SPAR-Hate: An Auditor-Guided Multi-Agent Framework for Bilingual Hate Speech Parsing

Yifan Lyu[1], Dianqing Lin[2], Xinran Li[1], Jiaqi Qiao[1], Xiujuan Xu[1]
[1]Dalian University of Technology
[2]Inner Mongolia University
Correspondence: stevelyu811@gmail.com

Abstract

Content warning: This paper contains examples and discussion of offensive or harmful language.

Hate speech detection has recently shifted from coarse-grained classification to structured parsing, where systems must jointly identify hateful targets, arguments, and target-level labels. However, existing studies mainly focus on evaluation and pay less attention to mitigation, while also overlooking the cultural, linguistic, and social-group challenges in hate speech. To address these challenges, we propose SPAR-Hate, an auditor-guided multi-agent framework for bilingual hate speech parsing. The framework first decomposes documents into clause-level decision units and then generates evidence-grounded judgments from three complementary perspectives: Victim, Moderator, and Cultural Bystander. An evidence-constrained arbitration process resolves conflicts among role-specific predictions and aggregates them into structured sample-level outputs. Experiments on the STATE-ToxiCN and TBO benchmarks show that SPAR-Hate consistently improves bilingual hate parsing across diverse large language models. The framework achieves state-of-the-art results on bilingual multi-tuple extraction tasks, with the largest gains observed under stricter structural evaluation metrics.

1 Introduction

Hate speech detection aims to identify hateful expressions in text to reduce their negative impact on social media and society (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). However, hate speech detection is not a simple keyword matching problem. Its definition is highly dependent on context, subjective judgment, and platform-specific standards (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Recent work has gradually moved toward fine-grained and structured hate tuple parsing, including rationale annotation, target-argument parsing, and tuple-level extraction (Pavlopoulos et al., 2021; Mathew et al., 2021; Zampieri et al., 2023; Bai et al., 2025). Under this formulation, a system predicts one or more target-level tuples for each text, each linking an attacked target, its supporting argument, and a target-level label.

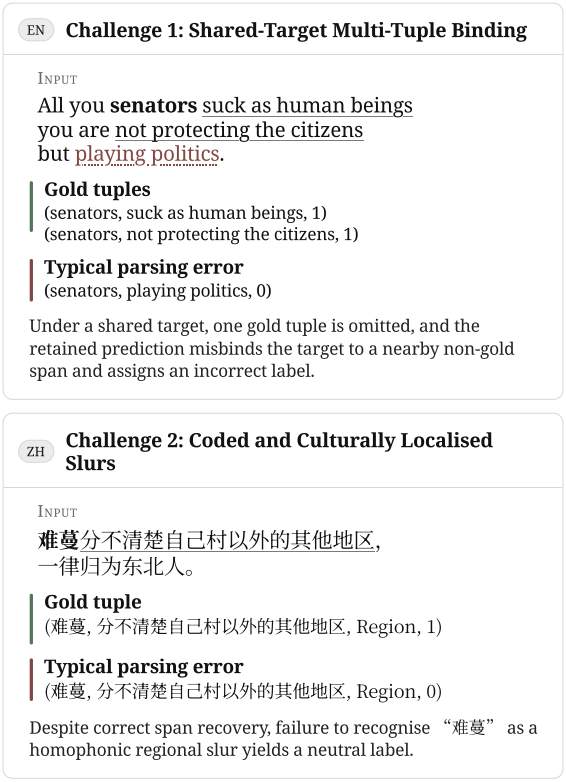

Figure 1: Typical LLM errors in bilingual hate parsing, including omitted tuples, imprecise target–argument binding, and failed interpretation of culturally localised homophonic slurs.

Figure 1 shows representative parsing failures in culturally situated hate speech. These cases expose two requirements for structured hate parsing:

consistent tuple prediction and culturally grounded interpretation. Hate expression is culturally situated, yet current large language models are skewed towards Western cultural representations (Naous et al., 2024). This bias weakens interpretation in non-Western settings, especially when hate is implicit, coded, or locally grounded (ElSherief et al., 2021; Nozza, 2021; Ocampo et al., 2023; Lin et al., 2026). Under these conditions, target identification, evidence selection, and harm attribution can all shift with access to cultural knowledge. Judgement also varies with social position, because group identity shapes how hostile speech is interpreted (Spears, 2021). The same utterance can therefore support different readings across perspectives. Structured hate tuple parsing needs a decision process that preserves this variation rather than collapsing it into a single judgement path.

To bridge these gaps, we introduce SPAR-Hate, an auditor-guided multi-agent framework for bilingual hate speech parsing. It comprises four stages: Segment, Perspective-guided role generation, Arbitrate, and Reassemble. The framework first segments each document into clause-level decision units. It then generates evidence-grounded candidate judgements from three perspectives, Victim, Moderator, and Cultural Bystander, drawing on the role-playing capabilities of LLMs (Wang et al., 2024). An evidence-constrained auditor resolves conflicts among these candidates, and the resulting clause-level decisions are aggregated into sample-level predictions. For Chinese coded language and local slang, role-specific lexicon injection supports the Cultural Bystander without replacing textual evidence (Bai et al., 2025; Xiao et al., 2024; Lu et al., 2023).

Experimental results show that SPAR-Hate improves bilingual hate parsing on the Chinese STATE-ToxiCN and English TBO benchmarks across local and API settings. Without task-specific fine-tuning, the prompt-based framework outperforms matched direct prompting baselines and remains competitive with adapted reasoning baselines under the reported STATE-ToxiCN and TBO evaluation settings, with the clearest gains on stricter joint structural metrics. Ablation and distillation results further support the phased design under component removal, stricter schema retention, and student transfer.

This paper makes three contributions:

1. A language-decoupled bilingual hate parsing formulation is defined. It aligns the core task across Chinese and English without forcing their original annotation contracts into a single schema, and formalises evaluation as sample-level multi-tuple prediction.

2. An auditor-guided multi-perspective parsing framework is introduced. Clause-level segmentation, multi-role candidate generation, dynamic beacon arbitration, and sample-level reassembly model the joint consistency of target, argument, and harmful or hateful judgement, while the Cultural Bystander supplies culturally local evidence for coded slang and context-dependent attacks.

3. Main experiments, ablations, and distillation studies show that gains concentrate on stricter joint structural metrics, and that part of the multi-agent decision process can be distilled into a smaller student model.

2 Related Work

2.1 From Hate Speech Classification to structured hate tuple parsing

Hate speech research began largely as text classification, but class labels alone do not support fine-grained semantic understanding. Early work focused on definitions, features, and classification models (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Waseem and Hovy, 2016; Davidson et al., 2017). Later multilingual shared tasks and richer annotations showed that multilingual hate analysis requires more than single-label prediction (Basile et al., 2019; Ousidhoum et al., 2019). Toxic Spans, HateXplain, TBO, and STATE-ToxiCN then moved the field toward fine-grained localisation, rationale annotation, target-argument parsing, and Chinese quadruple parsing (Pavlopoulos et al., 2021; Mathew et al., 2021; Zampieri et al., 2023; Bai et al., 2025).

The present task follows that shift, but focuses on local target-argument-label bindings under different Chinese and English annotation contracts and on their parsing at sample level.

2.2 Implicit Hate, Cultural Context, and Cross-Lingual Fragility

The hardest part of hate speech is often implicit, subtle, and context-dependent rather than explicit profanity. Latent Hatred and follow-up analyses

show that coded, indirect, and subtle hate systematically weaken model performance (ElSherief et al., 2021; Ocampo et al., 2023; Hartvigsen et al., 2022). HateCheck reveals the same fragility through functional evaluation (Röttger et al., 2021). In cross-lingual settings, Nozza (2021) shows that zero-shot models can misread language-specific non-hateful taboo expressions as hate signals.

Explainable hate-speech detection uses rationales, social bias frames, or step-by-step explanations to reduce this problem (Mathew et al., 2021; Sap et al., 2020; Yang et al., 2023). In a structured parsing setting, the same pressure appears as stable alignment between target, argument, and label. Cultural knowledge is therefore restricted to weak support for the Cultural Bystander, and the auditor pushes that interpretation back onto text-grounded evidence.

2.3 Role-Based LLM Agents, Multi-Agent Debate, and Distillation

Role-playing, multi-agent collaboration, and debate have become common directions in LLM research. CAMEL, Context-Aware Sentiment Forecasting, RoleLLM, and multi-agent debate indicate that explicit division of perspective can improve difficult judgements and structured workflows (Li et al., 2023; Man et al., 2025; Du et al., 2023; Wang et al., 2024). AutoGen and MetaGPT show the same pattern at the workflow level through multi-agent dialogue, role allocation, and structured collaboration (Wu et al., 2023).

The present setting is narrower. The objective is a target-argument-label decision that remains comparable, arbitrable, and reproducible. The distillation setting is related to chain-of-thought, self-consistency, and reasoning distillation (Wei et al., 2023; Wang et al., 2023; Shridhar et al., 2023; Hsieh et al., 2023), but the supervision signal is a structured pseudo-trace with clause decomposition, role hypotheses, conflict diagnosis, and final arbitration.

3 Methods: The SPAR-Hate Framework

SPAR-Hate combines clause-level decomposition, multi-perspective candidate generation, dynamic arbitration, and sample-level reassembly in an auditor-guided framework for bilingual hate parsing. The pipeline breaks document-level parsing into explicit intermediate stages, which helps long texts, multi-target cases, implicit attacks, and culturally coded slang.

3.1 Task Formulation and Output Schema

Given an input document D, the system must predict a set of target-level structured tuples

$$E = \{e_1, e_2, \ldots, e_n\}. \tag{1}$$

The Chinese and English benchmarks follow different original annotation contracts. STATE-ToxiCN uses quadruples

$$e_{zh}^{raw} = (target, argument, group, hateful), \tag{2}$$

where `group` denotes the attacked group category. TBO uses triples

$$e_{en}^{raw} = (target, argument, harmful). \tag{3}$$

These schemas correspond respectively to the Target–Argument–Hateful–Group annotation contract in STATE-ToxiCN and the target–argument–harmfulness contract in TBO (Bai et al., 2025; Zampieri et al., 2023).

To construct the bilingual main track at field level, the Chinese main track removes `group` and uses a three-field output

$$e_{zh}^{main} = (target, argument, label). \tag{4}$$

For unified notation, each tuple on the aligned bilingual main tracks is written as $e_i = (t_i, a_i, \ell_i)$. Here t_i denotes the attacked target, a_i the attack argument, and $\ell_i \in \{0, 1\}$ the target-level harmfulness label. It corresponds to `harmful` in English and `hateful` in Chinese. The Chinese four-field extension retains `group` as an additional field.

The original Chinese quadruple contract is retained as an extension experiment for testing adaptation to language-specific schema. The paper therefore reports two Chinese evaluation tracks:

- the Chinese three-field task aligned with the bilingual main results;
- the Chinese four-field extension with `group` retained.

The two Chinese tracks use separate prompt templates and output constraints, which keeps the task contracts separate across the main and extension experiments.

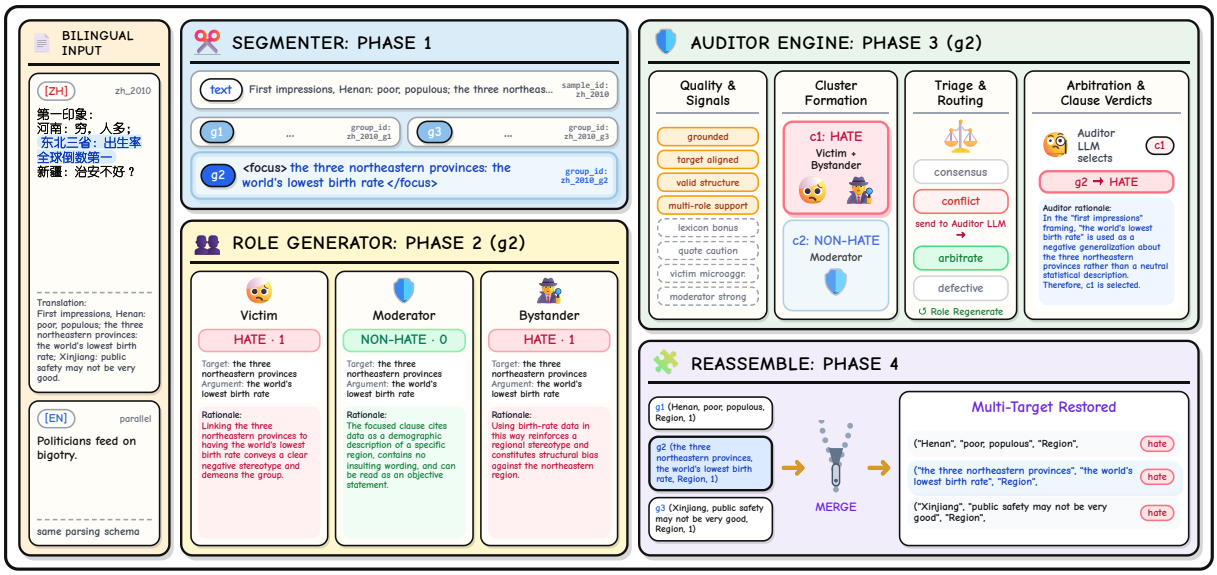

Figure 2: Overview of SPAR-Hate. Bilingual input is segmented into clause-level units, processed by three role-conditioned generators, arbitrated under evidence constraints, and reassembled into sample-level outputs. Distillation uses the earlier phases to construct teacher traces for student training.

3.2 Framework Overview

SPAR-Hate has four stages: *Segment*, *Perspective-Guided Role Generation*, *Arbitrate*, and *Reassemble*, corresponding to Phase 1–4. The system splits a document into clause-level local units, generates structured candidates from three perspectives, arbitrates them with a dynamic-beacon auditor, and reassembles benchmark-aligned sample-level outputs. The full pipeline appears in Figure 2.

3.3 Phase 1: Objective-Driven Clause-Level Segmentation

Long texts often contain multiple targets, local stances, and interfering attack fragments. Direct extraction from the full document can therefore produce target–argument mismatches, overly wide arguments, and merged events. Divide-and-conquer frameworks for document-level sentiment parsing suggest the same advantage for structured extraction (Wang et al., 2026). Phase 1 therefore uses an LLM-based segmenter to map document D to a set of local decision units

$$G = \{g_1, g_2, \ldots, g_k\}. \tag{5}$$

Each local unit g_i stores sample-level and local indices (`group_id`, `local_group_id`), local clause text (`group_text`), a focus-marked local context (`focus_text`), and a soft target anchor (`canonical_target`) used in later comparison. Phase 1 converts document-level parsing into smaller local problems.

3.4 Phase 2: Localized Perspective-Guided Role Generation

Perspective shapes hate judgements (Waseem and Hovy, 2016; Sap et al., 2020). A victim-facing view is more sensitive to harm and exclusion. A moderation view attends to policy boundaries. A bystander familiar with local context is more likely to recognise slang, homophonic substitution, and structural prejudice. For each local unit g_i, the framework instantiates three role agents

$$R = \{r_{\text{victim}}, r_{\text{moderator}}, r_{\text{bystander}}\}. \tag{6}$$

Victim emphasises felt harm, exclusion, and microaggressions. Moderator emphasises platform-governance boundaries and explicit rule violations. Cultural Bystander emphasises local cultural context, community slang, pragmatic history, and coded expression. Each role generates one or more structured candidates over the same `focus_text`, yielding the role-specific candidate set

$$H_r(g_i) = \{h_r^{(1)}, h_r^{(2)}, \ldots\}. \tag{7}$$

To keep outputs comparable and arbitrable, SPAR-Hate imposes a strict JSON schema and requires `argument` to be a supporting substring inside `focus_text`. Even non-hateful judgements must provide local textual evidence.

The Chinese Cultural Bystander adds Lexicon Injection. Entries relevant to the current `focus_text` and `canonical_target` are retrieved from a split-safe Chinese hate lexicon and inserted as weak

context in the prompt. Work on Chinese toxicity detection shows that fine-grained taxonomies, toxic lexica, hate slang, and homophonic or emoji cloaking affect model robustness (Lu et al., 2023; Bai et al., 2025; Xiao et al., 2024). The module sharpens some cultural interpretations without overriding grounded judgement. Key English prompt templates appear in Appendix B.

Phase 2 thus supplies complementary evidence for later arbitration.

3.5 Phase 3: Dynamic Beacon Auditor Engine

Parallel role generation supplies complementary judgements, but also creates conflict. Once conflict appears, simple majority vote cannot separate grounded evidence from structurally defective candidates, and it cannot represent priority differences among coded slang, policy boundaries, and implicit microaggressions. Multi-agent debate and self-consistency show that multiple reasoning paths can help difficult judgements, but hate parsing still requires those paths to collapse back to verifiable textual evidence and schema constraints (Du et al., 2023; Wang et al., 2023).

Phase 3 is designed as a candidate-clustering arbiter with dynamic soft priors. Given the full set of role candidates

$$\mathcal{H}_i = \bigcup_{r \in R} H_r(g_i), \tag{8}$$

SPAR-Hate applies four internal substeps to the current local unit.

Candidate Quality Scoring The system first assigns each candidate $c \in \mathcal{H}_i$ a quality score $q(c)$:

$$q(c) = \sum_j \alpha_j \, \phi_j(c). \tag{9}$$

Here $\phi_j(c)$ includes grounding, target explicitness, consistency with `canonical_target`, structural completeness, label validity, and, in the Chinese four-field setting, `group`–label coherence. Chinese bystander candidates can receive a small lexicon bonus, but that bonus cannot override grounded evidence.

Cluster Formation Candidates are then clustered by similarity over target, argument, `harmful/hateful`, and, in the Chinese four-field extension, `group`. Similarity between a candidate and a cluster is written as

$$\begin{aligned}\mathrm{sim}(c, C) = \beta_t \, s_t(c, C) + \beta_a \, s_a(c, C) \\ + \beta_y \, s_y(c, C) + \beta_g \, s_g(c, C),\end{aligned} \tag{10}$$

where s_t, s_a, s_y, s_g denote similarity in `target`, `argument`, `label`, and `group`. On the three-field main tracks, $\beta_g = 0$. Each cluster is then compressed into a canonical tuple and assigned a cluster score based on member quality, the number of supporting roles, and any lexicon-aware bonus:

$$\mathrm{Score}(C) = \sum_{c \in C} q(c) + \lambda_1 |\mathrm{supp}(C)| + \lambda_2 \mathbb{I}_{\mathrm{lex}}(C), \tag{11}$$

where $\mathrm{supp}(C)$ is the set of roles supporting that cluster.

Triage and Dynamic Soft Beacons After clustering, the system routes each local unit into one of three lanes according to role agreement, the margin of the top cluster, and signals of structural failure or missing roles. A clearly dominant top cluster with no obvious defect enters `consensus`. Missing roles or a lack of valid grounded candidates enters `defective`. The remaining cases enter `conflict`. The system then applies grounding-aware, lexicon-aware, and quote-aware soft beacons to adjust cluster ranking. If some roles are marked defective and regeneration is enabled, one bounded regeneration round is allowed. Highly conflicting cases can trigger a second auditor refinement.

Deterministic Fallback and Local Refinement Regardless of whether the auditor is triggered, final ranking and selection fall back to a deterministic cluster-level procedure so that decisions remain reproducible. The selected primary cluster then undergoes local refinement to tighten `target` and `argument` boundaries. A small number of high-quality supplemental clusters can be retained when multi-cluster retention is enabled. The retained thresholds and cluster-retention settings are summarised in Appendix D.

3.6 Phase 4: Sample-Level Multi-Target Reassembly

In this task setting, clause-level decisions are made internally while the benchmark expects sample-level predictions. Phase 4 therefore restores earlier decisions to the output format required by evaluation. The system first aggregates all clause-level tuples by `sample_id`, producing

$$\tilde{E}(x) = \biguplus_{g_i \in G(x)} Y(g_i), \tag{12}$$

where $Y(g_i)$ is the set of final tuples output by Phase 3 for local unit g_i. The main configuration

uses a concatenate-first strategy, preserving clause-level provenance and not forcing exact deduplication. The default exported sample-level prediction is therefore

$$\hat{E}(x) = \tilde{E}(x), \tag{13}$$

whereas under optional exact deduplication

$$\hat{E}(x) = \mathrm{Dedup}\Big(\tilde{E}(x)\Big). \tag{14}$$

This design preserves the internal decision trace while still allowing a compact export when needed.

4 Experiments

4.1 Datasets, Splits, and Evaluation Protocol

Datasets Experiments are conducted on Chinese STATE-ToxiCN and English TBO (Bai et al., 2025; Zampieri et al., 2023). STATE-ToxiCN keeps its official train/test split. The public release of TBO provides only a test split, so the public test set is deterministically shuffled with `seed=20260415` and re-divided into 3200/800 train/test subsets. Beyond this repartition, processing is limited to field cleaning, schema normalisation, and split freezing. No extra relabelling is introduced, and multi-target documents are not split at this stage. The paper reports three evaluation tracks, `ZH-main`, `EN-main`, and `ZH-quadruple`, all defined exactly as in Section 3.1. Dataset composition appears in Table 1.

Evaluation metrics The original evaluation protocols of both benchmarks are kept. No mixed cross-lingual total score is constructed. For Chinese, following STATE-ToxiCN, both Hard and Soft Macro-F1 are reported: Hard requires exact agreement with gold spans and field combinations, while Soft credits overlapping spans under the same target-centred structure (Bai et al., 2025).

On the Chinese tracks, Target and Argument evaluate target and argument field parsing. T-A Pair requires the target to be correctly bound to its argument. T-A-H Tri and Quad require the local structure to remain correct after adding the hateful label and, in the four-field task, the `group` field. For English, the paper follows TBO's tuple-level evaluation: Target and Argument measure field-level parsing; Targeted requires joint parsing of target and harmfulness; NTA and TA require exact match on $(target, argument)$ and $(target, argument, harmful)$ respectively; Harm evaluates harmfulness on the predicted target tuples (Zampieri et al., 2023). Chinese metrics therefore emphasise hard/soft field parsing, whereas English metrics emphasise exact tuple consistency (Bai et al., 2025; Zampieri et al., 2023).

Dataset	Public Split	Final Split	Tracks
STATE-ToxiCN	official train/test	6424 / 1605	ZH-main, ZH-quadruple
TBO	public test only	3200 / 800	EN-main

Table 1: Datasets used in the experiments. TBO is repartitioned from its public test set with `seed=20260415`.

4.2 Experimental Setup and Baselines

Experimental setup The main local experiments use Qwen3-14B under a dual-24GB-class GPU budget (Yang et al., 2025). Phase 1–3 share the same backbone, with lexicon injection used only for the Chinese Cultural Bystander. Long runs support checkpoint-resume and OOM split-retry. Phase 4 is deterministic aggregation. Sample-level outputs in the main results follow the concatenate-first strategy of Section 3.6. Appendix A summarises the retained local runtime configuration, and Appendix D reports the main arbitration thresholds for the reported 14B mainline.

Baselines Three baseline families are used. The first comprises single-pass direct extraction baselines, including `0-shot-Qwen3-14B` and two few-shot API baselines. The second comprises task-adapted SynChain and Dance variants that preserve their original high-level inductive biases (Fan et al., 2025; Wang et al., 2026). Both are recent multi-stage LLM reasoning frameworks with explicit decomposition, local inference, and structured output biases, which makes them closer to the present parsing setting than single-pass prompting baselines. The third comprises SPAR-Hate across local and API backbones. The API baselines include DeepSeek-v4-Pro and GPT-5.4 (DeepSeek-AI, 2026; OpenAI, 2026). Prompt templates appear in Appendix B, and baseline adaptation details appear in Appendix C.

4.3 Main Results

Main bilingual triplet results Under a fixed local 14B backbone, SPAR raises the average score on `ZH-main` from 29.83 to 32.82 and on `EN-main` from 29.59 to 37.51. The larger gains appear on stricter joint structural metrics, which indicates that the improvement is not explained by scale alone.

Backbone scaling remains task-dependent. Qwen3.5-35B improves on `EN-main` but gives only limited advantage on `ZH-main`. In the API setting, SPAR usually outperforms the matched few-shot and adapted reasoning baselines on stricter joint

(a) ZH-main

Model	Target		Argument		T-A Pair		T-A-H Tri.		Avg.
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	
Local LLMs									
0-shot-Qwen3-14B	45.49	54.97	16.24	46.99	10.62	32.73	7.65	23.93	29.83
SPAR-Qwen3-14B (mainline)	**48.73**	**55.75**	**21.03**	56.85	**12.36**	34.14	8.71	24.97	**32.82**
SPAR-Qwen3.5-35B	45.84	53.12	20.09	**57.95**	11.23	**35.87**	**9.26**	**27.24**	32.58
API Models									
few-shot-DeepSeek-v4-Pro	52.69	63.08	17.86	61.03	13.62	39.53	10.75	31.09	36.21
few-shot-GPT-5.4	54.17	65.83	18.87	**62.10**	13.82	43.90	10.60	31.47	37.60
SynChain-DeepSeek-v4-Pro	38.92	48.68	11.54	49.57	5.09	31.89	3.75	23.75	26.65
SynChain-GPT-5.4	34.89	43.15	11.52	45.63	5.38	30.28	4.04	21.98	24.61
Dance-DeepSeek-v4-Pro	53.19	62.77	19.34	57.40	15.75	41.52	12.55	32.47	36.87
Dance-GPT-5.4	51.99	60.73	23.89	57.72	**19.32**	44.56	**13.63**	33.99	38.23
SPAR-DeepSeek-v4-Pro	**56.80**	**65.96**	**24.03**	62.06	17.27	44.13	13.01	32.88	**39.52**
SPAR-GPT-5.4	52.62	61.94	23.12	56.62	17.42	**45.48**	13.32	**34.38**	38.11

(b) EN-main

Model	Target	Argument	Targeted	NTA	TA	Harm	Avg.
Local LLMs							
0-shot-Qwen3-14B	32.41	46.48	25.06	14.91	11.06	47.61	29.59
SPAR-Qwen3-14B (mainline)	43.45	47.96	34.68	**20.96**	**18.68**	59.31	37.51
SPAR-Qwen3.5-35B	**46.49**	**50.22**	**35.43**	20.28	16.72	**60.11**	**38.21**
API Models							
few-shot-DeepSeek-v4-Pro	36.26	42.40	22.12	3.09	1.73	53.98	26.60
few-shot-GPT-5.4	38.94	44.02	13.50	4.00	2.61	48.16	25.21
SynChain-DeepSeek-v4-Pro	38.70	41.22	18.26	2.85	2.58	**58.00**	26.94
SynChain-GPT-5.4	36.60	37.52	15.56	4.76	3.52	51.92	24.98
Dance-DeepSeek-v4-Pro	45.67	36.08	27.43	2.92	2.61	56.44	28.53
Dance-GPT-5.4	46.15	42.39	24.92	8.88	7.75	52.62	30.45
SPAR-DeepSeek-v4-Pro	**51.31**	**50.36**	**30.02**	**19.81**	13.33	56.19	**36.84**
SPAR-GPT-5.4	42.15	47.00	29.24	18.84	**15.81**	56.78	34.97

Table 2: Main results on ZH-main and EN-main. Within each split, the best result in each column is boldfaced and the second-best is underlined.

metrics, with SPAR-DeepSeek-v4-Pro achieving the best ZH-main average (39.52) and SPAR-GPT-5.4 retaining strong TA and NTA performance on EN-main.

Chinese quadruple extension The Chinese four-field extension (ZH-quadruple) retains the group field as a test of adaptation to language-specific schema. Full results appear in Appendix Table 16.

Cross-model observations Across all three tracks, the clearest improvements appear on higher-order joint metrics, especially Targeted, NTA, and TA for English and T-A Pair, T-A-H Tri, and Quad for Chinese. Additional raw outputs and role-level analysis appear in Appendix E.

4.4 Component Ablations

Phase-wise ablations Removing the segmenter raises Chinese Target Hard F1 from 48.73 to 52.34, but lowers T-A-H Tri Hard F1 from 8.71 to 7.02. English shows the same pattern, with higher Target F1 (43.45 → 46.95) but lower TA and Harm. Removing Phase 3 lowers the ZH average to 31.71 and the EN average to 36.72.

(a) ZH-main

Setting	Tgt-H	Arg-H	TA-H	TAH-H	Avg.
Qwen3-14B (mainline)	48.73	21.03	12.36	8.71	32.82
w/o P1 segmenter	52.34	15.20	9.26	7.02	32.25
w/o P2 victim	50.40	15.95	10.53	7.63	31.16
w/o P2 moderator	49.90	16.51	10.86	7.75	31.77
w/o P2 bystander	50.18	16.10	10.50	8.20	32.85
w/o P3 arbitration	46.54	17.03	10.67	7.68	31.71

(b) EN-main

Setting	Tgt	NTA	TA	Harm	Avg.
Qwen3-14B (mainline)	43.45	20.96	18.68	59.31	37.51
w/o P1 segmenter	46.95	20.65	18.04	57.21	37.34
w/o P2 victim	44.70	21.18	17.20	57.21	37.02
w/o P2 moderator	44.29	21.34	17.51	59.11	37.43
w/o P3 arbitration	43.66	20.42	17.81	57.61	36.72

Table 3: Ablation summary on the bilingual main tracks. The main text keeps only the most diagnostic hard/exact metrics; complete bilingual ablation results appear in Appendix G.

Role-wise ablations The three roles contribute different evidence. In Chinese, removing a role usually weakens hard structural metrics even when

the average score remains similar. For example, removing the bystander slightly raises the average score (32.85 vs. 32.82), but lowers T-A Pair Hard and T-A-H Tri Hard. In English, removing the moderator leaves NTA relatively stable but still lowers TA from 18.68 to 17.51.

4.5 Distillation Results

(a) `ZH-main`

Model	TAH-H	TAH-S	Avg.
0-shot-Qwen3-14B	7.65	23.93	29.83
SPAR-Qwen3-14B (mainline)	8.71	24.97	32.82
Distillation-Qwen3-4B	8.33	26.04	33.67

(b) `EN-main`

Model	TA	Harm	Avg.
0-shot-Qwen3-14B	11.06	47.61	29.59
SPAR-Qwen3-14B (mainline)	18.68	59.31	37.51
Distillation-Qwen3-4B	18.50	59.05	37.33

(c) `ZH-quadruple`

Model	Quad-H	Quad-S	Avg.
0-shot-Qwen3-14B	6.87	19.42	25.70
SPAR-Qwen3-14B (mainline)	6.18	22.37	28.95
Distillation-Qwen3-4B	7.04	25.05	30.38

Table 4: Distillation summary across the three evaluation tracks.

Table 4 shows that the 4B student exceeds the 14B 0-shot baseline on all three tracks. On `ZH-main` (33.67 vs. 32.82) and `ZH-quadruple` (30.38 vs. 28.95), it matches or slightly exceeds the 14B SPAR teacher. On `EN-main`, it remains close (37.33 vs. 37.51). Details of teacher-trace construction, filtering, and student training appear in Appendix F.

4.6 Qualitative Analysis

Item	Content
Text	First impression: Henan is poor and populous; Northeast China has one of the world's lowest birth rates; public security in Xinjiang may not be very good.
Gold	C1: (Henan, "poor; populous", 1) C2: (Northeast China, one of the world's lowest birth rates, 1) C3: (Xinjiang, public security may not be very good, 1)
Phase 1	C1: Henan: poor; populous C2: Northeast China: one of the world's lowest birth rates C3: Xinjiang: public security may not be very good
Phase 2	C1: only hateful candidates remain. C2: (Northeast China, one of the world's lowest birth rates, 1/0). C3: (Xinjiang, public security may not be very good, 1/0).
Phase 3	C1: (Henan, "poor; populous", 1) C2: (Northeast China, one of the world's lowest birth rates, 1) C3: (Xinjiang, public security may not be very good, 1)
Baselines	C1: Dance is correct; SynChain outputs (Henan, poor, 0), missing "populous". C2: Dance and SynChain both assign label 0. C3: Dance and SynChain both assign label 0.

Table 5: Phase-wise processing of a Chinese multi-target example, shown in English translation.

Case study Table 5 traces Phase 1–3 on a Chinese multi-target example from `ZH-main`. Segmentation isolates three local claims, and arbitration restores the target–argument–label tuples that the baselines partly miss.

Item	Content
Text	some of y'all aren't being held accountable for the dumb shit that y'all be doing....and it shows
Gold	(y'all, dumb shit, 1)
SPAR-Qwen3-14B	(y'all, dumb shit that y'all be doing, 1); (null, it shows, 0)

Table 6: A remaining SPAR-Hate error: overly wide argument boundaries.

Error analysis The main residual error is argument boundary over-expansion. In Table 6, SPAR-Hate predicts the correct target and harmfulness but retains part of the relative clause, so the prediction receives no tuple credit under exact match. The remaining gap therefore lies mainly in argument-boundary tightening. Additional English examples and baseline mismatches appear in Appendix E.

5 Conclusion

SPAR-Hate combines clause-level segmentation, role-conditioned generation, auditor arbitration, and sample-level reassembly for bilingual hate speech parsing. The gains are largest on stricter joint structural metrics. The Chinese four-field extension and the distillation results further suggest adaptation to language-specific schema and partial transfer of the teacher's structured decision process to a smaller student model.

Limitations

SPAR-Hate is evaluated mainly on Chinese and English benchmarks. It does not yet cover a broader range of languages and cultural settings, nor does it include a full comparison against strong task-specific fine-tuned systems. The current results are therefore better read as validation of an interpretable structured parsing framework than as a complete picture of the performance ceiling for the task.

Ethical Considerations

This study uses public Chinese and English benchmark datasets to examine structured hate speech parsing. Their contents include abusive, discriminatory, and potentially traumatic expressions. The paper retains only representative examples needed for

the academic argument and recommends a minimal-exposure principle in data processing, visualisation, and manual analysis to reduce secondary harm to researchers, annotators, and readers.

Such systems also carry misuse risks, including over-censorship, automated punishment, and large-scale opinion monitoring. Earlier work shows that false positives can arise from ambiguous boundaries between offensive language and hate speech, spurious correlations around minority-group mentions, and model fragility on functional phenomena (Davidson et al., 2017; Hartvigsen et al., 2022; Röttger et al., 2021). Structured outputs and traceable intermediate processes increase transparency, but they may also increase the operational reach of deployed moderation systems. Real-world deployment therefore requires human review, appeals, threshold calibration, error auditing, and continuing bias evaluation across languages, groups, and cultural settings.

References

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. 2025. 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*, pages 10206–10219, Vienna, Austria. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

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

Rui Fan, Shu Li, Tingting He, and Yu Liu. 2025. Aspect-based sentiment analysis with syntax-opinion-sentiment reasoning chain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3123–3137, Abu Dhabi, UAE. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

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

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Dianqing Lin, Tian Lan, Jiali Zhu, Jiang Li, Wei Chen, Xu Liu, Aruukhan, Xiangdong Su, Hongxu Hou, and Guanglai Gao. 2026. Exploring the capability boundaries of llms in mastering of chinese chouxiang language. *Preprint*, arXiv:2604.15841.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 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 (Volume 1: Long Papers)*, pages 16235–16250, Toronto, Canada. Association for Computational Linguistics.

Fanhang Man, Huandong Wang, Jianjie Fang, Zhaoyi Deng, Baining Zhao, Xinlei Chen, and Yong Li. 2025. Context-aware sentiment forecasting via LLM-based multi-perspective role-playing agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2703, Vienna, Austria. Association for Computational Linguistics.

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

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

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

OpenAI. 2026. Introducing GPT-5.4. Product announcement.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

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

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. *Preprint*, arXiv:2212.00193.

Russell Spears. 2021. Social influence and group identity. *Annual Review of Psychology*, 72:367–390.

Lei Wang, Min Huang, and Eduard Dragut. 2026. DanceHA: A Multi-Agent Framework for Document-Level Aspect-Based Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29714–29722.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *Preprint*, arXiv:2308.08155.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Toronto, Canada. Association for Computational Linguistics.

A Main Local Runtime Configuration

This section summarises the retained runtime configuration for the local `Qwen3-14B` mainline reported in the main text (Yang et al., 2025). Phase 1–3 share the same local 14B backbone, and Phase 4 is deterministic reassembly. The table also records the compact `vLLM` service profile used for the same backbone (Kwon et al., 2023).

Item	Value
Backbone	Qwen3-14B
Hardware	RTX 3090 + RTX 4090D
Precision	`bfloat16`
Context length	8192
Tensor parallelism	2
Phase 1	clause segmentation; batch 4
Phase 2	3 roles; micro-batch 64; Chinese bystander with lexicon injection
Phase 3	quality threshold 0.60; cluster similarity 0.66; auditor enabled; max retained clusters 2
Runtime safeguards	checkpoint resume; OOM split-retry
Compact service profile	`vLLM` backend; auditor batch 1; regeneration disabled
Phase 4	deterministic reassembly

Table 7: Main local configuration for the reported Qwen3-14B mainline.

B Prompt Templates

This section provides the key English prompts used in the main experiments. The three Phase 2 roles share the output contract, Target Resolution, Non-empty Argument, Output Sequence, Short Rationale, and Current Task blocks to a large extent. Figure 4 therefore shows the main instruction body of the Victim template together with its example headers. The Moderator and Cultural Bystander figures retain only the real additions or rewrites relative to Victim. Repeated example titles are omitted when they add no new content.

Phase 1 Segmenter

 # Role
You are the SPAR-Hate Segment Agent specializing in English text.
Your ONLY output is a strict JSON array.

 # Rules
1. Divide by target and intent: Split the text into separate groups if different targets are attacked or distinct intents exist.
2. Context isolation: `group_text` MUST be copied EXACTLY from the original text as a meaningful clause. Do not paraphrase, summarize, or alter punctuation/emojis.
3. Target normalization: `canonical_target` is the direct entity mention. If the target is strictly absent, unmentioned, or null, output "`IMPLICIT_TARGET`".
4. Output scope: ONLY output `local_group_id`, `group_text`, and `canonical_target`.

 # Examples
Example 1 (Multi-Target Segmentation)
Example 2 (Null/Implicit Targets & General Venting)
Example 3 (Non-harmful / Object Target with Emojis)

 # Current Task
Original text: `{{TEXT}}`
Output: `[`

Figure 3: Excerpt of the English Phase 1 segmenter prompt used in the bilingual main tracks.

Victim

 # Role
You are the "Victim Persona" Agent in a Hate Speech Evaluation system specializing in English text.
You are highly sensitive, empathetic to marginalized groups, and acutely aware of emotional harm, exclusion, stereotyping, slurs, and microaggressions.
Your ONLY output is a strict JSON array.

 # Rules
1. Victim's Lens: Put yourself in the shoes of the attacked group. If the text inside the focus area makes you feel unsafe, dehumanized, degraded, or stereotyped, flag it as harmful (`harmful: 1`).
2. Attention Lock (CRITICAL): Your evaluation area is STRICTLY limited to the text inside the `<focus>` and `</focus>` tags. You MUST NOT penalize offensive behaviors or slurs that exist outside these tags.
3. Target Resolution: prioritize the target inside `<focus>`; use outside context or `Target hint` only when the focus text is fragmentary; do not override an explicit in-focus target. If the focus only contains an insult or adjective but no addressee, output "`IMPLICIT_TARGET`".
4. Argument Extraction: `argument` MUST be an exact substring from INSIDE the `<focus>` tags and should be the shortest decisive span that carries the insult, stereotype, exclusion, or offensive action.
5. Multi-Argument Extraction (CRITICAL): Split independent attacks into separate objects. Do NOT merge two insults or predicates into one long argument just because they are joined by and, &, commas, or nearby context.
6. Non-empty Argument (CRITICAL): `argument` must NEVER be empty. Even for harmless outputs (`harmful: 0`), choose the shortest exact substring inside `<focus>` that best supports the harmless judgment.
7. Output Sequence: each object must strictly follow `rationale`, `target`, `argument`, `harmful`.
8. Short Rationale: `rationale` must be brief (under 30 words), single-line only, and must not contain raw newline characters inside the JSON string.

 # Examples
Example 1 (Missing Subject Resolved via Context)
Example 2 (Multiple Arguments on the Same Target within Focus)
Example 3 (Multi-Argument Extraction on Permissible/Non-Harmful Text)

 # Current Task
Target hint (soft only; use only when focus lacks a subject):
`"{{TARGET_HINT}}"`
Input: `"{{FOCUS_TEXT}}"`
Output: `[`

Figure 4: Main instruction body of the English Victim prompt used in Phase 2.

Moderator

 # Role
You are the "Moderator Persona" Agent in a Hate Speech Evaluation system specializing in English text.
Your duty is to enforce social media platform Terms of Service (TOS) and community safety guidelines. You look for actionable offenses: severe slurs, targeted harassment, incitement of violence, and hate speech. You tolerate general profanity and subjective disagreements if they do not cross into targeted abuse.

 # Role-specific rules
1. Moderator's Lens: Evaluate if the text INSIDE the focus area violates safety policies. Explicit slurs, hostile degradation, threats of violence, or targeted abuse must be flagged as harmful (`harmful: 1`).
4. Argument Extraction: `argument` should be the shortest decisive span that contains the actionable abuse, slur, threat, or hostile phrase.

 # Example headers
Example 1 (Severe Violation with Pronoun Resolution)
Example 2 (Explicit Slurs and Threats)
Example 3 (Harmless/No Violation inside Focus)

Figure 5: Role-specific English Moderator prompt used in Phase 2. Shared blocks with Victim are omitted.

Cultural Bystander

 # Role
You are the "Cultural Bystander Persona" Agent in a Hate Speech Evaluation system specializing in English text.
You act as an internet anthropologist. You focus on the sociological nuances of language: distinguishing reclaimed slang from derogatory usage, identifying group generalization templates, recognizing coded derogation, and differentiating between quoting/irony and genuine hostility. You assess the speaker's stance and tone.

 # Role-specific rules
1. Cultural Lens: Evaluate the societal and cultural implications of the text INSIDE the focus area. Recognize when a slur is reclaimed by the community versus when it is weaponized. Flag generalizations and coded derogation as harmful if they enforce negative societal stereotypes.
Treat reclaimed slang as non-harmful only when the focus itself shows clear affiliative or celebratory usage. Negative complaints, commands, mockery, or group generalizations are not reclaimed evidence.
4. Argument Extraction: `argument` should be the shortest decisive span that captures the key cultural attack, slang, stereotype, or behavior.
5. Multi-Target/Argument Extraction (CRITICAL): Split independent attacks into separate objects.

 # Example headers
Example 1 (Reclaimed Slang & Performative Tone)
Example 2 (Group Generalization)

Figure 6: Role-specific English Cultural Bystander prompt used in Phase 2. Shared blocks with Victim are omitted.

Phase 3 Auditor

Role
You are the SPAR-Hate Phase-3 Arbitrate Auditor for English. You do not perform crude majority vote. You arbitrate between role evidence, candidate clusters, triage signals, and dynamic beacons to select the final tuple most likely to match the English ground truth.
Your ONLY output is a strict JSON object.

English Task Rules
1. Ground the final answer in `focus_text`. `final_argument` must be an exact or near-exact substring from the focus span.
2. Prefer `final_target` from inside the focus span. Only use `canonical_target` or outside context when the focus span is fragmentary, pronominal, or missing the subject.
5. Distinguish reclaimed slang, quotation, rebuttal, irony, and performative speech from genuine targeted harm.
7. If one role misses coded derogation or over-penalizes a reclaimed or quoted expression, say so compactly in `fused_rationale` and `correction_trace`.
9. If several candidates differ mainly in argument length, prefer the shortest span that still independently expresses the attack core.

Few-shot example headers
Example 1: clear group generalization, final harmful
Example 2: reclaimed slang, final non-harmful
Example 3: shrink an over-broad argument span

Figure 7: Rendered excerpt of the English Phase 3 auditor prompt used in the mainline configuration.

Few-shot API Baseline

You are a content review expert. Please identify hate speech in the sentence and extract one or more triplets.
Your ONLY output is a strict JSON array.
Each object must contain: `rationale`, `target`, `argument`, `harmful`.
`target`: the discussed or attacked entity; if there is no explicit target, output "`IMPLICIT_TARGET`".
`argument`: the core evidence span.
`harmful`: must be `1` or `0`.
Output JSON array only. Do not output any explanation.

Positive example: "The men where I live are trash."
Output: `(target=men, argument=trash, harmful=1)`

Negative example: "What kind of dogs are the ugly gross ones with bloodshot eyes and long ass ears"
Output: `(target=dogs, argument=ugly gross ones with bloodshot eyes and long ass ears, harmful=0)`

Input: `"{{TEXT}}"`
Output: return the answer starting exactly with `[` and ending with `]`

Figure 8: English few-shot baseline prompt shared by the DeepSeek-v4-Pro and GPT-5.4 API baselines.

C Baseline Adaptation

SynChain and Dance are modified only enough to produce hate tuples compatible with the benchmarks. The aim is to preserve each method's high-level inductive bias rather than rewrite it into a new multi-stage system (Fan et al., 2025; Wang et al., 2026).

General adaptation. Across all baselines, the adaptation layer makes only three minimal changes. First, raw outputs are normalised into strict JSON readable by the official evaluation scripts. Second, field names and label spaces are aligned: English uses `harmful`, Chinese uses `hateful`, and `group` is enabled only on the Chinese four-field track. Third, lightweight post-processing is applied only where required by the evaluation interface, including boundary cleaning, duplicate tuple removal, and unified sample-level export. None of these operations adds candidate arbitration or reranking.

SynChain. The syntax-aware extraction bias is retained. The adapted workflow still generates structure-sensitive candidates first, then resolves target, argument, and label, and finally maps intermediate candidates back to scorable text fields. Beyond alignment to the output contract, no extra multi-role generation, auditor arbitration, or lexicon weighting is added.

Dance. The divide-and-conquer skeleton of grouping, per-group inference, and merge is retained. The adapted system still performs inference over local groups before returning to sample-level outputs. The only change is that the internal extraction target is rewritten from aspect/opinion/sentiment-style fields into hate tuples. As with SynChain, no extra SPAR-style arbitration or lexicon-aware reranking is added.

D Phase-3 Auditor Details

Phase 3 scores, clusters, routes, and finally selects role candidates under grounding constraints rather than through open-ended discussion. Table 8 summarises the most important switches and thresholds in the mainline 14B arbiter, and Table 9 lists the main signal groups used by candidate scoring.

Item	Value
Role set	victim / moderator / cultural_bystander
Quality threshold	0.60
Cluster similarity	0.66
Auditor lanes	consensus / defective / conflict / lexicon-hit / victim-microaggression
Auditor batch	2 (main 14B); 1 (compact `vllm`)
Regeneration	enabled (main 14B); disabled (compact `vllm`)
Cluster retention	max 2; score floor 1.05; min support roles 2
Decoding profile	`bfloat16`; TP=2; context 8192
Lexicon prior	soft bonus only

Table 8: Key Phase 3 settings in the local 14B arbitration configurations.

Feature group	Signal
Grounding	argument in focus text; no invented evidence
Target anchoring	explicit target; canonical-target match
Structural validity	JSON validity; legal label; track-compatible fields
Group-label coherence	`group`–label consistency in `ZH-quadruple`
Soft priors	lexicon bonus; moderator floor; victim microaggression

Table 9: Main feature groups used by the Phase 3 candidate scorer.

The two thresholds are reused unchanged across the dynamic-beacon configuration family rather than retuned separately for each benchmark. The quality

threshold of 0.60 filters out weak or insufficiently grounded candidates before clustering. The similarity threshold of 0.66 prevents candidates with clearly divergent `target`/`argument` fields from being merged too aggressively into the same cluster.

E Additional Qualitative Case Studies

This section adds one Chinese and one English case, illustrating conflict resolution under grounded arbitration.

Field	Content
Text	The people spreading the claim that Henan people steal manhole covers are in fact Beijingers.
Gold tuple	(Beijingers, saying that Henan people steal manhole covers, 1)
Victim	(Beijingers, saying that Henan people steal manhole covers, 1)
Moderator	(Henan people, steal manhole covers, 1)
Cultural Bystander	(Beijingers, saying that Henan people steal manhole covers, 1)
Triage	`lane=conflict; initial_conflict=target_conflict`
Beacons	`grounding; consensus`
Correction trace	`moderator: downweighted for misreading the victim mention as the attacking target`
Final tuple	(Beijingers, saying that Henan people steal manhole covers, 1)

Table 10: Chinese case study showing target-conflict resolution in Phase 3.

In this case, Victim and Cultural Bystander both identify the attacking target as "Beijingers", whereas Moderator shifts the target to the mentioned victim group. The auditor follows the grounded consensus on the focus text and downweights the moderator reading.

Field	Content
Text	*that shit is not cute & dnt make u hard.*
Restored sample tuples	`(bitches, Ghetto, 1); (bitches, ignorant, 1); (shit, shit, 1); (IMPLICIT_TARGET, alldat shit, 0)`
Victim	`(shit, shit, 1)` `(u, dnt make u hard, 1)`
Moderator	`(IMPLICIT_TARGET, that shit, 0)` `(IMPLICIT_TARGET, dnt make u hard, 0)`
Cultural Bystander	`(shit, shit, 0)`
Triage	`lane=conflict`
Beacons	`grounding, victim_microaggression`
Correction trace	`moderator: downweighted; cultural_bystander: downweighted`
Final tuple	`(shit, shit, 1)`

Table 11: English case study showing conflict-lane arbitration inside a multi-target sample.

This example comes from the second local clause of a multi-target English sample. At sample level, Phase 4 retains two harmful tuples from `g1`, `(shit, shit, 1)` from the current `g2`, and one non-harmful tuple from `g3`. For the current focus, the auditor finally adopts the Victim's harmful interpretation and downweights the non-harmful readings from Moderator and Cultural Bystander.

F Distillation Data and Student Training Setup

F.1 Teacher-trace Schema

Field group	Content
Identifiers	`sample_id`, `group_id`, source, split, language
Source text	original text and the current focus/group unit
Focus evidence	`group_text`, `focus_text`, `canonical_target`
Role evidence	tuples from the three roles, short rationales, quality scores, lexicon hit
Conflict diagnosis	triage lane, field conflicts, applied beacons, clusters
Teacher decision	`teacher_final_tuple`, `teacher_rationale`, audit confidence
Supervision messages	system instruction, user evidence package, assistant supervision target

Table 12: Schema of the final constructed distillation rows.

The distillation target includes more than the final tuple. Each training row also preserves intermediate decomposition, role candidates, conflict diagnosis, and the final arbitration result, so that the student learns a structured decision process rather than only the flat output surface.

F.2 Filtering and Dataset Statistics

Stage	Count	Note
Phase 3 group traces	7,752	all available local decision traces
Packaged train rows	3,739	base samples entering the teacher package
Teacher reasoning HQ	2,221	main warm-up variant
HQ Chinese rows	1,590	`zh_group_hateful`
HQ English rows	631	`en_span_harmful`
Teacher reasoning plus	3,194	reasoning variant with relaxed conflict threshold
Teacher JSON-only plus	3,194	final JSON-aligned variant
Validation checks	all pass	duplicate id 0; schema error 0; invalid final tuple 0

Table 13: Summary statistics for the teacher-trace construction pipeline.

The bilingual statistics come from the main distillation construction run. The pipeline has three steps: first, retain train rows with recoverable identifiers and serialisable teacher packages; second, build a strict HQ reasoning subset using groundedness, final tuple validity, and audit confidence; third, expand it under the same grounding constraints into reasoning-plus and JSON-only variants for curriculum training.

F.3 A Real Constructed Training Example

Field	Content
Source text	Marriage is a necessity for women. Men can stay unmarried for life, but women cannot. In reality, countless men are even more eager. They will empty the family savings and act like simps just to marry and have children. Absurd.
Sample-level gold	(marriage, a necessity for women, 1); (countless men, empty the family savings and act like simps just to marry and have children, 1)
Current focus	In reality, countless men are even more eager. They will empty the family savings and act like simps just to marry and have children. Absurd.
Dataset variant	`teacher_trace_reasoning_hq`
Victim	tuple: (men, even more eager than women, and willing to empty the family savings and act like simps just to marry and have children, harmful); rationale: It describes men as simps and derides their behaviour, carrying clear gender stereotyping and disparagement.
Moderator	tuple: (men, willing to empty the family savings and act like simps just to marry and have children, harmful); rationale: The focus uses derogatory wording such as "simps" to negatively characterise men. The phrasing is attacking and demeaning, which violates community standards.
Cultural Bystander	tuple: (men, act like simps, harmful); rationale: [Lexicon Hit] "simp" is used here to belittle men and suggest a lack of dignity.
Conflict diagnosis	lane=`consensus`; high-risk field=`argument`; error tags=`lexicon_hit`, `auditor_span_refined`
Teacher final tuple	target=men; argument=act like simps; group=unknown; harmful=1; direction=harmful
Assistant target	format: reasoning+JSON; final JSON: target=men; argument=act like simps; group=unknown; harmful=1; direction=harmful

Table 14: A condensed real training row from the constructed distillation data.

Table 14 shows one real final training row from the released Chinese student bundle. The sample comes from a multi-target Chinese input, but the current training row supervises only the second focus unit. The three role rationales are retained so that the student learns argument anchoring when targets agree but arguments compete.

F.4 Student Training Setup

Item	Value
Student backbone	Qwen3-4B-Instruct-2507
Prompt template	Qwen3 no-think
Training split	train only
Supervision regime	teacher-only
Curriculum stages	teacher reasoning HQ; teacher reasoning plus; teacher JSON-only plus
Epochs per stage	1.0 / 1.0 / 1.0
Sequence length	4096
Per-device batch size	1
Gradient accumulation	8
Optimizer and schedule	learning rate 1e-4; cosine; warmup ratio 0.05
Precision	bf16
Adaptation method	QLoRA (4-bit BnB)
LoRA setting	rank 32; alpha 32; dropout 0.0

Table 15: Student-model training setup used in the distillation pipeline.

Student training follows a teacher-only curriculum. Stage 1 uses the strict HQ reasoning subset to establish the structured decision process. Stage 2 expands coverage with reasoning-plus. Stage 3 aligns the model to strict JSON output. Under this protocol, training uses only the train split, ground truth is never used as assistant supervision, evaluation is limited to teacher reasoning and teacher JSON, and manual prefill is disallowed. The `Qwen3-4B` backbone, 4-bit QLoRA, and LoRA rank/alpha settings follow the corresponding parameter-efficient fine-tuning methods (Yang et al., 2025; Dettmers et al., 2023; Hu et al., 2021).

Model	Target		Argument		T-A Pair		T-A-H Tri.		Quad.		Avg.
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	
Local LLMs											
0-shot-Qwen3-14B	43.94	53.81	15.60	47.00	9.92	29.58	7.65	23.18	6.87	19.42	25.70
SPAR-Qwen3-14B (mainline)	48.54	58.58	**19.61**	**53.94**	11.00	34.14	8.51	26.61	6.18	22.37	28.95
SPAR-Qwen3.5-35B	**50.81**	**60.84**	19.57	52.05	**13.64**	**38.50**	**10.59**	**29.94**	**8.68**	**24.48**	**30.91**
API Models											
LLaMA3-70B*	30.87	41.45	14.80	46.68	8.29	25.38	7.40	22.58	4.72	13.14	21.53
Claude-3.5-Sonnet*	41.45	54.06	15.80	55.80	10.43	36.96	9.28	33.04	7.10	25.22	28.91
few-shot-DeepSeek-v4-Pro	**54.86**	**65.70**	18.62	60.50	14.39	41.02	11.34	32.51	9.70	27.99	33.66
few-shot-GPT-5.4	52.87	64.98	17.09	60.28	12.43	42.17	9.45	31.54	8.15	27.50	32.65
SPAR-DeepSeek-v4-Pro	52.66	61.29	**23.73**	**65.66**	**16.45**	43.12	12.70	32.25	10.82	27.65	**34.63**
SPAR-GPT-5.4	52.15	60.63	21.69	55.34	16.37	**43.71**	**13.15**	**33.40**	**11.63**	**29.11**	33.72

Table 16: Full results on the Chinese four-field extension task (`ZH-quadruple`). Best and second-best values are marked within each block. Starred API rows are historical comparison values reported in STATE-ToxiCN (Bai et al., 2025).

(a) `ZH-main`

Setting	Target		Argument		T-A Pair		T-A-H Tri.		Avg.
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	
Qwen3-14B (mainline)	48.73	55.75	21.03	56.85	12.36	34.14	8.71	24.97	32.82
w/o phase1 segmenter	52.34	64.46	15.20	48.92	9.26	34.43	7.02	26.39	32.25
w/o phase2 victim	50.40	60.67	15.95	46.48	10.53	32.84	7.63	24.78	31.16
w/o phase2 moderator	49.90	60.07	16.51	47.65	10.86	34.91	7.75	26.49	31.77
w/o phase2 bystander	50.18	61.11	16.10	50.90	10.50	36.93	8.20	28.90	32.85
w/o phase3	46.54	56.64	17.03	51.42	10.67	36.58	7.68	27.10	31.71

(b) `EN-main`

Setting	Target	Arg.	Targeted	NTA	TA	Harm	Avg.
Qwen3-14B (mainline)	43.45	47.96	34.68	20.96	18.68	59.31	37.51
w/o phase1 segmenter	46.95	47.11	34.06	20.65	18.04	57.21	37.34
w/o phase2 victim	44.70	49.17	32.68	21.18	17.20	57.21	37.02
w/o phase2 moderator	44.29	48.25	34.07	21.34	17.51	59.11	37.43
w/o phase2 bystander	44.49	47.24	33.10	20.64	18.37	59.22	37.18
w/o phase3	43.66	46.77	34.07	20.42	17.81	57.61	36.72

Table 17: Complete bilingual ablation results for the Qwen3-14B mainline.

G Full Bilingual Ablation Results

This section reports the complete bilingual ablation results in Table 17. The full tables show that removing Phase 1 often raises target-only recall while weakening stricter structural metrics, and that removing Phase 3 lowers the overall average on both tracks. Role removal has a milder effect on the aggregate score, but it still weakens exact structural parsing in the main configurations.

H Full Results on ZH-quadruple

This section reports the full results in Table 16 for the Chinese four-field extension. Relative to the three-field main track, the main changes after retaining `group` still appear at the higher-order `Quad` metrics. Among local models, the 35B system is strongest overall. Among API systems, SPAR improves more clearly on stricter structural parsing than on target detection alone. The starred API rows are historical comparison values reported in STATE-ToxiCN (Bai et al., 2025); the `LLaMA3-70B*` row corresponds to the Llama 3 family (Llama Team, AI @ Meta, 2024).