these groups primarily focus on Person-type entities and fail to encompass other entity types. Given that Location, Organization, and Event entities exhibit strong characteristics of individual independence and are difficult to classify into finer subcategories, we directly incorporate these four entity types as distinct groups. This results in a total of 17 groups, comprising the 13 social groups and the 4 fundamental entity types, which serve as the basis for the coarse-grained knowledge base.

By associating fine-grained entities with their respective groups, we establish a logical bridge between multi-granular knowledge. This coarse-grained knowledge, characterized by

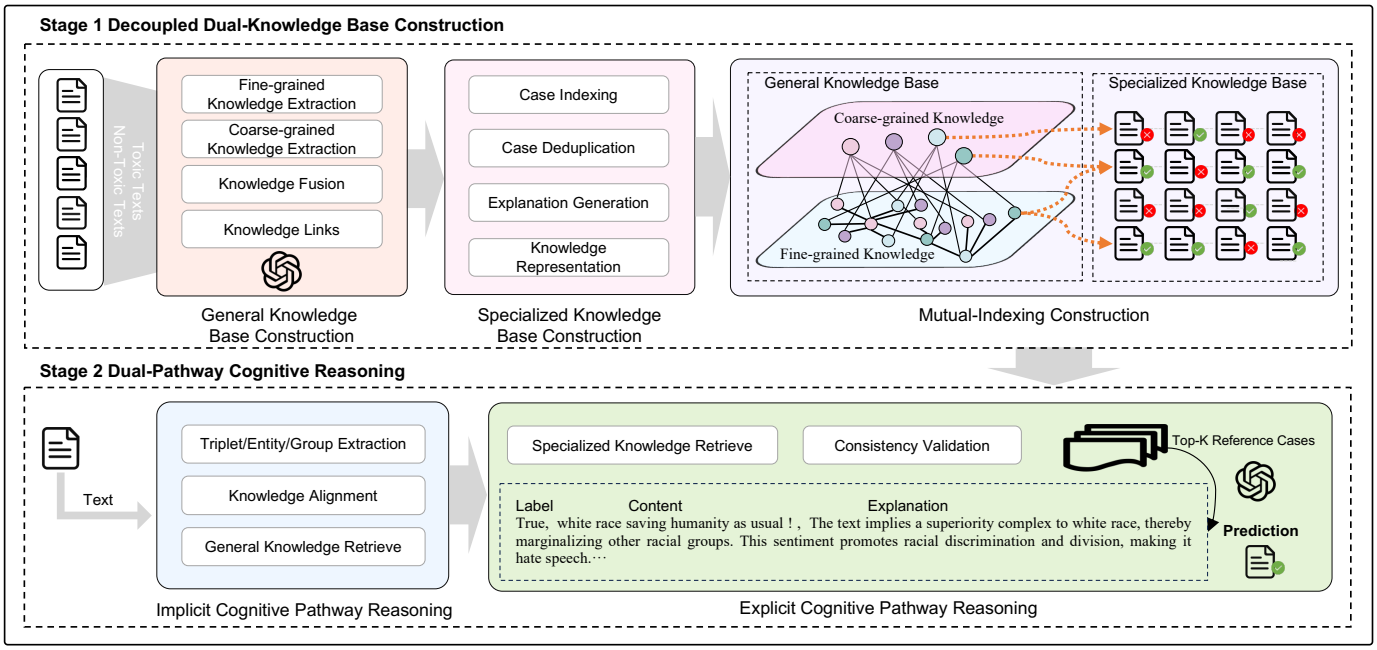

Fig. 2. The architecture of the DCAdapt framework. It features a decoupled dual-knowledge base construction stage linked by a mutual-indexing mechanism, followed by a dual-pathway cognitive reasoning stage.

a high level of abstraction, provides a broader universal semantic context, effectively compensating for the limitations of fine-grained clues in complex or weakly associated contexts, thereby ensuring the robustness of the knowledge-enhanced process. The text sources for constructing this general knowledge graph are derived from the training set of specific datasets. Since the general knowledge base does not involve toxicity supervision signals, it can be further supplemented and expanded using other real-world corpora or existing knowledge bases.

B. Specialized Knowledge Base Construction

The specialized knowledge base is designed to integrate domain-specific labeled datasets by mining the underlying discriminative logic embedded within samples, thereby addressing the absence of annotated signals in general-purpose knowledge bases. The annotated samples provide precise criteria and decision boundaries for toxicity classification. To further enhance the interpretability and logical coherence of these cases, rationales for toxicity judgments are generated for each labeled sample using LLMs. The original text, annotation labels, and judgment rationales are then structurally reorganized to form reference cases that are context-rich and annotated-signal-dense. The specialized knowledge base not only offers authoritative annotation references for specific datasets but also achieves explicit representation of domain expertise through rationale injection, enabling models to reach consistent judgments at complex decision boundaries.

C. Mutual-Indexing Construction

Drawing inspiration from the decoupling philosophy of KAG [40], the mutual-indexing mechanism establishes a loosely coupled association between the commonsense nodes in the general knowledge base and the toxic precedents in the specialized knowledge base. This architecture allows the general knowledge base to continuously accumulate and expand semantic knowledge across diverse datasets while ensuring that the specialized knowledge base maintains specified judgmental criteria. Through the mutual-indexing mechanism, the framework dynamically maps abstract clues from the raw text to the most relevant reference precedents.

Specifically, the general knowledge base functions as a dual-layer semantic network, where fine-grained triplets and coarse-grained group knowledge provide objective facts that without direct toxicity annotation signals. In contrast, the specialized knowledge base, composed of annotated samples, serves as the primary source of judgmental criteria and supervisory signals. To bridge this gap, we employ the mutual-indexing mechanism to build bidirectional links between the entities and relations in the general knowledge base and the annotated samples in the specialized knowledge base. By implementing bidirectional associations between general knowledge base entities/relationships and specialized knowledge base annotated samples, the framework ensures that whether retrieval hits specific entities or abstract patterns, it can trace back to explanatory case precedents through the mutual-indexing mechanism. This loosely coupled connection grants DCAdapt significant flexibility: the framework can seamlessly switch between different dataset standards or respond to dynamic drifts in annotation criteria within the specialized knowledge base.

D. Implicit Cognitive Pathway Reasoning

During the implicit cognitive pathway reasoning stage, DCAdapt simulate human intuitive responses when encountering information, aiming to achieve intuition-driven clue expansion based on a general knowledge base through rapid semantic association. For text under detection, the framework first performs multi-granularity clue extraction: at the fine-grained level, it utilizes LLMs to extract fact triples and independent entities, generating context descriptive information to align with knowledge base content; at the coarse-grained level, the framework aggregates all mentioned entities in the text and maps them to predefined group according to target object, establishing multi-level initial semantic anchors.

Subsequently, the framework proceeds to the cross-graph knowledge awakening stage. It employs pre-trained language models to semantically encode the extracted clues and their descriptions, integrating them into the general knowledge base based on consistency principles. Centering on activated knowledge nodes, the framework executes a multi-hop neighborhood search to recall closely associated general knowledge. This process bypasses complex logical deduction, relying instead on semantic diffusion mechanisms to link fragmented and implicit adversarial intent in text with macro-level contextual knowledge. Through this intuition-based context completion, the implicit cognitive pathway effectively externalizes latent semantics underlying textual surfaces, establishing a robust contextual foundation for subsequent rational analysis by the explicit cognitive pathway.

E. Explicit Cognitive Pathway Reasoning

After contextual completion of the implicit pathway, the framework transitions to the explicit cognitive pathway. This pathway emulates a judge delivering a rigorous verdict based on precedents, aiming to impose logical constraints and rational verification on intuitive clues through external criteria. Leveraging the mutual-indexing mechanism, the framework precisely maps activation paths from the general knowledge base to annotated precedents in the specialized knowledge base. For fine-grained clues, the framework directly traces their corresponding precedent texts, while for coarse-grained clues, it aggregates the flow of precedents under specific group knowledge clusters. All retrieved cases are then reranked and filtered based on their semantic similarity to the target text, thereby extracting decision rationales highly relevant to the current context.

To eliminate potential biases arising from imbalanced sample distributions, the framework selects an equal number of toxic and non-toxic annotated cases as stance references. These reference cases, containing original texts, explicit toxicity labels, and in-depth judgment rationales, are integrated into prompt templates to guide the decision-making process of LLMs. At this stage, the LLM functions as a judge, performing a rigorous reasoning for toxic detection on the clues provided by the implicit pathway according to the annotation boundaries and logical principles defined in the reference cases. Through this precedent-based stance calibration, the framework ultimately outputs highly interpretable detection results that align with the specific situational constraints and annotation stances.

TABLE I
GENERAL KNOWLEDGE BASE STATISTICS

	Samples	Entities	Relations
HateXplain	15,383	16,980	19,242
IHC	19,332	16,273	18,880
MERGE	34,715	30,321	38,087

IV. EXPERIMENTS

This section introduces the datasets and evaluation metrics used in our study. We also outline the comparative methods and experimental settings. Finally, we analyze the experimental results and discuss the generalizability of the approach.

A. Experimental Settings

We construct two individual knowledge database based on HateXplain [41] and IHC [42], and the training-test split follows MetaTox [21]. The training set for the MERGE database is compiled by merging the training sets of the two aforementioned datasets, while the test set remains consistent with that of the individual datasets. Samples from training set are employed to build both general knowledge base and specialized knowledge base. The proposed framework is then evaluated using the test set. Table I provides the statistical information regarding the general knowledge bases.

We use macro F1-score to evaluate various classifiers for hate speech detection, which measures the harmonic mean of precision and recall. The macro F1-score is computed by averaging the individual F1-scores for toxic speech and non-toxic speech categories, thereby assigning equal significance to each class and offering a balanced performance evaluation. The baseline methods include Vanilla LLM, RAG, and KGQA. Specifically, we utilize MetaTox [21] as the KGQA method. To the best of our knowledge, it is the only approach that integrates LLM with knowledge enhancement for hate speech detection. The pre-trained BAAI/bge-m3 [43] model is employed as the semantic representation model for knowledge alignment and case filtering. During decoupled dual-knowledge base construction stage, the LLM utilized is Qwen2.5-14B-Instruct (Qwen) [44]. In the dual-pathway cognitive reasoning stage, both Qwen and Llama3.1-8B-Instruct (Llama) [45] are employed. For a fair comparison with current RAG and KGQA approaches, the main experiment employs 1-hop retrieval from the general knowledge base while retaining 3 pairs for the specialized knowledge base.

B. Main Results

The main experimental results of our method are presented in Table II, where the experimental results for Vanilla LLM, RAG, and MetaTox are sourced from MetaTox [21]. The DCAdapt method in the table utilizes the same dataset for both

decoupled dual-knowledge base construction stage and dual-pathway cognitive reasoning stage, whereas DCAdapt-ME denotes a scenario where the general knowledge base employs MERGE dataset while specialized knowledge bases use their respective datasets, with validation conducted on corresponding datasets. Compared to baseline methods, DCAdapt demonstrates significantly superior performance across two LLMs with different parameter scales on two datasets, highlighting the effectiveness of our proposed approach in perceiving and understanding toxic criteria. Under the DCAdapt-ME setting, general knowledge base aggregates knowledge clues from two databases, constructing a more comprehensive semantic network. The additional performance improvement achieved by DCAdapt-ME also indicates that associating reasoning with context that is clearly positioned, semantically complete, and multifaceted can enhance the detection of toxic speech.

TABLE II
MAIN RESULTS OF DCADAPT

Backbone LLM	Method	HateXplain		IHC	
		Acc	F1	Acc	F1
Qwen	Vanilla LLM	70.95	64.04	66.34	66.32
	RAG	73.13	70.18	66.71	66.67
	MetaTox	73.39	72.48	73.65	69.95
	DCAdapt	74.84	73.04	75.98	75.20
	DCAdapt-ME	**75.00**	**74.83**	**76.58**	**75.88**
Llama	Vanilla LLM	63.36	48.11	50.79	48.03
	RAG	69.59	62.02	64.29	64.23
	MetaTox	68.87	62.50	69.04	68.55
	DCAdapt	72.25	71.53	70.34	68.67
	DCAdapt-ME	**74.12**	**74.02**	**71.18**	**68.82**

C. Ablation Study

The foundation of the DCAdapt framework consists of a general knowledge base and a specialized knowledge base, which are interconnected via a mutual-indexing mechanism. To evaluate the significance of these components, we conducted ablation studies on the mutual-indexing mechanism (MI), the coarse-grained general knowledge (CGK), and the fine-grained general knowledge (FGK), respectively. Ablating the mutual-indexing mechanism implies that the framework can no longer access the specialized knowledge base, which is functionally equivalent to ablating the specialized knowledge base itself. The experimental results are presented in Table III.

It can be observed that the performance of "w/o MI" undergoes a significant degradation. This serves as a deliberate validation of our architectural validity: hate speech detection necessitates specific precedents to align with evaluation standards. Without MI, the model loses the judgmental criteria provided by the specialized knowledge base and is forced to rely solely on its internal knowledge, naturally regressing to a vanilla LLM. This underscores the critical role of both the mutual-indexing mechanism and the specialized knowledge base. Furthermore, the table indicates that the negative impact of ablating the FGK on framework performance is more pronounced than that of ablating the CGK. This disparity can be attributed to the fact that coarse-grained knowledge, due to its inherent abstraction, provides relatively ambiguous reference cases. In contrast, fine-grained knowledge directly links to similar or highly relevant cases, thereby providing superior guidance for hate speech detection.

TABLE III
ABLATION RESULTS OF DCADAPT

Backbone LLM	Method	HateXplain		IHC	
		Acc	F1	Acc	F1
Qwen	DCAdapt	**74.84**	**73.04**	**75.98**	**75.20**
	-w/o MI	70.95	64.04	66.34	66.32
	-w/o CGK	72.45	69.66	74.77	73.90
	-w/o FGK	72.13	69.30	72.23	71.70
Llama	DCAdapt	**72.25**	**71.53**	70.34	**68.67**
	-w/o MI	63.36	48.11	50.79	48.03
	-w/o CGK	71.83	71.37	70.78	66.62
	-w/o FGK	70.22	69.86	**70.86**	66.83

D. Parameters Sensitivity Analysis

Impact of Reference Scale: To investigate the influence of reference scale on dual-pathway reasoning, we constructed six groups of balanced sample pairs. As illustrated of Fig 3 (a), performance exhibits a rapid ascent followed by stabilization, with inflection points appearing at 3 pairs for HateXplain and 5 pairs for IHC. This demonstrates a saturation effect where a limited set of precise precedents suffices to rectify the intuitive biases of the implicit pathway. Conversely, excessive reference information yields diminishing marginal returns and introduces semantic noise, reflecting a cognitive bottleneck that hinders the distillation of core decision criteria. Thus, DCAdapt achieves high-precision calibration through strategic knowledge association at minimal computational cost.

Impact of General Knowledge Retrieval Depth: As illustrated of Fig 3 (b), regarding retrieval depth, the transition from 0 to 1 hop triggers a substantial surge in performance, confirming that literal text representations are insufficient without neighborhood context. While local semantic noise causes a plateau between 1 and 3 hops, a significant deep-seated transition occurs at 4 hops. This resurgence reveals the framework's capacity to manage long-range dependencies by activating global semantic clusters and macro-social backgrounds. By bridging isolated fragments into coherent logical chains, DCAdapt successfully distills critical evidence from high-order recalls, achieving high-precision alignment with domain-specific standards.

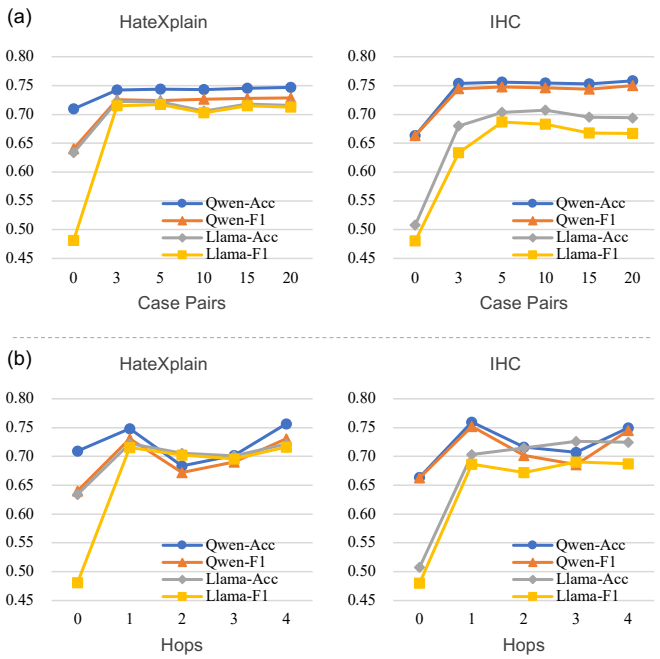

Fig. 3. An illustration of parameter sensitivity analysis of DCAdapt across different datasets. (a) Impact of reference scale. (b) Impact of general knowledge retrieval depth.

E. Case Study

To intuitively demonstrate the performance disparities among various models when processing implicit hate speech, we present a representative case in Table IV. The input test text is: "My traps help save the white race.", which comes from IHC dataset. Although this statement contains no direct profanity or slurs, it utilizes a classic white supremacist narrative and is officially classified as hate speech ([T]) in the ground-truth labels.

TABLE IV
CASE STUDY: EVALUATING THE EFFICACY OF DCADAPT

	Text: My traps help save the white race. **Ground Truth:** [T]
RAG	**Related Docs by KNN Search:** 1. looks white to me. [F] 2. wait you just liked your own tweet? oh man. the white race is in trouble after all. [F] 3. the key to maintaining a civilized society is maintaining the white race! deport the third worlders! [T] **Prediction:** [F]
MetaTox	**Related Triples by Graph Search:** 1. "The White Race" "BenefitFrom" "Black Man" 2. "White race" "Saving" "Humanity" 3. "God" "Helpout" "White Race" **Prediction:** [F]
DCAdapt	**Related Docs by Graph + KNN Search:** 1. true and they are using the black man for their benefit. [F] 2. white race saving humanity as usual! [T] 3. so it targets faggots and blacks looks like god is trying to help out the white race. [T] **Prediction:** [T]

Comparative analysis reveals that RAG and MetaTox fail (both [F]) due to neutral semantic noise and generalized knowledge limitations respectively. In contrast, DCAdapt leverages dual-cognition reasoning synergized with dual-knowledge bases: the implicit pathway employs semantic diffusion to uncover latent features amidst noise, while the explicit pathway utilizes case-based knowledge for logical calibration of the underlying incitement. By bridging perceptual association and rational verification, DCAdapt successfully deciphers multi-hop subtle clues, achieving an accurate prediction [T]. A broader data analysis also reveals that the framework successfully rescued several hard samples by correcting False Positives and False Negatives.

V. CONCLUSION

This paper proposes a dual-cognition driven framework for adaptive hate speech detection called DCAdapt and validates the effectiveness, flexibility, and cross-dataset adaptability of the DCAdapt framework through comprehensive experiments. In the future, we plan to further refine the process of constructing the knowledge base, optimize the quality of knowledge, and improve the knowledge-enhanced reasoning mechanism to boost the performance of hate speech detection.

REFERENCES

[1] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel, "The virality of hate speech on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–22, 2024.

[2] Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Lee, "Recent advances in online hate speech moderation: Multimodality and the role of large models," *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4407–4419, 2024.

[3] Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han, "Predict: multi-agent-based debate simulation for generalized hate speech detection," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 20963–20987.

[4] Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty, "Hate personified: Investigating the role of llms in content moderation pipeline for hate speech," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2024, pp. 15847–15863, Association for Computational Linguistics.

[5] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith, "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection," in *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2022, pp. 5884–5906.

[6] Neil Fasching and Yphtach Lelkes, "Model-dependent moderation: inconsistencies in hate speech detection across llm-based systems," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 22271–22285.

[7] Chu Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu, "Legally enforceable hate speech detection for public forums," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 10948–10963.

[8] Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh, "Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis," *arXiv preprint arXiv:2308.16705*, 2023.

[9] Badr Alkhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab, "Token: Task decomposition and knowledge infusion for few-shot hate speech detection," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2109–2120.

[10] Bogdan Mihai Guțu and Nirvana Popescu, "Exploring data analysis methods in generative models: From fine-tuning to rag implementation," *Computers*, vol. 13, no. 12, pp. 327, 2024.

[11] Yulong Wang, Hong Li, and Ni Wei, "Sahsd: Enhancing hate speech detection in llm-powered web applications via sentiment analysis and few-shot learning," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 3014–3025.

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

[13] Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović, "Angrybert: Joint learning target and emotion for hate speech detection," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2021, pp. 701–713.

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

[15] Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar, "Metahate: A dataset for unifying efforts on hate speech detection," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2024, vol. 18, pp. 2025–2039.

[16] Jan Fillies and Adrian Paschke, "Youth language and emerging slurs: tackling bias in bert-based hate speech detection," *AI and Ethics*, pp. 1–13, 2025.

[17] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill, ""hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media," *ACM Transactions on the Web*, vol. 18, no. 2, pp. 1–36, 2024.

[18] Fan Huang, Haewoon Kwak, and Jisun An, "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*, 2023, pp. 294–297.

[19] Richard Willats, Josh Pennington, Aravind Mohan, and Bertie Vidgen, "Classification is a rag problem: A case study on hate speech detection," *arXiv preprint arXiv:2508.06204*, 2025.

[20] Jianfa Chen, Emily Shen, Trupti Bavalatti, Xiaowen Lin, Yongkai Wang, Shuming Hu, Harihar Subramanyam, Ksheeraj Sai Vepuri, Ming Jiang, Ji Qi, et al., "Class-rag: Real-time content moderation with retrieval augmented generation," *arXiv preprint arXiv:2410.14881*, 2024.

[21] Yibo Zhao, Jiapeng Zhu, Can Xu, Yao Liu, and Xiang Li, "Enhancing LLM-based hatred and toxicity detection with meta-toxic knowledge graph," in *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, Eds., Vienna, Austria, July 2025, pp. 24747–24760, Association for Computational Linguistics.

[22] Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi, "Rag and recall: Multilingual hate speech detection with semantic memory," in *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, 2025, pp. 219–227.

[23] Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu, "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)*, 2024, pp. 12073–12086.

[24] Tsungcheng Yao, Ernest Foo, and Sebastian Binnewies, "Personalised abusive language detection using llms and retrieval-augmented generation," in *ICNLSP 2024: 7th International Conference on Natural Language and Speech Processing*. Association for Computational Linguistics, 2024.

[25] Daniel Kahneman, "Thinking, fast and slow," *Farrar, Straus and Giroux*, 2011.

[26] Jonathan St BT Evans and Keith E Stanovich, "Dual-process theories of higher cognition: Advancing the debate," *Perspectives on psychological science*, vol. 8, no. 3, pp. 223–241, 2013.

[27] Jia Yu, Zhiwei Huang, Lichao Zhang, Yu Lu, Yuming Yan, Zhenyang Xiao, and Zhenzhong Lan, "Benchmark for detecting child pornography in open domain dialogues using large language models," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–9.

[28] Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin, "State toxicn: A benchmark for span-level target-aware toxicity extraction in chinese hate speech detection," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 10206–10219.

[29] Shuzhou Yuan, Ercong Nie, Lukas Kouba, Ashish Yashwanth Kangen, Helmut Schmid, Hinrich Schütze, and Michael Färber, "Llm in the loop: Creating the paradehate dataset for hate speech detoxification," *arXiv preprint arXiv:2506.01484*, 2025.

[30] Xiaodong Wu, Hao Wu, and Pei He, "Leveraging domain-specific word embedding and hate concepts in hate speech detection," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[31] Angel Felipe Magnossão de Paula, Paolo Rosso, and Damiano Spina, "Mitigating negative transfer with task awareness for sexism, hate speech, and toxic language detection," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[32] Huan Ma, Changqing Zhang, Huazhu Fu, Peilin Zhao, and Bingzhe Wu, "Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning," *arXiv preprint arXiv:2310.03400*, 2023.

[33] Chen Han, Yijia Ma, Jin Tan, Wenzhen Zheng, and Xijin Tang, "Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion," *arXiv preprint arXiv:2511.07267*, 2025.

[34] Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su, "Hipporag: Neurobiologically inspired long-term memory for large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 59532–59569, 2024.

[35] S Ma, C Xu, X Jiang, et al., "Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[36] Y Mu, J Yang, T Li, et al., "Ha-gcen: Hyperedge-abundant graph convolutional enhanced network for hate speech detection," *Knowledge-Based Systems*, vol. 300, pp. 112166, 2024.

[37] M Cuccarini, L Draetta, B Fiumanò, et al., "Unveiling stereotypes: Combining knowledge graphs and llms for implied stereotype generation," in *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, 2025, pp. 314–326.

[38] Michał Bilewicz and Wiktor Soral, "Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization," *Political Psychology*, vol. 41, pp. 3–33, 2020.

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

[40] Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, et al., "Kag: Boosting llms in professional domains via knowledge augmented generation," in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 334–343.

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

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

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

[44] Qwen Team, "Qwen2.5: A party of foundation models," September 2024.

[45] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al., "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.