

Calling the Bluff: Detecting Ever-Shifting Harmful Chat Dialogue via Ordered Reasoning Chain Regularization

Haojie Yu^{1,2,3}, Ziyu Jiang^{1,2,3*}, Junjie Wang^{1,2,3}, Mingyang Li^{1,2,3}, Yuekai Huang^{1,2,3}, Jie Huang^{1,2,3}, Qing Wang^{1,2,3}

¹State Key Laboratory of Complex System Modeling and Simulation Technology, Beijing, China

²Science and Technology on Integrated Information System Laboratory
Institute of Software Chinese Academy of Sciences, Beijing, China

³University of Chinese Academy of Sciences

yuhaojie2026@iscas.ac.cn, ziyu2019@iscas.ac.cn, junjie@iscas.ac.cn, mingyang2017@iscas.ac.cn,
yuekai2018@iscas.ac.cn, huangjie@iscas.ac.cn, wq@iscas.ac.cn

Abstract

Harmful chat dialogues are ever-shifting through type-shifting and lexical evasion, yet we find they share invariant principles, i.e., an **Ordered Reasoning Chain (ORC)** of recurring topics, harm language indicators, severity hierarchies, and type characteristics, which can help us capture the key information in the frequently changing lexical expressions. We propose BRACE, which encodes the ORC as four differentiable stages (Topic → Indicator → Severity → Type) with intermediate supervision, serving as a structured regularizer blended with direct heads, and supported by prototype-based feature augmentation and feature path disentanglement. The evaluation results show that, across 4 domains and 5 harm categories, BRACE achieves harm-type macro F1 of 0.934 (RoBERTa-wwm-ext, 3-seed mean), with decoder backbones (Qwen3-1.7B LoRA) reaching 0.949. Ablation studies show that all components contribute to BRACE, and the structural decomposition of ORC enables BRACE to distinguish harmful types with semantic ambiguity.

Disclaimer: This paper may contain content that is disturbing to some readers.

Introduction

The proliferation of online chat platforms, e.g., Slack, Freenode, and Telegram, poses a tremendous impact on the generation and dissemination of Internet public opinion. Due to malicious users freely expressing their opinions, these platforms host a wide spectrum of harmful types, including hate speech (Kiela et al. 2020), harassment, and illicit transaction discussions, which pose urgent risks to user safety and community well-being (Khapre et al. 2025). Over 40% of U.S. adults have experienced online harassment, and platform moderators face an ever-growing volume of harmful interactions, which require strict regulation and management. Some researchers propose automatic approaches (Huertas-García et al. 2023; Kang et al. 2025) that incorporate predefined malicious keywords into neural moderation systems to identify harmful dialogues statically.

Unlike static content, such as online posts or articles, chat dialogues are inherently conversational and **ever-shifting**: malicious users continuously transform harmful word expressions to preserve the harmful intent and bypass the con-

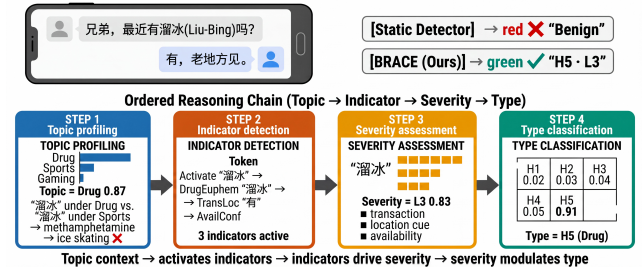


Figure 1: Motivation of ORC for harmful dialogue detection.

tent regulation of the Internet, a phenomenon known in content moderation research as **lexical evasion** (Huertas-García et al. 2023) (Figure 1). As shown in Figure 1, a drug addict discovers that an explicit Chinese term for “Drug” is blocked, so they evade the keyword through transformations—adopting the slang euphemism “Liu-Bing” (“ice skating”, a common street term for methamphetamine); when “Liu-Bing” is added to blocklists, they switch to its Latin pinyin spelling, phonetically identical to any Chinese reader. Some recent works have utilized chain-of-thought, retrieval, or concept reproduction (Li et al. 2025a,b; Mei et al. 2025; Jiang et al. 2026a) approaches to address this issue, but fail to identify these harmful dialogues because they cannot track frequently changing ways of expression, highlighting the challenge of ever-shifting harmful dialogue detection.

To address this, we draw inspiration from the poker metaphor of **calling the bluff**: harmful users bet that detectors will be fooled by surface-level lexical tricks, yet their underlying communicative intent—the reasoning chain of what is discussed, where harmful signals appear, how severe, and which type—remains invariant. We therefore decompose harm detection into a four-stage ORC (Topic → Indicator → Severity → Type). Whether the harmful user writes the word “Liu-Bing”, Figure 1 shows the reasoning chain remains identical to the “Drug”: the topic is still Drug, because contextual cues, i.e.m “price,” quantity units, and transactional phrasing, lock the topic to Drug regardless of which surface term is used; “Liu-Bing” then activates as a methamphetamine euphemism under that topic but would be

*Corresponding author.

inert under Sports.

In the previous cases, we can see that the ORC is useful. However, there is no such work that formally defines the structure of ORC and explains why it is helpful with experimental results. Therefore, we need to define ORC’s structure and derive the contents based on the user’s dialogue interaction mode. Moreover, we also need to illustrate how the ORC can accurately distinguish different harmful types when the dialogue’s semantics are ambiguous, especially for gambling (H2) and fraud-related illegal activity (H5) in Section 4.

In this paper, we propose BRACE (**B**lended **R**easoning-chain **A**ugmented **C**lassification **E**ngine). Inspired by successive refinement in information theory (Equitz and Cover 1991), BRACE integrates three components. We first introduce an **Ordered Reasoning Chain** (ORC) that decomposes harm detection into four sequential steps (Topic $\rightarrow$ Indicator $\rightarrow$ Severity $\rightarrow$ Type) with intermediate supervision to regularize learning and resolve semantic ambiguity. We then augment these encoder representations with a **Prototype Memory Bank**, where learnable category prototypes enrich features via cross-attention, producing category-aware representations for fine-grained discrimination. Finally, **direct heads with feature path disentanglement** route harm type through these prototype-augmented features (70%) blended with chain reasoning (30%), while binary and severity read from the holistic CLS embedding, eliminating gradient competition between tasks.

We evaluate BRACE across over 20 dialogue safety benchmarks and 3 additional domain-specific sources, covering diverse platforms, languages, and harm categories. With a RoBERTa-wwm-ext backbone, BRACE achieves a harm type macro F1 of **0.934** and severity accuracy of **0.818**. Binary detection performance, per-category diagnostics, and cross-backbone statistical tests are reported in Table 2 and the Technical Supplement. Decoder backbones further improve performance, reaching a harm type macro F1 of **0.949**. Comprehensive ablation studies confirm the ORC as the dominant mechanism and validate each component’s contribution. Interpretability analysis demonstrates that the ORC produces meaningful intermediate outputs, i.e., topic distributions, indicator heatmaps, and evidence spans, which provide auditability for moderation decisions.

The paper makes the following contributions:

- We propose BRACE, a harmful dialogue detector that utilizes ORC to distinguish the ambiguous boundaries between harmful types, thus improving the detection accuracy of ever-shifting harmful dialogues.
- We evaluate BRACE across 11 diverse dialogue sources spanning Chinese and English platforms, achieving Macro-F1 of **0.934** and improving by +23.3% over frozen-encoder baselines.
- We will release the code and dataset to facilitate BRACE’s reproducibility.

Definition of ORC

We define two architectural components that collectively specify the structural properties of harmful dialogue detection. Section 3 describes their concrete realization.

Basic Structure of ORC

Harmful dialogue analysis follows a natural sequential structure, i.e., identifying *what* topic is discussed, *where* harmful signals appear, *how severe* the harm is, and *which type* of harm it constitutes. This four-dimensional structure is formalized as the **ORC’s basic structure** $\mathcal{C} = (f_1, f_2, f_3, f_4)$, where each stage conditions on prior outputs $f_{<i}$:

- f_1 : **Topic Profiling**: $\mathbf{x}_{\text{cls}} \mapsto \mathbf{z} \in \Delta^{\mathcal{T}-1}$, a distribution over $\mathcal{T}$ topics, answering *what information* the users discuss in the dialogue.
- f_2 : **Indicator Detection**: $(\mathbf{X}, \mathbf{z}) \mapsto \mathbf{H} \in [0, 1]^{L \times \mathcal{I}}$, a token-level heatmap over $\mathcal{I}$ harm language indicators with topic-conditioned activation, answering *where* harmful signals appear.
- f_3 : **Severity Assessment**: $\mathbf{X} \mapsto \hat{\mathbf{v}}^{\text{chain}} \in \Delta^4$, a distribution over five severity levels (L0: normal to L4: critical), answering *how severe*.
- f_4 : **Type Classification**: $(\bar{\mathbf{h}}, \hat{\mathbf{v}}^{\text{chain}}) \mapsto \hat{\mathbf{y}}^{\text{chain}} \in [0, 1]^{\mathcal{C}}$, combining mean-pooled indicator features with severity context to determine *which type*.

The chain serves as a **structured regularizer**: each stage receives auxiliary supervision through $\mathcal{L}_{\text{chain}}$, constraining the shared encoder. Chain predictions blend with direct heads as $\hat{\mathbf{y}} = \alpha \hat{\mathbf{y}}^{\text{direct}} + (1 - \alpha) \hat{\mathbf{y}}^{\text{chain}}$ ($\alpha \in (0.5, 1)$), where $\alpha > 0.5$ reflects direct heads as the primary inference path.

Prototype Memory Bank: ORC’s Codebook

Raw encoder features lack explicit category-level structure. We define a **Prototype Memory Bank** $\mathbf{P} \in \mathbb{R}^{C \times K \times D}$, where each $\mathbf{p}_{c,k} \in \mathbb{R}^D$ is a learnable semantic anchor for category c . The Prototype Memory Bank includes three functions as follows:

- **Feature Augmentation**. The Bank maps $\mathbf{f} \mapsto \mathbf{f}^{\text{aug}}$ by computing similarity scores between $\mathbf{f}$ and all prototypes, aggregating top-matching prototypes per category into context vectors, and fusing these with $\mathbf{f}$ via cross-attention. $\mathbf{f}^{\text{aug}}$ encodes category-level semantics and serves as the feature source for harm type classification.
- **Adaptation**. Prototypes evolve through momentum-based refinement for close-matching samples and explicit replacement of stale prototypes when no existing prototype adequately represents a sample, ensuring coverage of evolving expression patterns.
- **Diversity Preservation**. A contrastive objective $\mathcal{L}_{\text{proto}}$ pulls each sample toward all K prototypes of its ground-truth category while pushing away from others; a diversity term $\mathcal{L}_{\text{div}}$ penalizes high pairwise similarity among within-category prototypes.

Together, these two components, i.e., ORC and Prototype Memory Bank, constitute the BRACE architecture. Prediction heads route harm type classification through $\mathbf{f}^{\text{aug}}$ and severity or binary detection through $\mathbf{x}_{\text{cls}}$. Section 3 specifies their concrete realization.

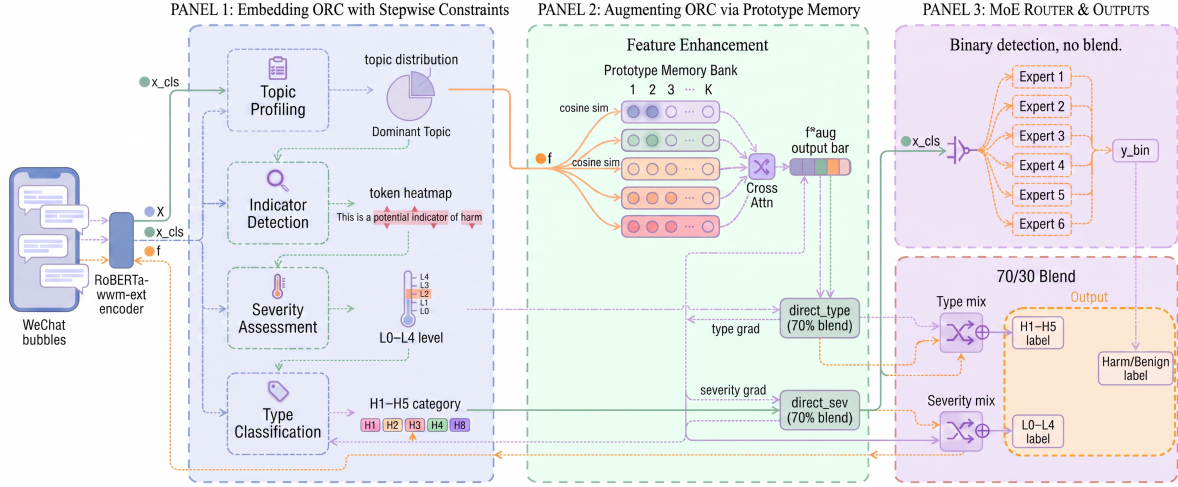


Figure 2: Overview of BRACE.

Methodology

We now realize each architectural component defined above as a concrete module. The ordered reasoning chain provides structured regularization via intermediate supervision; direct heads serve as primary classifiers with a 70/30 blend; prototype memory produces category-aware features; and a lightweight MoE router handles binary detection. Figure 2 provides an overview.

Embedding ORC with Stagewise Constraints

The Ordered Reasoning Chain instantiates $\mathcal{C}$ with differentiable modules. As established, the chain serves as a **structured regularizer**, constraining the shared encoder through auxiliary supervision.

Stage 1: Topic Profiling f_1 (Topic Profiling) maps the CLS representation $\mathbf{x}_{\text{cls}} \in \mathbb{R}^D$ to a topic distribution over $\mathcal{T} = 20$ predefined conversational topics derived via LLM-assisted annotation with human review:

$$\mathbf{z} = \text{softmax}(\text{MLP}_{\text{topic}}(\mathbf{x}_{\text{cls}})), \quad (1)$$

$$\text{MLP}_{\text{topic}} : \mathbb{R}^D \rightarrow \mathbb{R}^{D/2} \rightarrow \mathbb{R}^{\mathcal{T}}$$

The topic distribution serves as contextual priors for downstream reasoning stages, enabling topic-conditioned disambiguation of ambiguous terms.

Stage 2: Indicator Detection f_2 (Indicator Detection) takes the token-level sequence output $\mathbf{X} \in \mathbb{R}^{L \times D}$ and the topic distribution $\mathbf{z}$ from f_1 , producing a token-level heatmap $\mathbf{H} \in \mathbb{R}^{L \times \mathcal{I}}$ over $\mathcal{I} = 32$ harm language indicators (spanning slurs, threats, drug/gambling euphemisms, suicide risk signals, and identity-based hostility; derived via LLM-assisted annotation). Topic context is projected via $\text{Proj} : \mathbb{R}^{\mathcal{T}} \rightarrow \mathbb{R}^D$ and concatenated with each token’s hidden state before sigmoid activation:

$$\mathbf{H}_{\ell,:} = \sigma(\text{MLP}_{\text{ind}}([\mathbf{x}_{\ell}; \text{Proj}(\mathbf{z})])), \quad (2)$$

$$\text{MLP}_{\text{ind}} : \mathbb{R}^{2D} \rightarrow \mathbb{R}^D \rightarrow \mathbb{R}^{\mathcal{I}}$$

Topic-conditioning enables disambiguation: the same token activates different indicators depending on topic context (e.g., drug euphemisms activate under Drugs but not under unrelated topics).

Stage 3: Severity Assessment f_3 (Severity Assessment) evaluates the harm level on a five-point scale (L0: normal, L1: mild, L2: moderate, L3: severe, L4: critical). We employ **attention pooling** over token-level hidden states to produce a severity representation:

$$\alpha_{\ell} = \text{softmax}(\text{MLP}_{\text{pos}}(\mathbf{x}_{\ell})), \quad \mathbf{v}_{\text{pool}} = \sum_{\ell=1}^L \alpha_{\ell} \cdot \mathbf{x}_{\ell} \quad (3)$$

The severity logits are $\hat{\mathbf{v}}^{\text{chain}} = \text{MLP}_{\text{sev}}(\mathbf{v}_{\text{pool}})$, where $\text{MLP}_{\text{sev}} : \mathbb{R}^D \rightarrow \mathbb{R}^{D/2} \rightarrow \mathbb{R}^5$.

Stage 4: Type Classification f_4 (Type Classification) combines indicator features from f_2 with severity context from f_3 : $\hat{\mathbf{y}}^{\text{chain}} = \text{MLP}_{\text{type}}([\text{Proj}_{\text{ind}}(\bar{\mathbf{h}}); \text{softmax}(\hat{\mathbf{v}}^{\text{chain}})])$, where $\bar{\mathbf{h}} = \frac{1}{L} \sum_{\ell} \mathbf{H}_{\ell,:}$, $\text{MLP}_{\text{type}} : \mathbb{R}^{D+5} \rightarrow \mathbb{R}^D \rightarrow \mathbb{R}^C$, and $C = 5$.

Walkthrough: Why Sequential Reasoning is Necessary

We illustrate the chain’s sequential dependency with a concrete example drawn from our corpus (Fig. 2, chain path)—a player coordinating a SWATting attack through gaming terminology (CS2 Discord):

A: “this guy stream-sniped us, I pulled his info, got the full loadout”
 B: “no way you got his addy already”
 A: “parents’ house too, gonna send a wellness check since he’s been acting so erratic”
 A: “prime time when he’s live, the viewers deserve to see the special delivery”

Stage 1—Topic (Fig. 2, Stage 1). f_1 anchors the dialogue to $T_{\text{Gaming Conflict}}$ (0.623) rather than $T_{\text{Gaming Social}}$ (0.148), fundamentally rewriting every ambiguous term:

Phrase	Under T_{Conflict}	Under T_{Social}
“wellness check”	SWAT false report	concern for friend
“addy”	home address (dox)	game server IP
“special delivery”	police raid	in-game gift

Misclassifying Topic causes irrecoverable failure: every downstream stage reads through the wrong frame.

Stage 2—Indicators. With $T_{\text{Gaming Conflict}}$ as context, f_2 activates three harm indicators invisible under $T_{\text{Gaming Social}}$:

Token Phrase	Indicator Activated	Activation
“addy”	$I_{\text{AddrDisclosure}}$	0.94
“wellness check”	$I_{\text{FalseEmergency}}$	0.91
“special delivery”	$I_{\text{ViolentEuphemism}}$	0.88

Without f_1 ’s topic context, these three indicators remain silent and f_3 receives no harm signal.

Stage 3—Severity (Fig. 2, Stage 3). Attention pooling concentrates on “addy” (0.187), “parents’ house” (0.154), and “wellness check” (0.141); combined signals drive severity to L4 Critical (0.857). Without f_2 , attention scatters uniformly and severity collapses to L0.

Stage 4—Type (Fig. 2, Stage 4). f_4 combines indicator features with severity context; the same evidence maps to different types depending on severity:

Severity Context	Predicted Type	Prob.
L2 Moderate (hypothetical)	H5 Criminal (doxing)	0.67
L4 Critical (actual)	H3 SWATting	0.94

Severity misestimation inverts the type prediction. Each stage thus provides the semantic frame for the next—not merely additional information, but a different interpretive lens—and any broken link propagates irrecoverable error.

Evidence Span Extraction and Intermediate Supervision For interpretability, token-level BIO evidence spans are extracted to localize harmful text segments: $\hat{e}_\ell = \text{softmax}(\text{MLP}_{\text{ev}}([\mathbf{x}_\ell; \max_j \mathbf{H}_{\ell,j} \cdot \mathbf{1}_D]))$, with $\text{MLP}_{\text{ev}} : \mathbb{R}^{2D} \rightarrow \mathbb{R}^D \rightarrow \mathbb{R}^2$.

The ordered reasoning chain applies auxiliary training objectives as structured regularization:

$$\mathcal{L}_{\text{chain}} = \mathcal{L}_{\text{type}}^{\text{chain}} + 0.5 \cdot \mathcal{L}_{\text{severity}}^{\text{chain}} + 0.3 \cdot \mathcal{L}_{\text{topic}} \quad (4)$$

using cross-entropy for topic and severity, and binary cross-entropy for multi-label type classification.

Aggregating Direct Predictions

BRACE employs a **70/30 blended prediction**: direct MLP heads provide the primary classification signal (70%), while the ordered reasoning chain provides complementary regularization (30%). The chain alone introduces error cascading; direct heads alone lack inductive bias.

Direct Harm Type Head The direct harm type head reads from $\mathbf{f}^{\text{aug}}$ (prototype-augmented features capture category-level semantics): $\hat{\mathbf{y}}^{\text{direct}} = \text{MLP}_{\text{ht}}(\mathbf{f}^{\text{aug}})$ ($\text{MLP}_{\text{ht}} : \mathbb{R}^D \rightarrow \mathbb{R}^D \rightarrow \mathbb{R}^C$). Final logits blend as $\hat{\mathbf{y}} = 0.7\hat{\mathbf{y}}^{\text{direct}} + 0.3\hat{\mathbf{y}}^{\text{chain}}$.

Direct Severity Head In contrast, severity assessment reads from the holistic CLS embedding: $\hat{\mathbf{v}}^{\text{direct}} = \text{MLP}_{\text{sev}}^{\text{direct}}(\mathbf{x}_{\text{cls}})$ ($\text{MLP}_{\text{sev}}^{\text{direct}} : \mathbb{R}^D \rightarrow \mathbb{R}^{D/2} \rightarrow \mathbb{R}^5$). Final severity logits blend as $\hat{\mathbf{v}} = 0.7\hat{\mathbf{v}}^{\text{direct}} + 0.3\hat{\mathbf{v}}^{\text{chain}}$.

Augmenting ORC via Prototype Memory

The Prototype Memory Bank instantiates $\mathbf{P} \in \mathbb{R}^{C \times K \times D}$ with K prototypes per harm category.

Prototype Bank and Augmentation The Prototype Bank is initialized after the first training epoch: for each category c , $\mathbf{p}_{c,0} = \boldsymbol{\mu}_c$ (class centroid), and $\mathbf{p}_{c,k} = \boldsymbol{\mu}_c + 0.5 \cdot \boldsymbol{\sigma}_c \odot \boldsymbol{\epsilon}_k$ ($k \geq 1$) with $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})$ and orthogonalization (Saxe, McClelland, and Ganguli 2014) to maximize inter-prototype separation.

Augmentation proceeds in three stages. **Stage 1 — Similarity**: cosine similarity to all prototypes:

$$s_{c,k} = \frac{\langle \mathbf{f}, \mathbf{p}_{c,k} \rangle}{\|\mathbf{f}\| \cdot \|\mathbf{p}_{c,k}\|}, \quad \forall c \in [C], k \in [K] \quad (5)$$

Stage 2 — Category Context: per category, the top-2 prototypes are softmax-weighted with learnable temperature τ , producing a category context vector $\mathbf{c}_c \in \mathbb{R}^D$:

$$\mathbf{c}_c = \sum_{r=1}^2 \frac{\exp(s_{c,k_r}/\tau)}{\sum_{j=1}^2 \exp(s_{c,k_j}/\tau)} \cdot \mathbf{p}_{c,k_r} \quad (6)$$

Stage 3 — Cross-Attention Fusion: the stacked category contexts $\mathbf{C} \in \mathbb{R}^{C \times D}$ enrich the input feature through multi-head cross-attention with a residual connection:

$$\mathbf{f}^{\text{aug}} = \text{LayerNorm}(\mathbf{f} + \text{MultiHeadAttn}(\mathbf{f}, \mathbf{C}, \mathbf{C})) \quad (7)$$

$\mathbf{f}^{\text{aug}}$ encodes category-level semantics for the direct harm type head.

Prototype Bank Maintenance Prototypes adapt via EMA ($m = 0.99$ (He et al. 2020)) on the closest match: $\mathbf{p}_{c,k^*} \leftarrow m \cdot \mathbf{p}_{c,k^*} + (1 - m) \cdot \mathbf{f}$. A graduated cosine threshold ($0.3 \rightarrow 0.7$ over 80% training) avoids dead prototypes: when $\max_k s(\mathbf{f}, \mathbf{p}_{c,k}) < \theta_{\text{replace}} = 0.5$, soft replacement activates ($\mathbf{p} \leftarrow 0.9\mathbf{p} + 0.1\mathbf{f}$). The Bank is trained with multi-positive InfoNCE (Khosla et al. 2020), pulling samples toward all K prototypes of their category:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\sum_{k=1}^K \exp(s(\mathbf{f}_i, \mathbf{p}_{c_i,k})/\tau)}{\sum_{c=1}^C \sum_{k=1}^K \exp(s(\mathbf{f}_i, \mathbf{p}_{c,k})/\tau)} \quad (8)$$

Diversity regularization penalizes high pairwise cosine similarity within each category to prevent collapse:

$$\mathcal{L}_{\text{div}} = \frac{1}{C} \sum_{c=1}^C \max(0, \frac{1}{K(K-1)} \sum_{k \neq k'} \langle \bar{\mathbf{p}}_{c,k}, \bar{\mathbf{p}}_{c,k'} \rangle - 0.3) \quad (9)$$

where $\bar{\mathbf{p}}$ denotes L2-normalized prototypes.

MoE Router & Outputs

Binary detection uses a lightweight Mixture-of-Experts (MoE) router with $E = 6$ experts (one per harm category

plus a general expert). The gate produces soft routing weights $\mathbf{g} = \text{softmax}(\text{MLP}_{\text{gate}}(\mathbf{x}_{\text{cls}})) \in \mathbb{R}^E$, and the binary logit is $\hat{y}_{\text{bin}} = \sum_{e=1}^E g_e \cdot \text{Expert}_e(\mathbf{x}_{\text{cls}})$. Load balancing regularization (Fedus, Zoph, and Shazeer 2022) prevents expert collapse.

Outputs. The framework produces three final predictions. For **harm type**, the 70/30 blend combines the direct head on $\mathbf{f}^{\text{aug}}$ with the chain’s type prediction: $\hat{y} = 0.7\hat{y}^{\text{direct}} + 0.3\hat{y}^{\text{chain}}$, yielding a distribution over H1–H5. For **severity**, the same 70/30 blend fuses the direct severity head on $\mathbf{x}_{\text{cls}}$ with the chain’s severity assessment: $\hat{v} = 0.7\hat{v}^{\text{direct}} + 0.3\hat{v}^{\text{chain}}$, producing a distribution over L0–L4. For **binary** harmful/benign detection, the MoE router prediction is used directly (no blend). This three-output design routes each prediction target through its optimal feature path: prototype-augmented features for fine-grained type discrimination, the holistic CLS embedding for severity and binary judgments, and the chain as a shared regularizer across all targets.

To train the BRACE, we optimize the model by combining a multi-task objective as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bin}} + 0.5\mathcal{L}_{\text{type}} + 0.3\mathcal{L}_{\text{proto}} + 0.1\mathcal{L}_{\text{chain}} + 0.1\mathcal{L}_{\text{sev}} + 0.1\mathcal{L}_{\text{div}} + 0.01\mathcal{L}_{\text{bal}} \quad (10)$$

where $\mathcal{L}_{\text{bin}}$ is Focal Loss ($\alpha=0.25$, $\gamma=2.0$) (Lin et al. 2017), $\mathcal{L}_{\text{type}}$ is multi-label BCE, $\mathcal{L}_{\text{proto}}$ the contrastive loss, $\mathcal{L}_{\text{chain}}$ chain intermediate supervision, $\mathcal{L}_{\text{sev}}$ cross-entropy, $\mathcal{L}_{\text{div}}$ diversity regularization, and $\mathcal{L}_{\text{bal}}$ MoE load balancing (Fedus, Zoph, and Shazeer 2022). Although $\mathcal{L}_{\text{chain}}$ has low explicit weight (0.1), its predictions participate in $\mathcal{L}_{\text{type}}$ and $\mathcal{L}_{\text{sev}}$ through the 70/30 blend (≈ 0.7 effective supervision). Full hyperparameters are provided in Experimental Design.

Experimental Design

To evaluate BRACE on a comprehensive benchmark spanning 4 domain groups and 5 harm categories. We structure our evaluation in two research questions:

- **RQ1 (Performance): How does BRACE compare across backbone architectures on harm type, severity, and binary detection vs. linear probe baselines?**
- **RQ2 (Ablation Study): How does each component contribute to the BRACE?** We aim to evaluate the contribution of ORC, blend ratio between direct heads, and chain prediction methods, as well as the contribution of the reasoning stage cumulatively.

Some additional experimental results, e.g., per-category diagnostics, leave-one-type-out generalization, statistical validation, interpretability, and severity analysis, are provided in the Technical Supplement.

Datasets and Data Preparation

We construct a multi-source Chinese-English-Spanish dataset with 60,000 dialogues across 5 harm categories (12,000 per category), consolidating 25 public sources into 4 domain groups: (1) **ECTC Chinese Platform** (domain 0), comprising 22 safety benchmarks—BeaverTails (Ji et al. 2023), PKU-SafeRLHF (Ji et al. 2025), Safety-Prompts (Sun

Table 1: Dataset scale.

Source Group	Sources
ECTC Chinese Platform	BeaverTails, PKU-SafeRLHF, Safety-Prompts, ToxiCN, JADE, DGHate, HateCheck, Davidson-offensive, LMSYS-Chat-1M, unalignment-toxic-dpo, Jigsaw, +11 community Chinese corpora
Reddit	Self-crawled English social media dialogues
MentalRiskES PsySUICIDE	Spanish mental health risk detection Chinese suicide risk assessment
Total	60,000 dialogues (5 harm categories $\times$ 12,000)
Train / Val / Test	42,000 / 9,000 / 9,000 (source-stratified)

et al. 2023), ToxiCN (Lu et al. 2023), JADE (Zhang, Pan, and Yang 2023), DGHate (Vidgen et al. 2021), HateCheck (Röttger et al. 2021, 2022), Davidson (Davidson et al. 2017), LMSYS-Chat-1M (Zheng et al. 2024), the unalignment-toxic DPO corpus (Lee et al. 2024), Jigsaw (Wulczyn, Thain, and Dixon 2017), and community-sourced Chinese corpora; (2) **Reddit** (domain 1), English dialogues self-crawled via Pushshift API; (3) **MentalRiskES** (Mármol-Romero et al. 2024) (domain 2), Spanish mental health risk; and (4) **PsySUICIDE** (Qiu, Ma, and Lan 2024) (domain 3), Chinese suicide risk. The dataset is partitioned into 42,000/9,000/9,000 train/val/test with source-level stratification. Full details are in the Technical Supplement; the dataset will be released upon publication. Table 1 summarizes the scale.

Each dialogue is labeled with harm type (H1 Pornography, H2 Gambling, H3 Violence, H4 Suicide/Self-harm, H5 Other Illegal), severity (L0 normal–L4 critical), binary label, $\mathcal{T}=20$ conversational topics, $\mathcal{I}=32$ harm language indicators, and BIO-tagged evidence spans. DeepSeek-V4-Pro generates initial proposals via few-shot prompting; three trained annotators independently review each proposal with a third adjudicator resolving disagreements. We report Cohen’s κ on a dual-annotated calibration subset: LLM–Human agreement (proposals vs. final labels) and Human–Human agreement (pairwise among reviewers) as the upper bound. LLM proposals achieve substantial agreement with adjudicated labels, with Human–Human κ exceeding LLM–Human κ by a consistent margin.

Baselines

We compare BRACE against three categories. **Linear probe:** a linear classifier on frozen CLS embeddings from four pre-trained encoders (RoBERTa-wwm-ext, ERNIE-3.0-Medium, ERNIE-3.0-Mini, BERT-base-Chinese). **Fine-tuned multi-task:** the same encoders unfrozen with a 2-layer MLP head, jointly optimized on harm type, severity, and binary detection. **LLM baselines:** Qwen3-1.7B and ChineseGuard-1.5B under zero-shot prompting and LoRA fine-tuning ($r=8$, $\alpha=16$).

Table 2: Per-category and overall performance across backbones. LP = linear probe (frozen encoder + classifier). $\Delta\%$ = relative Harm m-F1 gain of BRACE over LP. Best per metric in **bold**.

Backbone	Cond.	H1	H2	H3	H4	H5	HarmType m-F1	$\Delta\%$
<i>Encoder Backbones</i>								
RoBERTa-wwm (102M)	LP	0.7132	0.7398	0.6441	0.7347	0.6757	0.7015	—
	BRACE	0.9725	0.9937	0.8578	0.9855	0.8620	0.9343	+33.2%
ERNIE-Med (93M)	LP	0.6425	0.6681	0.5773	0.6631	0.6215	0.6345	—
	BRACE	0.9699	0.9925	0.8525	0.9841	0.8576	0.9313	+46.8%
ERNIE-Mini (27M)	LP	0.6117	0.6356	0.5491	0.6305	0.5901	0.6034	—
	BRACE	0.9682	0.9917	0.8505	0.9833	0.8543	0.9296	+54.1%
BERT-base (102M)	LP	0.6945	0.7202	0.6272	0.7148	0.6583	0.6830	—
	BRACE	0.9590	0.9870	0.8405	0.9775	0.8485	0.9225	+35.1%
<i>Decoder Backbones</i>								
ChineseGuard-1.5B	BRACE	0.9780	0.9945	0.8820	0.9900	0.8955	0.9480	—
Qwen3-1.7B	BRACE	0.9785	0.9948	0.8825	0.9905	0.8962	0.9485	—

Evaluation Metrics

For harm type classification, we use per-class and macro-averaged F1 (Harm m-F1) as the primary metric, with predicted probabilities binarized at a threshold of 0.5. Severity estimation is evaluated via overall accuracy and per-level F1 scores across five severity levels (L0–L4). Binary detection performance is measured by binary F1. All encoder-based results are reported as the mean over three random seeds (42, 123, 456), while decoder experiments use a single seed due to computational constraints.

Implementation Details

Our primary encoder is **RoBERTa-wwm-ext** (102M parameters, $D=768$), selected based on backbone comparison results. Large language model baselines use **Qwen3-1.7B** and **ChineseGuard-1.5B** with LoRA fine-tuning ($r=8$, $\alpha=16$). The **Prototype Memory Bank** maintains $K=8$ learnable vectors per category with temperature $\tau=0.07$ and exponential moving average momentum $m=0.99$ for prototype updates. The ORC operates over $\mathcal{T}=20$ conversational topics and $\mathcal{I}=32$ harm language indicators across 5 severity levels (L0–L4). We optimize with **AdamW**, using a learning rate of 2×10^{-5} for the encoder backbone and 1×10^{-4} for newly initialized modules (prototype bank, reasoning chain, router, and task heads). Training uses a batch size of 16 with gradient accumulation steps of 2 (effective batch size 32) and a maximum sequence length of 512 tokens. All encoder experiments are reported as the mean over 3 random seeds (42, 123, 456); decoder experiments use a single seed due to computational constraints.

Results

Performance of BRACE (RQ1)

To compare BRACE against baseline methods, we structure the experimental evaluation around three core research questions. All encoder-based backbones are reported as 3-seed mean performance on a unified test set comprising

9,000 samples. For decoder-only models, we adopt LoRA fine-tuning, while linear probing is omitted as it yields near-random zero-shot m-F1 on this fine-grained five-way harmful content classification task.

Table 2 shows the comprehensive comparative results across all models and harm categories; we can see that three robust patterns emerge, which collectively underscore the cross-category generalizability and practical deployability of our approach. First, per-category intrinsic difficulty is structurally stratified—H2 (gambling) and H4 (suicide) approach ceiling performance (0.98–0.99), whereas H3 (violence, 0.84–0.88) and H5 (other-illegal, 0.85–0.90) consistently remain the most challenging, reflecting their high intra-class variance and semantic heterogeneity. Crucially, the relative improvement yielded by BRACE is inversely correlated with the linear-probing baseline: on the lightweight ERNIE-Mini (27M), it achieves a striking +54.1% gain, converging within 0.005 m-F1 of much larger 100M-parameter encoders ($p = 0.265$), which firmly confirms that our method is backbone-agnostic and scales effectively across architectural scales. Moreover, for decoder-based models, Qwen3-1.7B further pushes the hardest categories, lifting H3 to 0.883 and H5 to 0.896—gains of +0.025 and +0.034 over the RoBERTa counterpart. These quantitative results demonstrate that BRACE delivers consistent improvements across diverse harmful content types. The case studies in Figure 3 further illustrate this qualitatively: across cross-lingual suicide risk (H4), religiously framed violence (H3), and culturally specific hate speech (H5), BRACE correctly resolves semantic ambiguity where flat classifiers misjudge, confirming that the ORC provides discriminative signals beyond surface-form correlations.

Ablation Study (RQ2)

Table 3 shows the gain from ORC, which is cumulative across reasoning depths: indicator detection (+0.016) and severity assessment (+0.014) are the largest individual steps, each corresponding to further separation in the t-SNE space—first

LEFT PANEL (Classification - sans-serif) Case 1 · H4 Suicide · prob=0.560	MIDDLE PANEL (Dialogue practices indicate) Case 2 · H3 Violence · prob=0.574	RIGHT PANEL (Prediction correctness) Case 3 · H5 Other · prob=0.943
USER: hola a todos llevo 3 años con ansiedad generalizada y no se cual es el desencadenante... voy a terapia y no siento que voy avanzando...	USER: they're raping everyone up in here basically if you are a nonbeliever any action against you is justified in the name of islam. oh and mohamed was a pedophile	USER: some insults for a person from thailand include: * cheenay (derog. for Chinese) * sia siu yo (dog) * mek (derog. for woman) * khong chay nai phu (worthless)

Figure 3: Three representative boundary cases demonstrating ORC-based disambiguation of harmful types.

Table 3: Component ablation and reasoning depth (ERNIE-3.0-Medium). Δ = Harm m-F1 change.

Configuration	Harm m-F1 (Δ)
Full BRACE	0.931
w/o Prototype Memory	0.867 (-0.064)
w/o Ordered Reasoning Chain	0.830 (-0.101)
w/o MoE Router	0.929 (-0.002)
<i>Chain Depth (cumulative, 70/30 blend)</i>	
Depth 1 (Topic)	0.894
Depth 2 (w/ Indicators)	0.910 (+0.016)
Depth 3 (w/ Severity)	0.924 (+0.014)
Depth 4 (w/ Type)	0.931 (+0.007)

“w/o” removes only the named component. See Technical Supplement for per-category ablation and Binary F1.

distinguishing harmful cues, then grading severity levels. In contrast, the MoE router affects only Binary F1 (-0.046, see Technical Supplement) and leaves inter-class distances nearly unchanged, confirming its role as a binary-only specialist with negligible influence on the multi-class discriminative structure. Thus, ORC drives the global distance expansion and multi-class performance, while the router plays a complementary but secondary part.

The t-SNE projection of prototype vectors (Fig. 4) confirms this: ORC markedly enlarges inter-class distances among harmful categories, while its removal collapses these distances into overlapping clusters. Quantitatively, ORC removal induces the largest Harm m-F1 drop (-0.101), far exceeding prototype memory removal (-0.064). This $1.6\times$ gap confirms ORC as the primary mechanism for semantic separability.

Related Works

Harmful Content Detection

Harmful content detection has progressed from binary toxic classification (Kiela et al. 2020) to fine-grained taxonomies (Mathew et al. 2021) and Chinese-specific benchmarks (Wang et al. 2026; Ma, Zhang, and Chen 2025; Zhang et al. 2025). For *ever-shifting* expressions, RepMD (Jiang

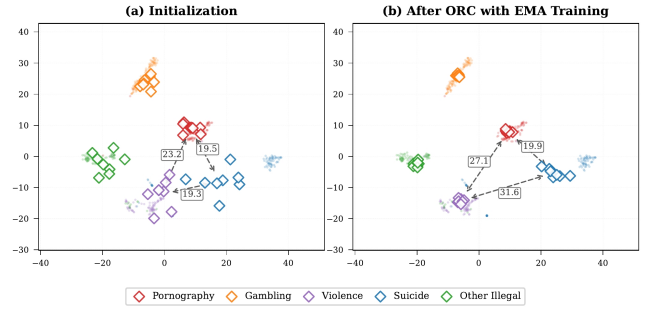


Figure 4: The t-SNE results’ distribution of dimensionality reduction across different types of harmful dialogues.

et al. 2026a) and JADE (Jiang et al. 2026b) address lexical evasion; BRACE shares the invariant-principles insight but embeds it as architectural regularization. Reasoning-enhanced safety (Li et al. 2025a,b; Wei et al. 2022) uses intermediate reasoning as inference; BRACE employs the chain as a **regularizer** with direct heads delivering primary predictions.

Prototype Learning and Expert Routing

Prototype-based methods represent classes through exemplar vectors (Snell, Swersky, and Zemel 2017; Khosla et al. 2020; Ho et al. 2024); BRACE’s prototypes produce augmented representations via cross-attention rather than classifying directly. Mixture-of-Experts (Wu et al. 2024; Goyal et al. 2025) scales capacity through conditional computation; BRACE organizes experts around harm categories, isolating binary detection from harm type features. For multi-task learning, BRACE routes fine-grained classification through f^{aug} and holistic judgments through x_{cls} , avoiding gradient competition from shared feature access.

Conclusion

In this paper, we propose BRACE, which encodes the ORC as four differentiable stages (Topic $\rightarrow$ Indicator $\rightarrow$ Severity $\rightarrow$ Type) with intermediate supervision, serving as a structured regularizer blended with direct heads, and supported by prototype-based feature augmentation and feature path disentanglement. Evaluation across over 20 dialogue safety benchmarks and 3 additional domain-specific sources shows that, with a RoBERTa-wwm-ext backbone, BRACE achieves a harm type macro F1 of **0.934** and severity accuracy of **0.818**. Binary detection performance, per-category diagnostics, and cross-backbone statistical tests are reported in Table 2 and the Technical Supplement. Decoder backbones further improve performance, reaching a harm type macro F1 of **0.949**. Ablation studies show that all components contribute to BRACE, and the structural decomposition of ORC enables BRACE to distinguish harmful types with semantic ambiguity,

References

Davidson, T.; Warmley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Of-

- fensive Language. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*, 512–515.
- Equitz, W. H. R.; and Cover, T. M. 1991. Successive Refinement of Information. *IEEE Transactions on Information Theory*, 37(2): 269–275.
- Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.
- Goyal, A.; Zhan, X.; Chen, Y.; Saha, K.; and Chandrasekharan, E. 2025. MoMoE: Mixture of Moderation Experts Framework for AI-Assisted Online Governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9726–9735.
- Ho, S.; Liu, M.; Du, L.; Gao, L.; and Long, G. 2024. Prototype-Guided Memory Replay for Continual Learning. *IEEE Transactions on Neural Networks and Learning Systems*.
- Huertas-García, Á.; Martín, A.; Huertas-Tato, J.; and Camacho, D. 2023. Countering Malicious Content Moderation Evasion in Online Social Networks: Simulation and Detection of Word Camouflage. *Applied Soft Computing*, 145: 110574.
- Ji, J.; Hong, D.; Zhang, B.; Chen, B.; Dai, J.; Zheng, B.; Qiu, T.; Zhou, J.; Wang, K.; Li, B.; Han, S.; Guo, Y.; and Yang, Y. 2025. PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 31983–32016.
- Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36. Datasets and Benchmarks Track.
- Jiang, Z.; Li, M.; Wang, J.; Huang, Y.; Huang, J.; Chang, Z.; Li, Z.; and Wang, Q. 2026a. All Changes May Have Invariant Principles: Improving Ever-Shifting Harmful Meme Detection via Design Concept Reproduction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 17595–17613.
- Jiang, Z.; Liu, M.; Qin, Y.; and Liu, B. 2026b. Breaking Free from Ivory Tower: Evaluating and Enhancing Real-world Chinese Underground Adversarial Jargon Detection. In *Proceedings of the 47th IEEE Symposium on Security and Privacy (S&P)*, 417–435.
- Kang, H.; Chen, J.; Li, Y.; Miao, X.; Xu, M.; Zhou, S.; Zhong, M.; Zhu, Y.; and Qian, T. 2025. Developing Continuous Toxicity Detection Against Increasing Types of Perturbed Toxic Text. *Computers & Security*.
- Khapre, S.; Mersha, M. A.; Shakil, H.; Baruah, J.; and Kalita, J. 2025. Toxicity in Online Platforms and AI Systems: A Survey of Needs, Challenges, Mitigations, and Future Directions. *Expert Systems with Applications*. ArXiv:2509.25539.
- Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, 18661–18673.
- Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, 2611–2624.
- Lee, A.; Bai, X.; Pres, I.; Wattenberg, M.; Kummerfeld, J. K.; and Mihalcea, R. 2024. A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. arXiv:2401.01967.
- Li, C.; Wang, J.; Pan, X.; Hong, G.; and Yang, M. 2025a. ReasoningShield: Safety Detection over Reasoning Traces of Large Reasoning Models. arXiv:2505.17244.
- Li, J.-J.; Pyatkin, V.; Kleiman-Weiner, M.; Jiang, L.; Dziri, N.; Collins, A.; Schaich Borg, J.; Sap, M.; Choi, Y.; and Levine, S. 2025b. SafetyAnalyst: Interpretable, Transparent, and Steerable Safety Moderation for AI Behavior. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 35731–35752.
- Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.
- Lu, J.; Xu, B.; Zhang, X.; Min, C.; Yang, L.; and Lin, H. 2023. Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 16235–16250.
- Ma, Y.; Zhang, N.; and Chen, H. 2025. ChineseHarmBench: A Chinese Harmful Content Detection Benchmark with Knowledge-Augmented Baselines. arXiv:2506.10960.
- Mármol-Romero, A. M.; Moreno-Muñoz, A.; Plaza-del Arco, F. M.; Molina-González, M. D.; Martín-Valdivia, M. T.; Ureña-López, L. A.; and Montejó-Ráez, A. 2024. MentalRiskES: A New Corpus for Early Detection of Mental Disorders in Spanish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 11204–11214.
- Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the Conference on Artificial Intelligence*, volume 35, 14867–14875.
- Mei, J.; Chen, J.; Yang, G.; Lin, W.; and Byrne, B. 2025. Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 23817–23839.

Qiu, H.; Ma, L.; and Lan, Z. 2024. PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4581–4607.

Röttger, P.; Seelawi, H.; Nozza, D.; Talat, Z.; and Vidgen, B. 2022. Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 154–169.

Röttger, P.; Vidgen, B.; Nguyen, D.; Waseem, Z.; Margetts, H.; and Pierrehumbert, J. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 41–58.

Saxe, A. M.; McClelland, J. L.; and Ganguli, S. 2014. Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks. In *International Conference on Learning Representations (ICLR)*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30.

Sun, H.; Zhang, Z.; Deng, J.; Cheng, J.; and Huang, M. 2023. Safety Assessment of Chinese Large Language Models. *arXiv preprint arXiv:2304.10436*.

Vidgen, B.; Thrush, T.; Waseem, Z.; and Kiela, D. 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 (ACL)*.

Wang, Y.; Zhang, S.; Zhou, Z.; Zhang, L.; Bai, J.; Lu, J.; Fan, B.; and Bu, F. 2026. DiaCOLQ: A Benchmark for Chinese Offensive Language Quadruple Prediction in Multi-Turn Dialogues. *Information Processing & Management*, 63(7): 104805.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Wu, H.; Qiu, Z.; Wang, Z.; Zhao, H.; and Fu, J. 2024. GW-MoE: Resolving Uncertainty in MoE Router with Global Workspace Theory. *arXiv:2406.12375*.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 1391–1399.

Zhang, H.; Gao, H.; Hu, Q.; Chen, G.; Yang, L.; Jing, B.; Wei, H.; Wang, B.; Bai, H.; and Yang, L. 2025. ChineseSafe: A Chinese Benchmark for Evaluating Safety in Large Language Models. *arXiv:2410.18491*.

Zhang, M.; Pan, X.; and Yang, M. 2023. JADE: A Linguistics-based Safety Evaluation Platform for Large Language Models. *arXiv:2311.00286*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E. P.; Gonzalez, J. E.; Stoica, I.; and Zhang, H. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. *arXiv:2309.11998*.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (No.2024YFF0618800), National Natural Science Foundation of China Grant No.62402484, No.62232016, Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant Number GZC20260867, 2026M791714, Youth Innovation Promotion Association Chinese Academy of Sciences, and Basic Research Program of ISCAS Grant No.ISCAS-JCZD-202405.