SPAR-Hate: Auditor-Guided Multi-Perspective Role Reasoning 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

Hate speech research has moved from coarse-grained classification towards structured parsing, where systems jointly identify targets, supporting arguments, and target-level labels. Documents with multiple targets, conflicting local readings, or culturally coded language make these bindings difficult to recover. SPAR-Hate is an auditor-guided multi-perspective role-reasoning framework for bilingual hate speech parsing. It decomposes each document into local focus units, elicits evidence-grounded candidates from Victim, Moderator, and Cultural Bystander perspectives, resolves candidate conflicts under grounding and schema constraints, and reassembles sample-level predictions. Experiments on STATE-ToxiCN and a controlled TBO split show gains across local and API backbones, concentrated on strict joint target–argument–label metrics. Full-test integrated-prompt controls, component ablations, and bounded-arbitration diagnostics identify the contribution of separated perspective generation and arbitration. Structured teacher traces also support training a smaller student model.

1 Introduction

Hate speech detection identifies hateful expressions whose interpretation depends on context, social judgement, and platform standards (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Fine-grained work now includes rationale annotation, target–argument parsing, and tuple extraction (Pavlopoulos et al., 2021; Mathew et al., 2021; Zampieri et al., 2023; Bai et al., 2025). Structured hate parsing predicts one or more target-level tuples, each linking an attacked target, its supporting argument, and a target-level label.

Figure 1 shows omitted tuples, faulty target–argument binding, and a missed homophonic slur. These failures require consistent tuple prediction and culturally grounded interpretation. Large language models remain skewed towards Western cultural representations (Naous et al., 2024), weakening performance on implicit, coded, and locally grounded hate (ElSherief et al., 2021; Nozza, 2021; Ocampo et al., 2023; Lin et al., 2026). Cultural knowledge can change target identification, evidence selection, and harm attribution. Social position also shapes interpretations of hostile speech (Spears, 2021), so the same utterance can support different grounded readings.

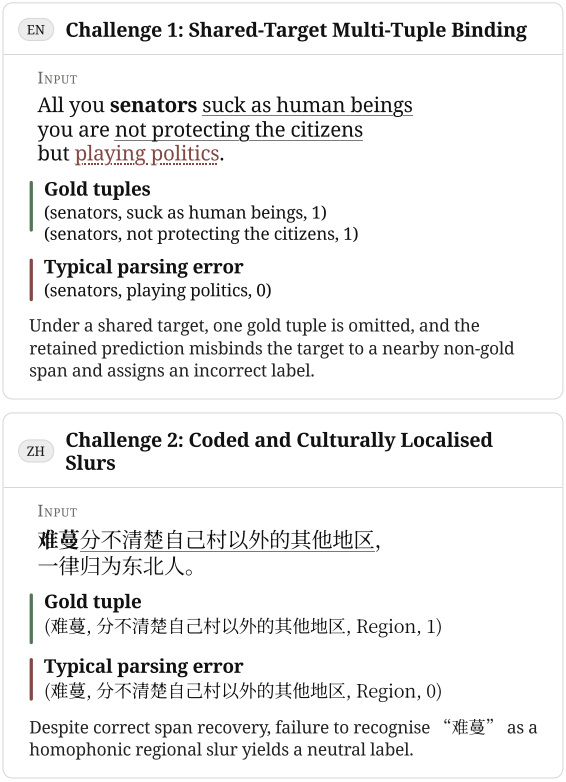

Figure 1: Typical LLM errors in bilingual hate parsing, including omitted tuples, imprecise target–argument binding, and failed interpretation of culturally localised homophonic slurs.

SPAR-Hate comprises four stages: Segment, Perspective-Guided Role Generation, Arbitrate, and Reassemble. Phase 1 produces local focus

units. Victim, Moderator, and Cultural Bystander generators produce evidence-grounded candidates for each unit. Phase 3 clusters and selects candidates under grounding and schema constraints, and Phase 4 restores sample-level predictions. The Chinese Cultural Bystander receives weak lexical context for coded language and local slang (Bai et al., 2025; Xiao et al., 2024; Lu et al., 2023).

Across local and API settings, SPAR-Hate improves the stricter joint structural metrics on Chinese STATE-ToxiCN and the controlled English TBO split. Component ablations test each phase, integrated-prompt controls test separated perspective generation and arbitration, and distillation transfers the structured traces to a smaller student model.

Contributions. The task formulation aligns Chinese and English structured parsing at field level while retaining their original annotation contracts. The framework combines local focus-unit segmentation, three perspective-conditioned generators, constrained arbitration, and sample-level reassembly. Main experiments, full-test integrated-prompt controls, ablations, and diagnostics test strict structural recovery and structured teacher-trace transfer.

2 Related Work

2.1 From Hate Speech Classification to Structured 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).

SPAR-Hate focuses on local target–argument–label bindings under separate Chinese and English annotation contracts and evaluates their sample-level reconstruction.

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

Implicit, subtle, and context-dependent hate remains difficult to detect. Coded and indirect expressions weaken model performance (ElSherief et al., 2021; Ocampo et al., 2023; Hartvigsen et al., 2022), while HateCheck identifies related functional failures (Röttger et al., 2021). Cross-lingual zero-shot models can also misread language-specific non-hateful taboo expressions as hate signals (Nozza, 2021).

Explainable hate-speech detection uses rationales, social bias frames, and stepwise explanations (Mathew et al., 2021; Sap et al., 2020; Yang et al., 2023). Structured parsing adds the requirement that target, argument, and label remain jointly aligned. SPAR-Hate supplies cultural knowledge as weak context and retains only text-grounded candidates.

2.3 Role-Conditioned Reasoning and Constrained Aggregation

RoleLLM and multi-perspective role-playing show that role conditioning elicits distinct knowledge and reasoning biases (Wang et al., 2024). Proposer–aggregator methods coordinate several model outputs (Du et al., 2023). AutoGen and MetaGPT organise role allocation and structured workflow handoffs, with MetaGPT encoding these handoffs through standard operating procedures (Wu et al., 2023; Hong et al., 2024).

SPAR-Hate constrains role outputs to comparable target–argument–label candidates before aggregation. Its distillation traces retain focus-unit decomposition, role hypotheses, conflict diagnosis, and final arbitration, linking the method 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).

3 Methods: The SPAR-Hate Framework

SPAR-Hate combines local focus-unit 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 Contracts

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 Chinese three-field task supports the bilingual main comparison. A four-field extension retains `group` and tests adaptation to the original language-specific schema. Separate prompt templates and output constraints preserve both contracts.

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 local focus units, generates structured candidates from three perspectives, arbitrates them with a dynamic-beacon procedure, and reassembles benchmark-aligned sample-level outputs. Figure 2 presents the full pipeline.

3.3 Phase 1: Local Focus-Unit 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 text (`group_text`), a focus-marked local context (`focus_text`), and a soft target anchor (`canonical_target`). A focus unit is defined around a target and local intent while retaining enough context for argument grounding. Its boundary need not coincide with a syntactic boundary.

3.4 Phase 2: Perspective-Guided Role Generation

Perspective shapes hate judgements (Waseem and Hovy, 2016; Sap et al., 2020). The task-motivated role set covers affected-group harm, platform-governance boundaries, and community interpretation of implicit or culturally coded hate. For each local unit g_i, the framework instantiates three role-conditioned generators

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

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

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

SPAR-Hate imposes a strict JSON schema and requires `argument` to be a supporting substring inside `focus_text`. Every judgement includes local textual evidence.

The Chinese Cultural Bystander receives entries from the auxiliary STATE-ToxiCN lexicon. Its

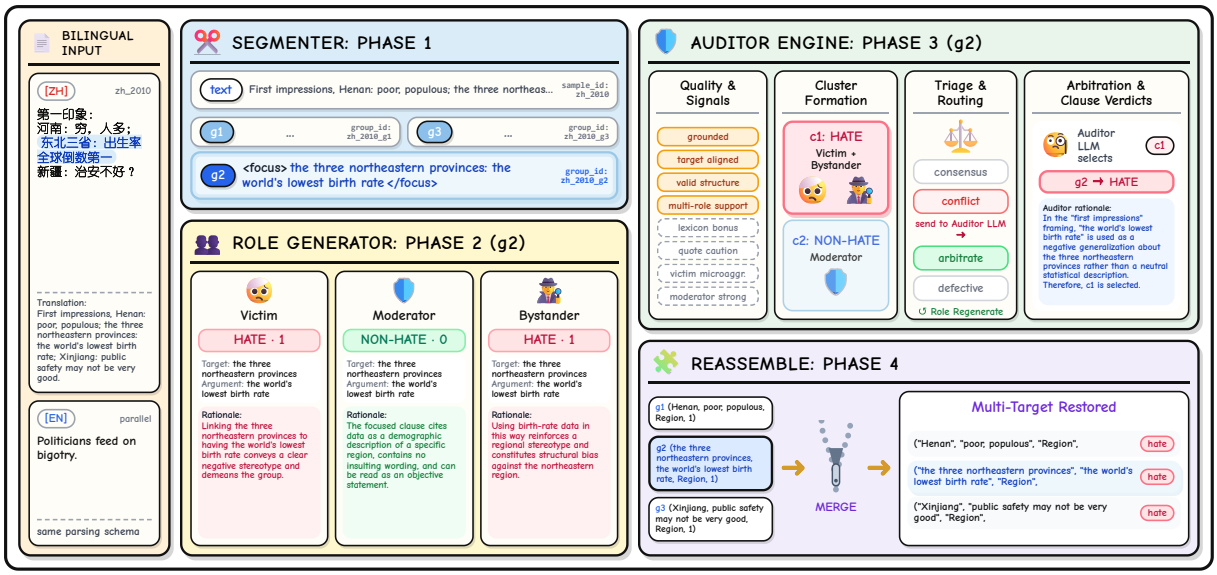

Figure 2: Overview of SPAR-Hate. Bilingual input is segmented into local focus units, processed by three role-conditioned generators, arbitrated under evidence constraints, and reassembled into sample-level outputs. Distillation uses the earlier phases to construct teacher traces for student training.

829 `term/category/definition` records contain no sample identifiers, gold tuples, or instance labels. Retrieved entries provide weak context for the current `focus_text` and `canonical_target`. Candidates remain subject to focus-span grounding, and lexicon-only singleton candidates are discarded. Lexicon hits occur in 10.1% of `ZH-main` groups and 9.6% of `ZH-quadruple` groups. Chinese toxicity research motivates this support for slang, homophones, and emoji cloaking (Lu et al., 2023; Bai et al., 2025; Xiao et al., 2024). Appendix B gives the prompt templates.

3.5 Phase 3: Local Candidate Clustering and Auditor Arbitration

Phase 3 filters, clusters, and ranks the role candidates for each local unit. Given

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

the procedure applies quality scoring, cluster formation, triage, and deterministic selection.

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. Candidates below quality 0.60 are filtered before clustering. An ungrounded or missing argument fails validation.

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, $\beta_g = 0$. The reported field weights are $0.45/0.45/0.05/0.05$, respectively. 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. The mainline support-role and lexicon-cluster bonuses are 0.12 and 0.08; the candidate-level lexicon bonus is 0.05.

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 Final ranking uses a deterministic cluster-level fallback. The selected primary cluster undergoes local refinement of `target` and `argument` boundaries. The reported configuration retains at most two clusters, requires score ≥ 1.05 and at least two supporting roles, and allows one regeneration round for a defective role. Across retained runs, 55.9–61.7% of focus units enter the conflict lane and 65.1–70.7% invoke auditor refinement. Appendix D reports the complete scoring, routing, and boundedness diagnostics.

3.6 Phase 4: Sample-Level Multi-Target Reassembly

In this task setting, local focus-unit 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 local 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 local 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

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

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 Experiments

4.1 Datasets 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. TBO numbers are controlled within-paper comparisons on this deterministic split only; they are not directly comparable to studies evaluated on the original public test-only release.

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 met-

API zero-shot model	EN	ZH
DeepSeek-v4-Pro	23.18	22.80
GPT-5.4	21.01	28.13

Table 2: Full-test average scores for the added API zero-shot baselines.

rics therefore emphasise hard/soft field parsing, whereas English metrics emphasise exact tuple consistency (Bai et al., 2025; Zampieri et al., 2023).

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 follow the concatenate-first strategy of Section 3.6. Appendix A gives the runtime configuration and artifact map; Appendix D gives the arbitration settings.

Baselines The baselines cover direct extraction, task-adapted SynChain and Dance, and SPAR with local and API backbones (Fan et al., 2025; Wang et al., 2026; DeepSeek-AI, 2026; OpenAI, 2026). Table 2 reports the added full-test API zero-shot averages. Prompt templates and baseline adaptations appear in Appendices B and C.

4.3 Main Results

Table 3 reports the complete bilingual main-track comparison.

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 occur on stricter joint structural metrics.

Qwen3.5-35B records 38.21 on `EN-main` and 32.58 on `ZH-main`. Among API systems, `SPAR-DeepSeek-v4-Pro` records the highest `ZH-main` average (39.52), while `SPAR-GPT-5.4` records EN NTA/TA scores of 18.84/15.81.

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 25.

Backbone	Track	Integrated	SPAR
DeepSeek-v4-Pro	EN NTA / TA	4.87 / 3.89	19.81 / 13.33
GPT-5.4	EN NTA / TA	11.41 / 8.18	18.84 / 15.81
DeepSeek-v4-Pro	ZH TA / TAH-H	6.31 / 4.40	17.27 / 13.01
GPT-5.4	ZH TA / TAH-H	14.08 / 10.56	17.42 / 13.32

Table 4: Full-test fixed-Phase-1 integrated-prompt controls. The integrated condition retains the three lens descriptions (and Chinese lexicon access) but removes separated candidates, clustering, and auditor arbitration.

(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 5: 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 H.

Cross-model observations The largest differences occur on Targeted, NTA, and TA for English and on T-A Pair, T-A-H Tri, and Quad for Chinese. Appendix F provides additional outputs and cases.

4.4 Framework Analysis

Integrated-prompt controls Four full-test integrated-prompt controls fix the Phase 1 focus units and place all three perspective descriptions in one call. The Chinese controls retain the same lexicon access. Table 4 reports lower strict tuple-binding scores after removing separated role outputs, candidate clustering, and auditor arbitration. GPT-5.4 Target scores are 46.01 for integrated prompting and 42.15 for SPAR on EN, with corresponding ZH scores of 54.40 and 52.62.

Component ablations Table 5 summarises the most diagnostic hard and exact metrics; complete results are deferred to Appendix H.

(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 3: 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.

Phase-wise ablations Removing the segmenter raises Chinese Target Hard F1 from 48.73 to 52.34 and lowers T-A-H Tri Hard F1 from 8.71 to 7.02. On English, Target F1 rises from 43.45 to 46.95, while TA and Harm fall. Removing Phase 3 lowers the ZH and EN averages to 31.71 and 36.72.

Role-wise ablations Removing the bystander changes the Chinese average from 32.82 to 32.85 and lowers T-A Pair Hard and T-A-H Tri Hard. Removing the moderator lowers English TA from 18.68 to 17.51, while NTA changes from 20.96 to 21.34.

Diagnostic validation Supplementary exact-set and blinded semantic diagnostics appear in Appendix E. They are reported as secondary checks, not primary ranking metrics. The attacked-group breakdown in the same appendix locates a remaining multi-group binding error.

4.5 Distillation Results

Table 6 compares the distilled 4B student with the zero-shot and SPAR-Qwen3-14B systems on all three evaluation tracks.

(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 6: Distillation summary across the three evaluation tracks.

The distilled student improves over SPAR-Qwen3-14B on both Chinese tracks and remains close to its teacher on `EN-main`, despite using a substantially smaller backbone. Appendix G reports teacher-trace construction, filtering, and student training.

4.6 Qualitative Analysis

Case study Table 7 traces three Phase 1 local focus units from a Chinese multi-target example. Phase 2 produces conflicting labels for g_2 and g_3, and Phase 3 selects the gold label for both units. For g_1, SPAR retains the full argument "poor; populous", which SynChain truncates.

Error analysis The main residual error is argument boundary over-expansion. In Table 8, SPAR-Hate predicts the correct target and harmfulness but retains additional contextual material, so the prediction receives no tuple credit under exact match. Appendix F provides further cases.

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

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	g_1: (Henan, "poor; populous", 1) g_2: (Northeast China, one of the world's lowest birth rates, 1) g_3: (Xinjiang, public security may not be very good, 1)
Phase 1	g_1: Henan: poor; populous g_2: Northeast China: one of the world's lowest birth rates g_3: Xinjiang: public security may not be very good
Phase 2	g_1: only hateful candidates remain. g_2: (Northeast China, one of the world's lowest birth rates, 1/0). g_3: (Xinjiang, public security may not be very good, 1/0).
Phase 3	g_1: (Henan, "poor; populous", 1) g_2: (Northeast China, one of the world's lowest birth rates, 1) g_3: (Xinjiang, public security may not be very good, 1)
Baselines	g_1: Dance is correct. SynChain outputs (Henan, poor, 0), missing "populous". g_2: Dance and SynChain both assign label 0. g_3: Dance and SynChain both assign label 0.

Table 7: Phase-wise processing of three local focus units. g_1, g_2, and g_3 denote Phase 1 units. One unit may yield several sample-level tuples after Phase 4. The Chinese input is shown in English translation.

5 Conclusion

SPAR-Hate combines local focus-unit segmentation, role-conditioned generation, constrained auditor arbitration, and sample-level reassembly. Its largest gains occur on stricter joint structural metrics under the controlled benchmark settings. Integrated-prompt controls record lower tuple-binding scores after collapsing perspective generation and arbitration into one call. The Chinese four-field extension and distillation results cover language-specific schema and structured teacher traces.

Limitations

Evaluation covers Chinese and English and omits a full comparison with task-specific fine-tuned systems. The task-motivated role set has not been exhaustively searched, and the lexicon has no complete removal ablation. Weak lexical context can still over-flag dialectal, culturally loaded, or reclaimed terms. The threshold replay, blind audit, and attacked-group breakdown have limited scope and cannot support broad claims about statistical optimality, fairness, or cultural validity. TBO results use the deterministic internal split and do not support direct comparison with results on the original public test-only release.

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 moder-

ation 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.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Yuheng Xiawu, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

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.

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.

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.

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 Implementation and Artifact Details

Table 9 gives the retained `Qwen3-14B` configuration (Yang et al., 2025). Phase 1–3 share the same local backbone, and Phase 4 is deterministic. The compact profile uses `vLLM` (Kwon et al., 2023).

Item	Value
Backbone	Qwen3-14B
Hardware	RTX 3090 + RTX 4090D
Precision	`bfloat16`
Context length	8192
Tensor parallelism	2
Phase 1	local focus-unit 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 9: Main local configuration for the reported Qwen3-14B mainline.

A.1 Runtime and Call Diagnostics

Retained EN, ZH-main, and ZH-quadruple runs take 499.0, 617.6, and 623.3 minutes. Phase 2 makes 4,524, 6,978, and 6,933 role calls at 0.25–0.28 role-groups per second. Auditor refinement is triggered for 982, 1,645, and 1,610 groups. These figures describe retained runs; incomplete billing logs preclude a monetary cost estimate.

A.2 Artifact Map and Recoverability

The artifact roots `SPAR/`, `SPAR_anonymous_submission/`, and `transfer_bundles/` contain the deterministic TBO split and Phase 0 preparation path, Phase 1–3 prompts and configurations, Phase 3 scoring and fallback code, direct and API baseline outputs, SynChain and Dance adaptations, diagnostic summaries, and teacher/student exports. Historical bilingual distillation records are partly archival. Preserved reports and official evaluation exports support the reported construction and evaluation statistics, while the Chinese student bundle retains the independently recoverable training example.

B Prompt Templates

The three Phase 2 roles share the output contract, Target Resolution, Non-empty Argument, Output Sequence, Short Rationale, and Current Task blocks. Figure 4 gives the Victim instruction body and example headers. The Moderator and Cultural Bystander figures retain their role-specific additions. Repeated example titles are omitted.

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 local context. 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.

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

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, refutation, 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 8: Rendered excerpt of the English Phase 3 auditor prompt used in the mainline configuration.

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 Auditor Scoring and Bounded Arbitration

Phase 3 scores, clusters, routes, and selects role candidates under grounding constraints. Table 10 gives the main settings, and Table 11 gives the scoring signals.

Item	Value
Role set	victim / moderator / cultural_bystander
Quality threshold	0.60
Cluster similarity	0.66
Similarity weights	target/argument/label/group = 0.45/0.45/0.05/0.05
Score bonuses	support role 0.12; candidate lexicon 0.05; cluster lexicon 0.08
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
Final selection	deterministic fallback after optional refinement

Table 10: 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 11: Main feature groups used by the Phase 3 candidate scorer.

The quality threshold of 0.60 filters weak or ungrounded candidates before clustering. The similarity threshold of 0.66 limits merges between candidates with divergent `target` or `argument` fields. Both values are reused across the reported languages and benchmarks.

Step	Operation	Bound
Validate	Check JSON, label, argument grounding, and target anchor	quality ≥ 0.60
Cluster	Compare target, argument, label, and optional group	similarity ≥ 0.66
Route	Assign consensus, defective, or conflict lane	one local unit
Refine	Apply soft beacons and optional auditor call	one regeneration
Select	Rank with deterministic fallback	at most two clusters

Table 12: Operational summary of local Phase 3 arbitration.

Measure	Retained-run summary
Raw-cluster p95	EN 3; ZH 4
Retained/final p95 and maximum	2
Conflict-lane share	55.9–61.7%
Auditor-use share	65.1–70.7%

Table 13: Boundedness and routing diagnostics across the retained runs.

E Robustness and Semantic Validation

E.1 Threshold Sensitivity

The actual auditor is replayed on fixed seed-10947 subsets of 200 complete samples per language, covering 385 EN and 281 ZH focus groups per setting. Quality varies over $0.55/0.60/0.65$ at similarity 0.66. Similarity varies over $0.60/0.66/0.72$ at quality 0.60. All other settings remain fixed. Table 14 reports the retained summary.

Language	Observed range	Default
EN	34.97–35.08	35.08
ZH	26.28–27.12	26.82

Table 14: Four-hard-metric averages in the fixed-subset threshold replay. The results are not full-test significance estimates.

E.2 Blind Semantic Audit

Two independent bilingual or native-speaker reviewers assess 60 stratified Chinese focus units with gold labels hidden. Separate final-tuple and Cultural-Bystander-rationale rubrics yield 120 judgements. Table 15 gives acceptance and agreement. Consensus cases reach 70% strict overall acceptance; disagreement-heavy lexicon–auditor cases receive lower acceptance.

E.3 Strict Sample Exact-Set Summary

E.4 Attacked-Group Breakdown

Table 17 reports ZH-quadruple scores by gold attacked group. The Racism+Sexism group has the largest TAH-to-Quad drop, locating the main loss in multi-group binding. The breakdown measures robustness across the annotated groups; fairness requires separate evidence.

Object	R1	R2	Agree.	κ
Final tuple	73.3	71.7	81.7	0.540
Bystander rationale	68.3	58.3	86.7	0.716

Table 15: Acceptance rates, percentage agreement, and Cohen's κ in the blind semantic audit.

Track	SPAR-DeepSeek	Direct references
EN	20/800 (2.50%)	0–9/800 (0.00–1.13%)
ZH-main	141/1605 (8.79%)	zero-shot: 37/1605; 66/1605

Table 16: Sample exact-set summary. Some ZH few-shot baselines exceed the SPAR-DeepSeek value.

Gold group	Samples	TAH-S	Quad-S
Sexism	302	36.66	34.49
Racism	223	40.56	35.57
Region	216	27.20	24.93
LGBTQ	93	34.75	32.13
Racism + Sexism	85	21.85	5.30

Table 17: ZH-quadruple attacked-group diagnostic. Scores are Soft F1 and groups are defined from gold labels.

F Qualitative Cases and Error Analysis

The Chinese case isolates a target conflict. The English case traces a conflict-lane decision within a multi-target sample.

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.

The current text is focus unit g_2 in a sample containing g_1, g_2, and g_3. Phase 4 retains two harmful tuples from g_1, `(shit, shit, 1)` from g_2, and one non-harmful tuple from g_3. One focus unit can emit several tuples. For g_2, the auditor selects the Victim reading and downweights the non-harmful Moderator and Cultural Bystander candidates.

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

Field	Content
Text	*that shit is not cute & dnt make u hard.*
Phase 4 restored sample prediction	`(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 19: English case study showing conflict-lane arbitration inside a multi-target sample.

G Distillation Data and Student Training

G.1 Teacher-trace Schema

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.

G.2 Filtering and Dataset Statistics

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.

G.3 A Real Constructed Training Example

Table 22 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.

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 20: Schema of the final constructed distillation rows.

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 21: Summary statistics for the teacher-trace construction pipeline.

G.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

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

Table 23: 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).

H Full Bilingual Ablation Results

Table 24 gives the complete bilingual ablations underlying the compact analysis in Section 4.4.

(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 24: Complete bilingual Qwen3-14B ablations.

I Full Results on ZH-quadruple

Table 25 reports the full four-field results used for the Chinese extension discussed in Section 4.3.

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 25: Full Chinese four-field results. Starred rows are historical STATE-ToxiCN values (Bai et al., 2025).