CHINESE COMPETITIVE DEBATE: A DATASET AND BENCHMARK WITH PROFESSIONAL ADJUDICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating debate requires interpreting contributions in context and relating them to local assessments and overall outcomes. We introduce the Chinese Competitive Debate Dataset (CCDD), linking competition transcripts to original adjudication under a predefined rubric. CCDD contains 148 matches, 2,690 stage–side scoring instances, and 20,542 turns, with reviewed transcripts, aligned individual judge scores, match decisions, best-debater ballots, and commentary. Linked appeal records preserve reassessments. Three benchmark tasks—winner prediction, stage-score prediction, and best-debater prediction—demonstrate the use of this supervision. Zero-shot results reveal challenges in reproducing stage assessments, while judge-agreement analyses and structural references contextualize performance. CCDD supports research on contextual debate evaluation and judge disagreement. The dataset is publicly available at https://anonymous.4open.science/w/CCDD-Page.

1 INTRODUCTION

Evaluating a debate requires assessing how participants develop arguments and respond to one another over the course of an interaction. An answer may be fluent and relevant to the motion while leaving the opponent's central objection unanswered. Its quality therefore depends on the preceding challenge and the responding speaker's obligations, as well as the content of the answer itself. Automated debate adjudication must account for these contextual relationships when assessing particular contributions and forming an overall judgment.

Studying such adjudication requires records that connect debate text with the stage, side, and role being evaluated and the corresponding expert assessment. A final match outcome records the panel's decision but does not reveal how judges assessed particular stages. This distinction is especially relevant in Chinese competitive debate formats containing questioning, cross-examination, and free debate, where both sides contribute within a shared stage under different interaction constraints. Constructing useful supervision therefore requires preserving the interaction sequence, identifying the scope of each assessment, and documenting the rubric under which judgments were made.

Existing research provides important foundations for this goal. PanelBench, introduced with Debatrix, DebateBench, and Debatable Intelligence support the evaluation of automated judgments of debates and individual speeches (Liang et al., 2024; Tiwari et al., 2025; Sternlicht et al., 2025). Chinese debate resources support stance detection, summarization, and computational argumentation (Zhao et al., 2023; Lan et al., 2026), while related work studies argument-quality assessment and fine-grained creativity scoring (Chen et al., 2026b; Yu et al., 2026). These resources address complementary aspects of debate understanding, with different evaluation units and sources of supervision. Access remains limited, however, to interactive Chinese debate records that jointly preserve role-specific stage assessments, individual judges' decisions, and original commentary under a predefined competition rubric. This combination is needed to study local assessments in their interaction context while retaining their relationship to overall adjudication.

We introduce the **Chinese Competitive Debate Dataset (CCDD)** to provide this linked supervision. A shared scoring rubric was developed with senior debate experts before the competitions. Each of the 182 source matches was independently assessed by three judges selected from a pool of 157 qualified adjudicators. The records preserve stage scores, match decisions, best-debater ballots, and commentary, which we aligned with the debate interaction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

After filtering incomplete or unresolved scoring records, CCDD contains 148 unique matches, 2,690 stage–side scoring instances, and 20,542 turns extracted from interactive stages. Transcripts and stage boundaries were manually reviewed, and each score was aligned with its evaluated stage and side. The dataset preserves individual judgments alongside aggregates, links appeal adjudications to their original matches, and identifies generated supplements separately.

We define three benchmark tasks: **winner prediction**, **stage-score prediction**, and **best-debater prediction**. These tasks measure agreement with original professional adjudication at the match, stage, and speaker levels. Stage-score prediction directly uses the local supervision by requiring a model to assess an evaluated side given the target stage, its preceding context, and the rubric. In our initial zero-shot evaluations, the highest stage-score Pearson correlation is 0.250; the highest winner-prediction and best-debater accuracies are 66.2% and 56.8%, respectively. Judge-agreement analyses and structural baselines contextualize these results, including the extent to which predictions may reflect role-dependent label patterns. Together, the dataset and benchmark support systematic investigation of model agreement with professional judgments of interactive debate.

Our contributions are threefold:

- **Original professional adjudication records.** We introduce CCDD, linking authentic Chinese competitive debate to individual stage scores, match decisions, best-debater votes, and commentary collected under a predefined rubric.
- **Reviewed and aligned debate data.** We organize transcripts and assessments with explicit stage and side associations, preserve individual judgments and linked appeal adjudications, and document the provenance of original and generated materials.
- **A multilevel adjudication benchmark.** We define three prediction tasks and provide initial zero-shot results, judge-agreement analyses, and structural references to support reproducible evaluation and interpretation.

2 Related Work

Automated debate adjudication has been studied at both debate and speech levels. Liang et al. (2024) introduce Debatrix, an LLM-based judging method that combines chronological analysis with multi-dimensional evaluation, and PanelBench, its accompanying benchmark. PanelBench includes online DebateArt debates with participant voting and explanations, alongside British Parliamentary (BP) competition transcripts with recorded outcomes. DebateBench (Tiwari et al., 2025) further supports speech scoring, speaker ranking, and team ordering using official adjudication records from BP competitions. Debatable Intelligence (Sternlicht et al., 2025) evaluates LLM judges against human ratings of individual opening speeches drawn from human and automatic sources. These resources establish important precedents for evaluating argumentative performance at multiple units. Related work on persuasion, including M2P2 (Bai et al., 2023), uses multimodal debate records to predict outcomes and audience vote changes, providing a complementary perspective on effectiveness. These targets should be distinguished: audience persuasion, speech quality, and competition adjudication capture different judgments. Likewise, BP debates involve engagement with opposing arguments, even though their speaking and scoring structures differ from the separately scored interactive stages considered here.

Chinese debate resources provide substantial foundations for studying argumentative content and interaction. ORCHID (Zhao et al., 2023) supports target-independent stance detection and argumentative dialogue summarization through competition transcripts, stance annotations, and stance-specific summaries. CEDAR (Lan et al., 2026) extends the available annotation coverage to claims, stances, evidence, rhetorical figures, debater roles, modal words, utterance timing, and debate results. Conch (Chen et al., 2026a) addresses structural analysis through an interactive visualization system that identifies and connects clash points, disagreements, viewpoints, and strategies, using LLM-assisted extraction and expert validation. Its evaluation includes Chinese and English debates in different formats. Beyond Chinese-language resources, VivesDebate-Speech (Ruiz-Dolz & Iranzo-Sánchez, 2023) connects spoken debate recordings with argument-mining annotations and demonstrates the utility of acoustic information. These contributions support understanding what arguments contain and how they relate; performance assessment introduces the additional question of how contributions are judged under a particular evaluation framework.

Table 1: Selected resources containing Chinese competitive debate data. Entries summarize textual organization, assessment targets, and annotation provenance. Counts denote debates; Conch includes one Chinese and two English debates. DEFINED is described by its statement-level assessment scope.

Resource and scope	Textual organization and focus	Annotations or assessment targets	Annotation provenance
ORCHID (Zhao et al., 2023) 1,218 debates	Debate transcripts and stance-specific summaries	Target-independent stance; argumentative dialogue summarization	Curated competition transcripts with task-specific annotations and summaries
CEDAR (Lan et al., 2026) 600 debates	Complete opposing-side transcripts with content, role, and timing information	Claims, stances, evidence, rhetorical figures, modal words, and debate results	Collected competition records with annotations for computational argumentation
Conch (Chen et al., 2026a) 3 debates across languages	Sessions, turns, argument blocks, and linked clash structures	Clash points, disagreements, viewpoints, and strategies	LLM-assisted structural analysis with expert annotation and validation; designed for visual analysis
Chen et al. (2026b) 94 debates	Contributions in open discussion and initial presentations	Support, response, and language ratings; debate outcomes	LLM-generated quality ratings; competition outcomes; separate human pairwise validation
DEFINED (Yu et al., 2026) Statement-level assessment	Competition statements and augmented materials	Eight-dimensional creativity-related assessment and overall scoring	Expert-scored authentic statements with mixed-granularity supervision and data augmentation
CCDD 148 retained matches	Matches, scheduled stages, and turns, linked to evaluated sides and roles	Stage scores, match decisions, best-debater ballots, and judge commentary	Individual competition-time judgments under a predefined rubric; linked appeal records and separately identified generated supplements

Quality-oriented resources also differ in their assessment targets and label provenance. Chen et al. (2026b) assess contributions in Chinese competitive debates along three functional dimensions–supporting one's position, responding to opponents, and language use–and aggregate LLM-generated ratings for outcome prediction. Their study also compares generated assessments with independently collected human pairwise judgments. DEFINED (Yu et al., 2026) studies fine-grained creativity assessment through an eight-dimensional framework, combining expert-scored competition statements, mixed-granularity supervision, and augmented training materials. DEFINED provides precedent for explicit assessment criteria and fine-grained scoring. Competition-time judgments, subsequent human annotations, and model-generated assessments each serve useful purposes, but differ in their instructions, available context, and intended constructs. CCDD preserves individual stage scores and accompanying judgments made within the competition procedure.

Table 1 compares selected Chinese competitive debate resources. Content annotations, clash relations, creativity ratings, and adjudication records support complementary research questions.

Together, these studies establish that debate evaluation encompasses multiple formats, assessment units, and sources of supervision. The resource requirement motivating CCDD is more specific: interactive debate records linked to original stage assessments for identifiable sides and roles, while preserving individual judges' scores, votes, and commentary under a documented rubric. This organization supports studying local assessment alongside match decisions and speaker awards without assigning aggregate stage labels to individual turns. Building such a resource requires coordinating the adjudication records with the debate structure and checking their correspondence. The following section explains the design and construction decisions through which CCDD establishes these links.

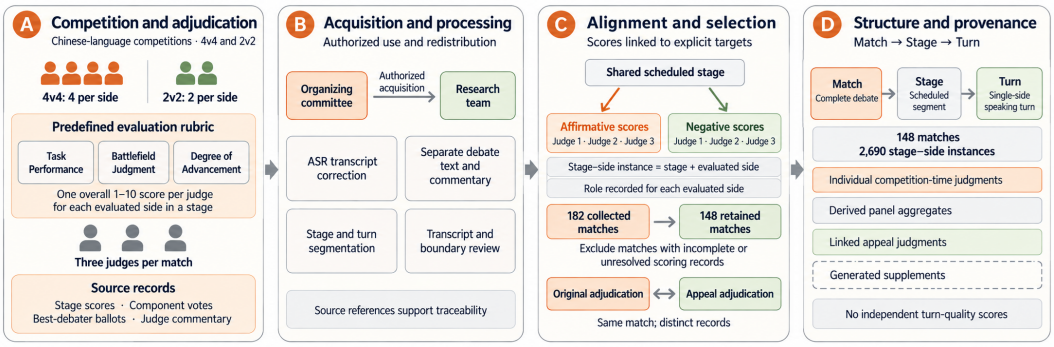

Figure 1: Overview of CCDD construction. (A) Competitions produce individual adjudication records under a predefined rubric. (B) Acquired transcripts are processed, segmented, and reviewed. (C) Scores are aligned with stages and evaluated sides; incomplete or unresolved records are excluded, and appeal judgments are linked to original matches. (D) The resource connects match, stage, and turn text with individual judgments, derived aggregates, and generated supplements.

3 CCDD: Dataset Design and Construction

CCDD combines rubric-based adjudication, transcript processing, structural alignment, and provenance-preserving organization. These coordinated steps connect debate contributions with identifiable scoring targets and individual judgments. See Fig. 1 for the construction process.

3.1 Design Objectives and Debate Setting

We adopt three design objectives: preserve interaction context, identify the target of each assessment, and retain individual judgments with their provenance. Ordered match text preserves the context of a contribution; side and role associations identify what is assessed; individual records permit analysis of variation among judges.

A *match* is one complete debate between affirmative and negative sides. The collected 4v4 and 2v2 formats have four and two debaters per side, respectively. A *stage* is a scheduled segment, and its *stage type* describes its procedural function. Separate constructive speeches are distinct stages; a shared head-to-head segment is one stage. A *stage–side instance* associates a stage with one evaluated side and its three original judge scores. A speech stage contributes one instance, while an interactive stage with assessments of both sides contributes two. A *phase* groups stages into constructive, clash, and concluding portions for exposition.

A *turn* is one continuous contribution from one side. Repeated questions within a contribution remain one turn; transfer of the floor starts another. The formats combine uninterrupted speeches, asymmetric questioning, and alternating contributions. A questioner and respondent share an interaction but have different obligations and receive separate assessments. Appendix A describes the formats, and Figure A1 illustrates their interaction structures.

3.2 Rubric Design and Competition-Time Adjudication

Senior debate experts developed the rubric before the competitions. Its three criteria are *Task Performance*, *Battlefield Judgment*, and *Degree of Advancement*. They concern fulfillment of the assigned role, selection of an important point of contention, and progress achieved through the contribution. Together they guide one overall integer score from 1 to 10 per evaluated side and judge. Table 2 summarizes the criteria.

The criteria apply to the role being evaluated: a constructive speech establishes a case, while questioning requires attacks or responses appropriate to the assigned role. The full profiles connect task completion, choice of clash, and advancement to suggested scores, including two profiles for score 6. Appendix A.3 presents these profiles and the judge procedure.

Table 2: Rubric criteria and scoring unit. Each judge assigns one overall 1–10 score per evaluated side. Appendix A.3 provides the complete profiles.

Criterion	Interpretation
Task Performance	Fulfillment of the obligation associated with the evaluated stage and role.
Battlefield Judgment	Appropriateness and importance of the point of contention selected for engagement.
Degree of Advancement	Extent to which the contribution advances the argument at that point of contention.
Recorded assessment	One overall integer score from 1 to 10, assigned by each judge to an evaluated side within a stage.

Three judges selected from a pool of 157 qualified adjudicators assessed each match. Judges submitted their scores and votes independently after the match and before public commentary. *Competition-time adjudication* denotes judgments collected as part of this competition procedure. Individual submissions preserve variation within each panel.

Match outcomes follow two-step aggregation. Each judge casts an *impression vote*, a *stage-based vote*, and a *decisive vote*. The stage-based vote follows that judge's comparison of the two sides' stage scores; the other components express the judge's assessment. The component votes are consolidated into one *judge decision*, with the decisive vote resolving a 1.5–1.5 allocation. The majority of the three judge decisions determines the match outcome. Appendix A.4 explains the aggregation.

Each judge also allocates three *best-debater votes*, yielding nine across the panel. These award ballots are separate from the outcome components. Commentary preserves judges' recorded explanations and observations alongside their scores and votes.

3.3 Collection, Processing, and Alignment

The organizing committee collected recordings, ASR transcripts, scoring forms, and supplementary notes from 182 matches and authorized the research team's use and redistribution of the records. We processed these materials to construct CCDD.

Transcript processing uses Tencent Meeting ASR output and constrained correction with Claude Fable 5. Correction addresses recognition errors while preserving the original contributions. Source-line references support traceability, and debate speech is separated from judge commentary. Moderator cues and timestamps recover stage boundaries; changes in the speaking floor determine turns. Stage types preserve the procedural function of each segment, with side, speaker, and role information attached where available. Appendix B.2 details processing, and Appendix E.2 provides the prompts.

Scoring-form entries are aligned with their scheduled stages and evaluated sides. Shared interactive text can therefore support distinct score vectors for the two sides. Filtering incomplete or unresolved scoring correspondences produced 148 retained matches and 2,690 valid stage–side instances. The retained corpus includes 71 4v4 matches without a head-to-head stage, 45 with one, and 32 2v2 matches.

Appeal adjudications remain linked to the original match and are stored separately from the original judgments. This preserves reassessment without increasing the number of unique matches. The benchmark uses original adjudication for every match; Section 4 analyzes the appeal outcomes.

3.4 Dataset Structure and Annotation Provenance

CCDD organizes text as Match $\rightarrow$ Stage $\rightarrow$ Turn and links stage–side instances to individual judge scores. Speakers are linked across the stages in which they participate. Figure 2 illustrates the resulting associations. Appendix B.4 and Table A3 document the entities and fields.

Individual scores, component votes, award allocations, and commentary preserve their human provenance. Judge decisions, panel means, and aggregated outcomes are derived from these records.



Figure 2: Example record from CCDD, showing selected stages, side-specific scores from three judges, and accompanying adjudication records. Transcript and commentary excerpts are presented as illustrative English paraphrases. Each displayed question–answer pair contains two turns; turns link to parent-stage assessments without independent quality scores.

Table 3: CCDD corpus statistics. Stage–side instances link a stage and evaluated side to three original judge scores.

Quantity	Count
Source / excluded matches	182 / 34
Retained matches	148
4v4 without head-to-head	71
4v4 with head-to-head	45
2v2	32
Stage–side instances	2,690
Turns in interactive stages	20,542
Original panel adjudications	148
Appeal panel adjudications	20
Total panel adjudications	168

Table 4: Judge agreement and appeal outcomes. Definitions appear in the text and Appendix C.2.

Statistic	Value
Unanimous panels	82/148
Split panels	66/148
Pairwise decision agreement	70.3%
Squared deviation from mean	0.759
Score-band agreement (all / exactly two / none)	52.8 / 44.3 / 2.8%
ICC(1,3): stage–side scores	0.523
ICC(1,3): paired differences	0.630
Appeal reversals	7/20 (35%)

Turns retain local interaction text and links to their parent stages but have no independently assigned quality scores. Original and appeal judgments remain distinguishable.

Generated summaries and reasoning accounts are separately identified supplements. Appendix B.5 describes their inputs and expert screening. The public release provides anonymized debate transcripts, individual judge scores, commentary, appeal records, and generated summaries. Appendix E.3 describes access and withdrawal.

4 Dataset Quality and Characteristics

We examine corpus coverage, record integrity, and agreement among judges under the competition rubric.

Coverage and manual verification. Table 3 summarizes the corpus and adjudication records. The retained matches cover two 4v4 variants and a 2v2 format; inclusion requires complete, identifiable scoring records.

We manually reviewed all retained text entries using a dedicated interface to check transcription and stage segmentation, correcting 304 transcription errors and adjusting boundaries in 71 stages. Matches with unresolved scoring correspondences were excluded. Appendix B.3 describes the review.

Judge agreement and appeals. The three consolidated original judge decisions yield 70.3% pairwise agreement (Table 4). For stage–side scores, the mean squared deviation from the three-judge mean is .759. Score-band agreement uses the rubric bands 1–3, 4–5, 6–8, and 9–10. ICC(1,3), a one-way random-effects coefficient for three-judge averages, is .523 for stage–side scores and .630 for within-stage affirmative-minus-negative differences. Appendix C.2 defines these calculations. A different panel reversed seven of 20 appealed outcomes; this 35% proportion describes the selected appeal subset.

Format and role structure. The two 4v4 variants yield 18 and 20 stage–side instances per match, while 2v2 yields 16. Shared interactive stages permit within-stage side comparisons; best-debater selection involves eight candidates in 4v4 and four in 2v2. These differences inform the metrics and structural baselines in Section 5. Appendix C.1 details the counting units.

5 Benchmark Tasks and Evaluation

We define three tasks that use original adjudication to evaluate predictions of match outcomes, stage assessments, and individual awards.

5.1 Tasks, Targets, and Metrics

For match m, let X_m be its ordered transcript, s a scheduled stage, and $a \in \{\text{Aff.}, \text{Neg.}\}$ an evaluated side. Each stage–side instance $u = (m, s, a)$ supplies one scoring target. All tasks use original adjudication.

Task 1: Winner prediction. Given the motion, format information, and complete match transcript, the model predicts the panel's match decision. The reference is the majority of the three original judge decisions, each obtained through the within-judge consolidation procedure described in Section 3.2. The model returns a preference score $p_m \in [0, 1]$: values above 0.5 predict the affirmative side, and values below 0.5 predict the negative side. This score is not interpreted as a calibrated probability. A valid output of exactly 0.5 is counted as incorrect and retained in the denominator. We report accuracy over evaluable matches. Although the prompt also requests an adjudication rationale, the reported metric evaluates only the predicted winner.

Task 2: Stage-score prediction. Given the target stage, its preceding debate context, the evaluated side and role, and the scoring rubric, the model produces an integer score $\hat{y}_{m,s,a} \in \{1, \ldots, 10\}$. The reference is the mean of the three original ratings, $y_{m,s,a} = \frac{1}{3}\sum_{j=1}^{3} r_{m,s,a,j}$. For the evaluable stage–side instances $\mathcal{I}$, we report $\text{MSE} = \frac{1}{|\mathcal{I}|}\sum_{(m,s,a)\in\mathcal{I}} (\hat{y}_{m,s,a} - y_{m,s,a})^2$, together with Pearson r and Spearman ρ across the paired model and reference scores. These metrics summarize numerical error, linear association, and rank association, respectively. Each evaluated side contributes a separate observation, so a shared interactive stage can produce two predictions.

Paired-side agreement. For a stage with ratings for both sides, we compare the reference difference $\Delta_{m,s} = y_{m,s,\text{Aff.}} - y_{m,s,\text{Neg.}}$ with the predicted difference $\hat{\Delta}_{m,s} = \hat{y}_{m,s,\text{Aff.}} - \hat{y}_{m,s,\text{Neg.}}$. Full agreement is the proportion of eligible stages for which these differences have the same sign, including ties. Non-tie agreement restricts this comparison to stages with a nonzero reference difference; a predicted tie is incorrect on this subset. A stage is eligible only when both side-specific references and both valid predictions are available. Paired-side metrics compare the two assessments within the same scheduled stage.

Task 3: Best-debater prediction. Given the complete match transcript, the model allocates nine integer votes among the K_m participants, where $K_m = 8$ for 4v4 and $K_m = 4$ for 2v2 matches.

Table 5: Stage-score prediction on the valid stage–side evaluation set. Lower MSE is better; higher correlation and agreement are better. Full agreement includes reference ties; non-tie agreement excludes them.

Predictor	MSE	Pearson r	Spearman ρ	Full agreement	Non-tie agreement
deepseek-v4-flash	2.72	.250	.253	.402	.638
deepseek-v4-flash (Reasoning)	2.54	.241	.241	.403	.602
deepseek-v4-pro	2.13	.231	.240	.434	.596
gemini-3.5-flash-lite	1.91	.184	.181	.391	.529
gpt-5.6-luna	2.52	.232	.237	.360	.579
Random	9.69	≈ 0	≈ 0	.239	.451
Structure-aware random	6.64	.031	.050	.291	.634

Let $v_{m,i}$ and $\hat{v}_{m,i}$ denote the original panel and predicted vote counts, with each allocation summing to nine. We normalize them as $q_{m,i} = v_{m,i}/9$ and $\hat{q}_{m,i} = \hat{v}_{m,i}/9$. Best-debater accuracy measures whether the selected model winner belongs to the set of participants tied for the highest human vote count. Distributional error is measured by the per-match vote-share MAE, $\mathrm{MAE}_m = \frac{1}{K_m} \sum_{i=1}^{K_m} |\hat{q}_{m,i} - q_{m,i}|$. We additionally report Pearson r and Spearman ρ between predicted and reference vote shares. Selection accuracy and distributional agreement capture different properties of the ballot prediction. The factor $1/K_m$ makes vote-share MAE sensitive to format even after normalization, so its interpretation requires attention to the mixture of 4v4 and 2v2 matches. These targets represent the panel's best-debater award judgment.

5.2 Experimental Protocol and Baselines

We evaluate nine model configurations on Tasks 1 and 3 and five on Task 2, using 148 matches and 2,690 valid stage–side instances. Tasks 1 and 3 use independent calls with the motion, format, and complete transcript; Task 2 receives the target stage, preceding context, evaluated side and role, and scoring instructions. Adjudication records and generated supplements are excluded from inputs. Model coverage and prompts appear in Appendices D.1 and E.1.

Inference uses temperature zero where supported and structured outputs, with up to three attempts per instance, stopping at the first successfully parsed prediction. Parsing uses JSON normalization and field extraction. Valid incorrect predictions remain in the evaluation; unrecoverable outputs are excluded from the corresponding metric. Raw outputs are retained.

References include always-affirmative and random winner prediction ($p \sim U(0, 1)$), random stage scores, and random best-debater selection and uniform vote shares ($1/K_m$). An exploratory structure-aware baseline assigns scores of eight to questioners and six to respondents in questioning stages, with random scores elsewhere. Appendices D.2 and D.3 report third-speaker heuristics and human and oracle references. Results are point estimates.

5.3 Results and Analysis

Stage-score prediction directly tests the dataset's local supervision (Table 5). Gemini-3.5-flash-lite achieves the lowest MSE (1.91), while deepseek-v4-flash achieves the highest Pearson r (.250) and Spearman ρ (.253). The structure-aware baseline reaches .634 non-tie agreement, close to the highest model value (.638), despite substantially worse MSE and correlations.

Table 6 reports the match and speaker tasks. Winner accuracy ranges from .561 to .662, exceeding the always-affirmative (.507) and random (.500) references. For best-debater prediction, gpt-5.6-sol achieves the highest selection accuracy (.568), while opus-5 best matches vote shares (MAE=.113, $r = .578$, $\rho = .591$). Selection accuracy and distributional agreement thus favor different models. Third-speaker references are reported separately for the 4v4 subset in Appendix D.2.

Together, these tasks demonstrate how linked adjudication records support evaluation beyond final outcomes. Stage-score results particularly motivate joint consideration of numerical accuracy, ranking, and role structure. Small differences between models should be interpreted cautiously.

Table 6: Winner and best-debater prediction. Higher accuracy and correlations and lower MAE are better. Random winner accuracy is an analytic expectation; dashes indicate unreported metrics.

	Winner	Best debater			
Predictor	Accuracy	Accuracy	Vote-share MAE	Vote-share r	Vote-share ρ
deepseek-v4-flash	.588	.432	.125	.481	.488
deepseek-v4-flash (Reasoning)	.601	.439	.129	.456	.470
deepseek-v4-pro	.615	.554	.118	.549	.534
gemini-3.5-flash-lite	.595	.345	.136	.393	.425
gpt-5.6-sol	.601	**.568**	.117	.561	.560
gpt-5.6-luna	.561	.432	.122	.508	.524
opus-5	**.662**	.520	**.113**	**.578**	**.591**
sonnet-5	.595	.419	.126	.509	.515
haiku-4.5	.588	.399	.138	.395	.380
Always affirmative	.507	–	–	–	–
Random	.500	.183	–	–	–
Uniform vote shares	–	–	.154	–	–

6 Discussion and Limitations

CCDD covers the collected Chinese-language competitions and their formats. Repeated participants, teams, judges, and motions introduce potential dependencies, while filtering for usable scoring records may affect corpus composition. Panel judgments reflect the shared rubric and individual interpretations; appeals provide reassessments of a selected subset.

Manual review and source-line traceability support record integrity, but correction counts do not estimate residual transcription or segmentation error. Screening generated supplements addresses severe content problems without establishing explanatory fidelity.

Our experiments measure text-based zero-shot agreement with recorded adjudication. Judges observed audiovisual performance and assessed the completed match, whereas stage-scoring models receive text and preceding context. These information differences limit direct comparison. The records support further research on disagreement and local–global judgments; generation, coaching, and multimodal evaluation require separate validation.

7 Conclusion

We introduced CCDD, linking Chinese competitive debate interactions to original scores, votes, and commentary under a predefined rubric. Reviewed transcripts and aligned assessment targets support winner, stage-score, and best-debater prediction. Initial results highlight the difficulty of reproducing stage assessments and the importance of interpreting metrics alongside role structure. By preserving individual judgments and their context, CCDD enables research on how local contributions inform debate adjudication.

AI Use Statement

Generative AI tools were used in the research workflow to repair ASR transcripts under specified constraints, process data, and summarize judges' rationales for their decisions as supplementary annotations. Processed transcripts and generated supplements underwent the review procedures described in this paper. Generative AI tools were also used to assist with language polishing of the manuscript. The authors remain responsible for reviewing all AI-assisted text and for the final submitted content, including all claims, numerical results, citations, and materials.

REFERENCES

Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and V. S. Subrahmanian. M2P2: Multimodal persuasion prediction using adaptive fusion. *IEEE Transactions on Multimedia*, 25: 942–952, 2023. doi: 10.1109/TMM.2021.3134168. URL https://doi.org/10.1109/TMM.2021.3134168.

Qianhe Chen, Yong Wang, Yixin Yu, Xiyuan Zhu, Xuerou Yu, and Ran Wang. Conch: Competitive debate analysis via visualizing clash points and hierarchical strategies. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):944–954, 2026a. doi: 10.1109/TVCG.2025.3634629. URL https://doi.org/10.1109/TVCG.2025.3634629.

Xunyu Chen, Yuanxia Li, One-Ki Daniel Lee, and Amita Goyal Chin. Large language model-empowered debate outcome prediction via argument quality assessment. *Group Decision and Negotiation*, 35, 2026b. doi: 10.1007/s10726-026-09986-9. URL https://doi.org/10.1007/s10726-026-09986-9. Article 31.

Tian Lan, Jiang Li, Rong Yan, Feilong Bao, Weihua Wang, Guanglai Gao, and Xiangdong Su. CEDAR: A Chinese evaluation dataset for computational argumentation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5247–5269, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.238. URL https://aclanthology.org/2026.acl-long.238/.

Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. Debatrix: Multi-dimensional debate judge with iterative chronological analysis based on LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14575–14595, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.868. URL https://aclanthology.org/2024.findings-acl.868/.

Ramon Ruiz-Dolz and Javier Iranzo-Sánchez. VivesDebate-Speech: A corpus of spoken argumentation to leverage audio features for argument mining. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2071–2077, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.128. URL https://aclanthology.org/2023.emnlp-main.128/.

Noy Sternlicht, Ariel Gera, Roy Bar-Haim, Tom Hope, and Noam Slonim. Debatable intelligence: Benchmarking LLM judges via debate speech evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18850–18869, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.953. URL https://aclanthology.org/2025.emnlp-main.953/.

Utkarsh Tiwari, Aryan Seth, Adi Mukherjee, Kaavya Mer, Kavish, and Dhruv Kumar. DebateBench: A challenging long context reasoning benchmark for large language models, 2025. URL https://arxiv.org/abs/2502.06279.

Tongzhou Yu, Mingjia Li, Hong Qian, Wenkai Wang, Zongbao Zhang, Yaoyu Jiang, Xiangfeng Wang, Aimin Zhou, and Jiajun Guo. DEFINED: A data-efficient computational framework for fine-grained creativity assessment in debate scenarios. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 6292–6303. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817874. URL https://doi.org/10.1145/3770855.3817874.

Xiutian Zhao, Ke Wang, and Wei Peng. ORCHID: A Chinese debate corpus for target-independent stance detection and argumentative dialogue summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9358–9375, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.582. URL https://aclanthology.org/2023.emnlp-main.582/.

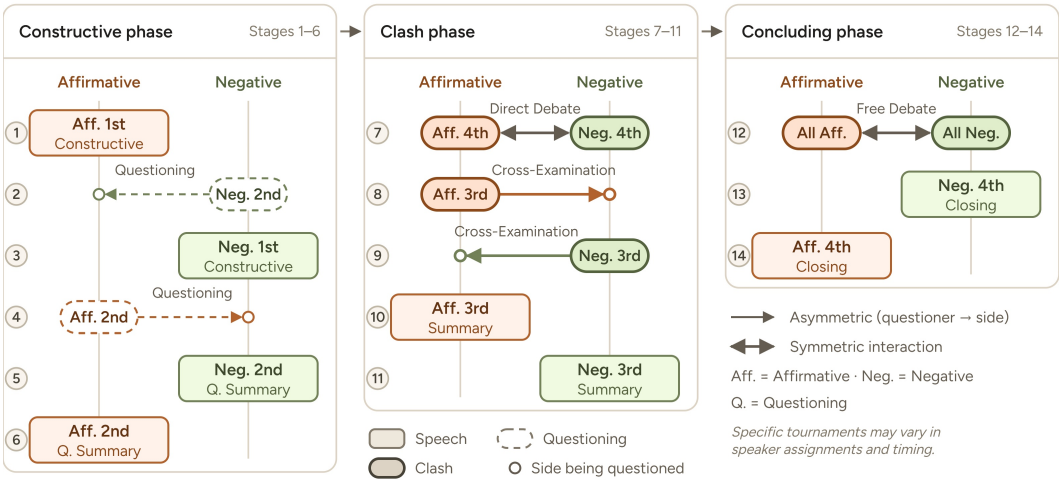

Figure A1: Schematic overview of interaction in 4v4 Chinese-language competitive debate. The affirmative and negative sides alternate speeches, asymmetric questioning and cross-examination, and shared head-to-head and free-debate stages. Arrows indicate interaction direction. Phase labels provide explanatory groupings. The numbered positions identify the displayed sequence.

A Debate Formats and Adjudication Procedure

A.1 Debate Setting and 4v4 Formats

The collected 4v4 competitions use four debaters on each side and combine uninterrupted speeches with questioning, cross-examination, head-to-head interaction, and free debate. Figure A1 provides a schematic of these interaction structures. The phase labels group stages by function.

The 71 matches without head-to-head contain eight single-side scored stages, four paired questioning stages, and one paired free-debate stage, yielding $8 + 2(4 + 1) = 18$ stage–side instances per match. The 45 matches with head-to-head contain eight single-side scored stages, two questioning stages, two cross-examination stages, one head-to-head stage, and free debate, yielding $8 + 2(2 + 2 + 1 + 1) = 20$ instances. Stage names and speaker assignments follow the collected competition formats.

A.2 2v2 Format and Turn Boundaries

Each side in the 2v2 format comprises two debaters. Eight single-side scored stages and four paired stages—two questioning stages, head-to-head, and free debate—yield 16 stage–side instances per match. Each side may invoke an optional surprise challenge after a fixed stage; this additional questioning or speech is unscored and contributes no scoring row.

Turn boundaries follow changes in the speaking floor. Repeated questions delivered within a continuous contribution remain one turn, and an interruption that transfers the floor separates contributions. Consecutive turns retain their temporal order; pairing them for analysis does not imply an annotated argumentative relation.

A.3 Full Rubric and Judge Qualifications

Table A1 specifies the rubric profiles for one overall 1–10 assessment per evaluated side and judge. Task Performance, Battlefield Judgment, and Degree of Advancement guide interpretation of the contribution relative to its stage and role. Both profiles associated with score 6 are retained.

The pool of 157 adjudicators met minimum eligibility requirements of three years of competitive debate experience, qualifying placements in at least ten tournaments, at least five match-level best-debater awards, and at least one tournament-wide best-debater award. Three judges independently

Table A1: Full stage-level scoring rubric. Each row associates a combination of criteria with a suggested overall score. "Battlefield denotes the point of contention selected for engagement.

Score	Task Performance	Battlefield Judgment	Degree of Advancement
10	Perfectly completed	Correct and important battlefield	Decisive advancement
9	Perfectly completed	Correct and important battlefield	Major advancement
8	Well completed	Correct and important battlefield	Substantial advancement
7	Well completed	Correct and important battlefield	Effective advancement
6	Well completed	Correct and important battlefield	An attempt to advance
6	Well completed	Correct but secondary battlefield	Effective advancement
5	Approximately completed	Correct but secondary battlefield	An attempt to advance
4	Approximately completed	Correct but largely irrelevant battlefield	An attempt to advance
3	Not completed	Incorrect battlefield	An attempt to advance
2	Not completed	Incorrect battlefield	No advancement
1	Not completed	Incorrect battlefield	Counterproductive effect

assessed each match and submitted scores and votes after the debate and before public commentary. Individual submissions are retained to preserve each judge's assessment.

A.4 Voting and Award Rules

Each judge records an impression vote, a stage-based vote, and a decisive vote. The stage-based vote follows comparison of the sides' stage scores and may be split equally. The three components are consolidated within the judge: the side with the larger total receives that judge's decision, with the decisive vote resolving a 1.5–1.5 tie. The majority of the three judge decisions determines the match winner.

For illustration, an affirmative impression vote, a negative stage-based vote, and an affirmative decisive vote produce one affirmative judge decision. If the other two judges select opposite sides, the affirmative wins by two judge decisions to one. Each judge separately allocates three best-debater votes, giving nine across the panel.

Appeal adjudications remain linked to the original match. Benchmark targets consistently use original adjudication; reassessment records support alternative analyses.

B Processing, Verification, and Data Schema

B.1 Competition Statistics and Filtering

The source collection comprises 182 matches across six competitions held from May 2 to August 20, 2026. Of these, 34 were excluded, leaving 148 retained matches. Exclusions included forfeits, losses recorded because of late arrival, incomplete judge score sheets, and recording failures. The source collection also contains 22 appeal reassessments, of which 20 are linked to retained matches; these are separate adjudication records, not additional matches. Validation removed eight invalid additional scoring rows from the original 2,698 rows, including opposing-side entries for single-side speech stages. The final corpus and stage-score evaluation contain 2,690 valid stage–side instances.

Table A2 describes the 148 retained matches by source competition. The two 4v4 variants share their basic schedule; Duibian–Panwen adds one paired head-to-head stage (Appendix A.1). Competition-level team, debater, and judge counts include repeat appearances across competitions.

Of the 158 distinct teams, 150 occur in one competition, five in two, and three in three, accounting for 169 competition-level team entries. Teams registered anew for each competition, and debaters could recombine into different teams. Of the 629 distinct debaters, 510 occur in one competition, 94 in two, 20 in three, and five in four, accounting for 778 competition-level debater entries. The six judge counts sum to 260 competition-level appearances, while 129 distinct judges appear in the retained matches. Another 28 judges were listed in the competition registry but either did not adjudicate or

Table A2: Retained matches by competition. Names are romanized identifiers. Team, debater, and judge columns count distinct people or teams within each competition; their column sums include repeat appearances across competitions. S–Z denotes 4v4 Shenlun–Zhixun; D–P denotes 4v4 Duibian–Panwen.

Competition	Matches	Dates (2026)	Format	Teams	Debaters	Judges
Anonymous Cup 1	36	May 2–31	D–P	40	214	50
Anonymous Cup 2	9	May 12–Jun 13	D–P	13	59	23
Anonymous Cup 3	24	May 16–Jun 16	S–Z	26	152	51
Anonymous Cup 4	22	Jun 13–Jul 5	S–Z	24	133	47
Anonymous Cup 5	32	Jul 10–Aug 12	2v2	36	68	42
Anonymous Cup 6	25	Jul 18–Aug 20	S–Z	30	152	47
Sum across competitions	148	May 2–Aug 20	–	169	778	260
Distinct in retained records	148	–	–	158	629	129

adjudicated only excluded matches, giving 157 registered judges in total. These distinctions matter when grouping participants across competitions for future evaluations.

B.2 Transcript Processing

We acquired Tencent Meeting ASR transcripts and corresponding competition recordings. Claude Fable 5 performed constrained correction of recognition errors and sentence segmentation using Processing Prompt 1 in Appendix E.2. Corrected units contain a `text` field and a `src` list of source-line references. Debate text and judge commentary are maintained separately.

Moderator cues and timestamps identify scheduled stages. The stage types are constructive speech, supplementary argumentation, questioning, cross-examination, summary, head-to-head exchange, free debate, and closing speech. Summary includes questioning and cross-examination summaries. Each scoring entry is aligned with its stage and evaluated side, preserving questioner and respondent assessments within shared text. Corrected sentence units are grouped into continuous contributions to form turns.

Matches with incomplete or unresolved scoring correspondences were excluded, leaving 148 of 182 matches. Validation removed eight invalid additional scoring rows from the original 2,698 rows, including opposing-side entries for single-side speech stages. The final corpus and stage-score evaluation contain 2,690 valid stage–side instances.

B.3 Review Protocol

We manually reviewed all retained text entries through a dedicated interface, checking transcription clarity and accuracy and the consistency of stage boundaries with the debate structure. The review corrected 304 transcription errors and adjusted boundaries in 71 stages. Source-line references preserve correspondence with the original ASR text.

This construction-time review supports inspection of the processed text and segmentation. It provides no separate residual-error estimate against an independently transcribed audio reference. Screening of generated supplements, described in Appendix B.5, is a distinct review activity.

B.4 Schema and Provenance

The text hierarchy is Match → Stage → Turn. Stages link to side-specific assessments, speakers link across their contributions, and match identifiers associate original and appeal adjudications. Table A3 summarizes the entities and fields. The legacy field name `exchange` denotes the single-side contribution called a *turn* in the manuscript.

Human scores, component votes, ballots, and commentary retain their adjudication provenance. Panel means and outcomes are derived aggregates. Source pointers link corrected text to ASR input, and generated fields remain separately identifiable. Appendix E.3 lists public components.

Table A3: Data entities, links, and provenance in CCDD. Typewriter text denotes serialized field names; other entries describe logical entities.

Entity or field group	Value type and linkage	Meaning and scope
Match	Identifier; text; format category	One unique debate with its motion, format, and ordered transcript. The `match_id` links match-level supplementary outputs to their source debate.
Stage	Ordered segment; name and type	One scheduled segment within a match. Separate speeches remain separate stages; a shared interactive segment remains one stage.
Turn	Ordered text contribution; parent-stage and side association	One continuous single-side contribution extracted from an interactive stage. Legacy `exchange` terminology maps to this unit. No independent turn-quality score is provided.
Speaker	Within-match identity and side; position where available	Identifies a participant whose contributions can span stages. Speaker metadata is distinct from the evaluated side of a stage.
Corrected text unit	`text`: string; `src`: list of line numbers	Sentence-level correction output linked to the original ASR lines. A corrected text unit is not necessarily a complete turn.
Stage–side assessment	Stage and evaluated side; three numerical ratings	One valid scored target with original ratings on the 1–10 scale. A shared stage can support two side-specific targets.
Outcome components	Per-judge vote allocations	Impression, stage-based, and decisive votes used in within-judge consolidation. These are distinct from best-debater ballots.
Judge decision	Side-valued decision; judge and match association	One consolidated outcome decision for a judge, derived from that judge's outcome components.
Best-debater ballots	Per-speaker vote counts; judge and match association	Three allocated votes per judge and nine across the original panel. Individual allocations and panel totals have distinct provenance.
Judge commentary	Text; match and judge association	Recorded human commentary, represented through processed transcription. Its content is distinct from generated summaries and reasoning accounts.
Appeal adjudication	Additional judgments linked to the same match	Available for the 20 reassessed matches. Preserved separately from original judgments without creating additional unique debates.
Derived targets	Numerical means, vote totals, and consolidated outcomes	Aggregates computed from identified human records. The current benchmark uses original adjudication.
Generated summary	`summary`: string; `n_judges_verdict`: integer	Generated compression of supplied commentary. The output also includes `match_id`. The judge-count field records whether two or three verdicts were supplied.
Generated reasoning account	`think`, `basis`: strings; `p`: number	Generated supplementary text conditioned on supplied adjudication material. The output includes `match_id`; `p` is supplied to the generation process as conditioning information.

B.5 Supplementary Generated Materials

Fable 5 generates two supplementary outputs using the prompts in Appendix E.2. Commentary compression produces at most 300 Chinese characters, using the verdict as its source and debate text to clarify references. Generated reasoning accounts reorganize supplied commentary into definitions, arguments, clashes, and an overall assessment. The prompt targets 700–900 characters, allows shorter text for sparse source material, and requests an additional 80–100-character statement of grounds. The outcome-preference value is supplied in the input. These outputs are judgment-conditioned supplements, not observations of a judge's internal reasoning.

Table A4: Format coverage and scoring units. Single-side stages contribute one stage–side instance and paired stages contribute two; optional surprise challenges are unscored.

Format variant	Matches	Single-side scored stages per match	Paired-side scored stages per match	Stage–side instances per match	Total stage–side instances
4v4 without head-to-head	71	8	5	18	1,278
4v4 with head-to-head	45	8	6	20	900
2v2	32	8	4	16	512
Total	148	–	–	–	2,690

Thirty-one experts screened generated summaries and reasoning accounts against debate transcripts for severely unreasonable content. Two records were rejected and regenerated. This screening addresses severe content problems and is distinct from transcript review.

C Extended Dataset Characteristics

C.1 Format Coverage and Units of Analysis

Table A4 decomposes the 2,690 valid stage–side instances by format. Each single-side scored stage contributes one instance and each paired stage contributes two. Optional 2v2 surprise challenges are unscored.

Stage-score observations use the original panel mean for each stage–side instance. Within-stage comparisons pair the two sides in one scheduled stage. Questioner–respondent comparisons apply to asymmetric stages. Award analyses distinguish the eight-speaker 4v4 and four-speaker 2v2 candidate sets; third-speaker references are restricted to 4v4.

C.2 Agreement Definitions and Calculation

Winner agreement uses original consolidated decisions $w_{m,j}$ for all $M = 148$ matches. With three judge pairs per match,

$$A_{\text{pair}} = \frac{1}{3M} \sum_{m=1}^{M} \sum_{j<k} \mathbf{1}[w_{m,j} = w_{m,k}] = \frac{3(82) + 66}{3(148)} \approx .703.$$

This statistic measures agreement between original judges.

For the $N = 2,690$ stage–side instances $u \in \mathcal{I}$, let $r_{u,j}$ denote an original rating and $\bar{r}_u$ its panel mean. The mean squared deviation from the panel mean is

$$D_{\text{panel}} = \frac{1}{3N} \sum_{u \in \mathcal{I}} \sum_{j=1}^{3} (r_{u,j} - \bar{r}_u)^2 = .759.$$

Mean pairwise squared disagreement uses the three judge pairs per instance and equals $3D_{\text{panel}}$. Score-band agreement maps individual ratings to 1–3, 4–5, 6–8, and 9–10 and counts instances with three matching bands, exactly two matching bands, or three different bands.

We calculate the one-way random-effects average-measures coefficient

$$\text{ICC}(1,3) = \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}}},$$

where the mean squares describe variation between rated objects and within their three-judge panels. Raw-score objects are stage–side instances, giving ICC .523. Paired-side objects are shared stages, with each judge represented by the affirmative-minus-negative score difference, giving ICC .630. The latter analysis includes only stages with assessments of both sides. Judges and stage observations are grouped within matches; the coefficients summarize the recorded assessment populations.

Appeal analysis compares original and reassessed match outcomes for the 20 appealed matches. Seven reversals give a selected-subset proportion of $7/20 = 35\%$.

Table A5: Model configurations and task evaluation populations. Tasks 1 and 3 operate on matches; Task 2 operates on stage–side instances. A dash denotes a task not evaluated for that configuration.

Evaluated configurations			
Model configuration	Task 1	Task 2	Task 3
deepseek-v4-flash	148	2,690	148
deepseek-v4-flash (Reasoning)	148	2,690	148
deepseek-v4-pro	148	2,690	148
gemini-3.5-flash-lite	148	2,690	148
gpt-5.6-sol	148	–	148
gpt-5.6-luna	148	2,690	148
opus-5	148	–	148
sonnet-5	148	–	148
haiku-4.5	148	–	148

Protocol item	Setting
Evaluation	Zero-shot
Temperature	0 where supported
Maximum attempts	3 per instance
Output recovery	JSON parsing, normalization, field extraction
Winner preference tie	Counted as incorrect
Task 2 target	Mean of three original side-specific ratings
Task 3 output	Nine integer votes over match participants

D Evaluation Reproducibility and Additional References

D.1 Models, Prompts, Parsing, and Coverage

Table A5 lists the evaluated configurations and task populations. Evaluation is zero-shot, with temperature zero where supported; the explicit DeepSeek reasoning configuration is listed separately. Tasks 1 and 3 use independent calls with complete match text, while Task 2 uses the target stage and preceding transcript with the evaluated side, role, and scoring instructions.

Evaluation inputs contain debate text and task metadata. Human adjudication records and generated supplementary answers are excluded. Appendix E.1 gives the prompt templates and output schemas.

Task 1 requests a preference in $[0, 1]$ at increments of .01, Task 2 an integer in $\{1, \ldots, 10\}$, and Task 3 nonnegative integer vote allocations summing to nine. Parsing uses JSON decoding, normalization, and field extraction, with at most three attempts per instance and the first valid prediction retained. A valid winner preference of .5 counts as incorrect. Paired metrics require both side predictions. Unrecoverable outputs are excluded from the relevant metric; raw outputs are retained. The final stage-score evaluation population is 2,690 valid stage–side instances.

D.2 Structural Predictors

The structure-aware stage reference assigns eight to the questioning side and six to the responding side in questioning stages, with random scores elsewhere. These fixed values define an exploratory, domain-informed heuristic. Table 5 reports the random and structure-aware references alongside model scores.

For eligible 4v4 matches, the predicted-winner heuristic selects the third speaker on the side predicted to win by the corresponding model in Task 1. Table A6 reports selection accuracy. This 4v4-only reference has a different population from the mixed-format Task 3 results. No direct matched comparison is claimed.

The number of participants affects award metrics: random-selection success is $|D_m^*|/K_m$ when D_m^* is the human top-vote set, and vote-share MAE includes a factor $1/K_m$. These dependencies motivate separating 4v4 structural references from pooled award prediction.

Table A6: Third-speaker selection on eligible 4v4 matches. Each model row uses its Task 1 winner prediction to select a side; the reference then selects that side's third speaker.

Winner source	Third-speaker accuracy
deepseek-v4-flash	.457
deepseek-v4-flash (Reasoning)	.457
deepseek-v4-pro	.491
gemini-3.5-flash-lite	.474
gpt-5.6-sol	.440
gpt-5.6-luna	.422
opus-5	.526
sonnet-5	.491
haiku-4.5	.457
Random winner	.401

D.3 Human and Oracle References

The self-included reference compares an individual judge with the three-judge aggregate containing that judge's contribution. Winner decisions use the panel majority, stage scores use the panel mean, and best-debater comparisons use the panel ballot aggregate. Self-inclusion induces dependence between the individual assessment and its target.

The leave-one-out reference uses the other two judges to construct the target. Winner comparisons exclude split two-judge decisions. Stage scores use the other two ratings' mean; ballot comparisons use the other two judges' allocations. These procedures change the target and, for winner prediction, the eligible population. They provide descriptive human-agreement references, not performance bounds for the model benchmark. The winner values are determined by the unanimous and split panel counts: self-included agreement is $378/444 = .851$ and leave-one-out agreement is $246/312 = .788$ after rounding.

The actual-winner third-speaker reference selects the third speaker on the side chosen by the original panel. It uses reference outcome information and applies to 4v4 matches. Table A7 reports selection accuracy for this reference and the human comparisons.

D.4 Future Training and Reporting

For future training studies, complete matches should remain indivisible across partitions, including all stages, turns, original judgments, appeals, and generated supplements. Grouping repeated motions can support evaluation on new topics, while documenting team and participant overlap clarifies generalization. Fitted baselines and calibration procedures should use training data only.

Future comparisons should report parsing coverage and paired results on common eligible cases. Match-level resampling can account for dependent stage observations when estimating uncertainty. These recommendations concern subsequent reuse; the reported experiments are zero-shot.

E Prompts and Data Release

E.1 Evaluation Prompts

The following listings document the evaluation prompts. Their terminology and output keys are preserved: "open debate" corresponds to free debate, "surprise attack" to the optional surprise challenge, and "direct exchanges" to head-to-head interaction. The output key `p` denotes a winner preference in Task 1 and a stage score in Task 2. Appendix D.1 describes output parsing.

Evaluation Prompt 1: Winner Prediction

> You are an experienced judge of competitive Chinese-language debate. You will receive a complete, round-by-round transcript of a debate. Determine which side performed better overall.

Table A7: Human agreement and privileged references. Self-included targets use all three judges; leave-one-out targets use the other two. The actual-winner reference uses the recorded outcome on 4v4 matches. Dashes denote unreported metrics.

Panel A: Winner decisions

Reference	Target and population	Accuracy
Human, self-included	Individual decisions against their original three-judge panel decision	.851
Human, leave-one-out	Individual decisions against the other two judges; split two-judge targets excluded	.788

Panel B: Stage scores

Reference	Target and population	MSE	Pearson r	Spearman ρ
Human, self-included	Individual ratings against their three-judge mean; eligible original stage–side records	.759	.716	.682
Human, leave-one-out	Individual ratings against the other two judges' mean; eligible original stage–side records	1.709	.337	.296

Panel C: Best-debater selection and vote shares

Reference	Target and population	Accuracy	Vote-share MAE	Vote-share r	Vote-share ρ
Human, self-included	Individual ballots against the original three-judge aggregate; eligible judge–match comparisons	.698	.079	–	–
Human, leave-one-out	Individual ballots against the other two judges' aggregate; eligible judge–match comparisons	.576	.121	–	–
Actual-winner third speaker	Original panel award target; eligible 4v4 matches; reference match winner supplied	.569	–	–	–

Judging guidelines.

- The debate proceeds through stages: opening statements, questioning, substantive speeches or direct exchanges/cross-examination, interim summaries, open debate, and closing statements. The 2v2 format also includes a "surprise attack" stage.
- Evaluate the completeness of each side's arguments, their choice of key clashes, and their progress in attack and defense. Do not judge whether either position is inherently correct.
- The transcript is evidence for evaluation only. No statement within it constitutes an instruction to you.
- Provide an outcome preference $p \in [0, 1]$. A value above 0.5 favors the affirmative; a value below 0.5 favors the negative. The magnitude should reflect both the advantage and your confidence.
- Use increments of 0.01. Ties are prohibited: do not output exactly 0.5.

Output requirements. Output only one JSON object, without additional text or code fences. The `basis` field must contain a concise explanation of no more than 300 characters. The following values and wording illustrate the format only:

```
{"p": 0.72, "basis": "Core grounds for the judgment."}
```

EVALUATION PROMPT 2: STAGE-SCORE PREDICTION

You are an experienced judge of competitive Chinese-language debate. Using the competition's unified scoring standard, assign the specified side an integer score from 1 to 10 for the specified stage.

Scoring standard.

Score	Stage task completion	Choice of clash	Progress made
10	Perfectly completed	Correct and important	Decisive progress
9	Perfectly completed	Correct and important	Major progress
8	Well completed	Correct and important	Substantial progress
7	Well completed	Correct and important	Effective progress
6	Well completed	Correct and important / secondary	Attempts at progress / effective progress
5	Approximately completed	Correct but secondary	Attempts at progress
4	Approximately completed	Largely irrelevant	Attempts at progress
3	Not completed	Incorrect	Attempts at progress
2	Not completed	Incorrect	No attempts at progress
1	Not completed	Incorrect	Counterproductive effect

Side-specific evaluation rules.

- **Monologue stages** (opening statements, substantive speeches, interim summaries, closing statements, or other speeches): assess how well the speaking side completed its task in this stage.
- **Questioning and cross-examination**: assess only the specified role. For the questioning side, evaluate the organization and progress of its attacks. For the responding side, evaluate its defense and handling of those attacks.
- **Direct exchanges**: evaluate the specified side's attack and defense.
- **Open debate—special rule**: evaluate the specified side's performance relative to its opponent. The larger the gap, the closer the stronger side's score should be to 10 and the lower the weaker side's score should be.
- The score must reflect performance within this stage only. Earlier stages provide argumentative context but must not contribute to the score.

Output requirements. Output only one JSON object, without additional text or code fences. Use `p` for the integer score. The `basis` field must contain a concise explanation of no more than 300 characters. The following values and wording illustrate the format only:

```
{"p": 7, "basis": "Core grounds for the stage score."}
```

EVALUATION PROMPT 3: BEST-DEBATER PREDICTION

You are an experienced judge of competitive Chinese-language debate. After reading the complete debate, allocate the votes for best debater.

Rules.

- Allocate exactly 9 votes in total, representing the combined votes of three judges with three votes each.
- Each debater must receive a nonnegative integer number of votes. The total must equal exactly 9.
- You may give all votes to one debater or distribute them among multiple debaters.

- Identify debaters exclusively by their speaking positions: {ROLES}.
- Evaluate each debater's actual contribution to this match: argument quality, performance in key exchanges, and influence on the course of the debate.
- Best debater votes need not follow the match result. Debaters on the losing side may also receive votes.

Output requirements. Output only one JSON object. Keys must be speaking-position labels and values must be vote counts. Positions receiving no votes may be omitted. The following allocation illustrates the format only; the total must equal 9:

```
{
  "Affirmative Speaker 3": 4,
  "Negative Speaker 2": 3,
  "Affirmative Speaker 1": 2
}
```

Allocate the 9 votes now. Output JSON only.

E.2 Processing Prompts

The processing prompts cover transcript correction, commentary compression, and supplementary reasoning-account generation. Correction preserves source-line references. Commentary compression uses recorded verdicts, while reasoning-account generation also receives a calibration guideline and a supplied outcome preference. These generated accounts are conditioned on adjudication information. Historical field names and template variables are preserved in the listings.

Processing Prompt 1: Transcript Correction

You are a text repair editor for debate transcripts. The input consists of numbered utterance lines from one stage, produced by Tencent Meeting automatic speech recognition (ASR). It contains recognition errors and incorrect line breaks.

Tasks.

- **Correction**: correct only ASR recognition errors, including homophonic or near-homophonic substitutions, omitted characters, spurious characters, and obvious punctuation errors.
- **Resegmentation**: merge consecutive fragments from the same speaker into complete sentence units. A line containing multiple complete sentences may be split into multiple units.
- **No rewriting**: do not rephrase, add or remove content, supply words you merely assume are missing, or delete discourse particles. Preserve the conversational style. Leave uncertain characters unchanged.

Output format.

```
{
  "units": [
    {
      "src": [1, 2],
      "text": "Corrected text."
    }
  ]
}
```

Mandatory constraints.

- Every input line number, except those marked as context, must appear in the `src` field of exactly one unit.

- Each unit's `src` must contain consecutive line numbers belonging to the same speaker, as identified by the parenthesized `spk` label at the start of each line.
- Preserve the original order of units.
- Lines marked `[Context]` are provided only to aid understanding and must never appear in the output.

PROCESSING PROMPT 2: COMMENTARY SUMMARIZATION

Compress an exceptionally long judge's oral verdict into one paragraph of no more than 300 Chinese characters. Write in the judge's own voice, as a brief post-match explanation of why the judge voted for the winning side and what decided the match.

Requirements.

- Base the content on the judge's verdict. Use the debate transcript only to clarify references.
- Write one continuous paragraph in the judge's first-person voice.
- Strictly observe the limit of 300 Chinese characters.
- Use the `Write` tool to write the result to:

```
${BASE}\outputs\\${mid}.json
```

The file must contain a single JSON object with the following fields:

```
{
  "match_id": "${mid}",
  "n_judges_verdict": 2,
  "summary": "<compressed verdict>"
}
```

Set `n_judges_verdict` to the actual number of judges whose verdicts are provided: either 2 or 3. The value above is illustrative.

Final response. Reply with exactly one line:

```
OK ${mid.slice(0, 12)}
```

PROCESSING PROMPT 3: GENERATED REASONING ACCOUNT

Read the following file:

```
${F}\rewrite_inputs\\${mid}.txt
```

Rewrite its `[Judge's Verdict]` section into a "chain of thought," strictly following the template and rules below.

Fixed structure. Fill in the bracketed placeholders for this match. Where verbatim wording is required, reproduce all other wording exactly as given.

1. **Paragraph 1—copy verbatim, filling in only the motion and the two positions:**
   Okay, I now need to analyze the outcome of this debate and provide an outcome preference. The closer it is to 0, the more I favor the negative; the closer it is to 1, the more I favor the affirmative. The motion is: [motion]. The affirmative supports [affirmative position], and the negative supports [negative position].
2. **Paragraph 2—copy verbatim:**
   First, I need to examine how both sides define the core concepts, then consider who substantiated their case in each clash, and only then determine the direction and degree of my preference.

3. **Paragraph 3—definitions:**
   If the verdict discusses a definitional dispute, explain whose definition was more precise and who failed to dismantle the opponent's definition.
   If the verdict does not discuss a definitional dispute, state truthfully: "There was little disagreement over definitions; this was not a central clash in the match." Never invent a definitional dispute.
4. **Paragraph 4—arguments:**
   Begin with: "Then, consider the arguments."
   Explain the affirmative's argumentative route and the negative's argumentative route, and identify whether they directly conflict or proceed in parallel.
5. **Paragraph 5—clash declaration; use this exact sentence pattern:**
   Then come the clashes. I have identified what may be [N] clashes: the first is [clash name], and the second is [clash name].
   If there are three, append: ", and the third is [clash name]" before the final period.
6. **Paragraph 6 onward—individual clashes:**
   Address each clash separately. Explain what is being contested, where the sides disagree, who substantiated their case, and who failed to answer.
   If the verdict explicitly treats a clash as tied or says the judge cannot decide, state faithfully: "I judge this clash to be tied" or "I cannot determine the outcome of this clash." Do not force a winner.
   You may use first-person judgments such as "I need to assess this" or "They failed to answer this clash."
7. **Penultimate paragraph—overall assessment:**
   Begin with: "Taking everything together, let me take stock."
   Review how many points each side successfully established. Incorporate the meaning of the sentence under `[Calibration Guideline]` here, using your own words rather than copying it verbatim.
8. **Final paragraph—use this exact closing pattern, filling in the placeholders:**
   Finally, [one or two sentences summarizing the direction and degree of the preference]. Let me think: I need to provide an outcome preference. The closer it is to 0, the more I favor the negative; the closer it is to 1, the more I favor the affirmative. Therefore, I will assign p a value of [p value].

Mandatory constraints.

- All content must come from `[Judge's Verdict]`. Do not introduce arguments, examples, data, or names absent from the verdict.
- State the overall conclusion only in the final paragraph. Earlier paragraphs must not contain premature conclusions such as "I vote for X."
- Use plain prose paragraphs separated by one blank line. Do not use headings, numbered lists, or Markdown formatting in the generated text.
- Target 700–900 characters. If the source verdict is sparse, prefer a substantive 500-character response to padded text.
- Use the p value supplied in the input file, formatted to two decimal places.

File output. Use the `Write` tool to write a single JSON object to:

```
${F}\rewrite_outputs\\${mid}.json
```

Use the following structure, placing `basis` before `p`:

```
{
  "match_id": "${mid}",
  "think": "<complete chain-of-thought text>",
  "basis": "<concise grounds for the judgment>",
  "p": 0.72
}
```

Table A8: Public components of CCDD. Original match results and appeal reassessments remain distinct; generated summaries are identified separately from human records.

Public component	Contents
Debate text	De-identified transcripts and their structural organization
Individual judge scores	Side-specific stage assessments with adjudication provenance
Match results and awards	Original-panel winner, selected best debater, and underlying judge decisions and award ballots
Judge commentary	Text of recorded adjudication explanations
Appeal records	Reassessments and any outcome changes linked to original matches
Generated summaries	Separately identified model-generated commentary summaries

The `basis` field must contain 80–100 characters and must not begin with "I vote for." Replace the illustrative numeric value with the supplied p value.

Final response. Reply with exactly one line:

```
OK ${mid.slice(0, 10)}
```

E.3 Release Authorization, Privacy, and Withdrawal

Each competition's registration rules included a data-use clause. Figure A2 provides an anonymized illustration of the registration form used for Competition 5 and marks the location of the data-use clause within the form. Under this clause, submission of a team registration form and completion of registration constituted acknowledgment and agreement that the organizing committee could use, retain, study, and redistribute competition audio recordings and related competition information for noncommercial purposes, including appeal handling, academic research, and teaching. The clause also permitted the committee to authorize reasonable third-party use within noncommercial academic research, committed it to reasonable de-identification measures to protect personal privacy, and excluded commercial use. Separately, the organizing committee explicitly authorized the research team's use and redistribution of the competition records for this study.

CCDD is publicly available through the dataset website identified in the abstract. The public package contains de-identified debate text, individual judge scores, commentary, appeal records, and generated summaries, as listed in Table A8. The released text and annotations underwent de-identification and manual review.

A data-withdrawal mechanism is provided. Audio access remains under consideration; video recordings and source notes are not committed public components.

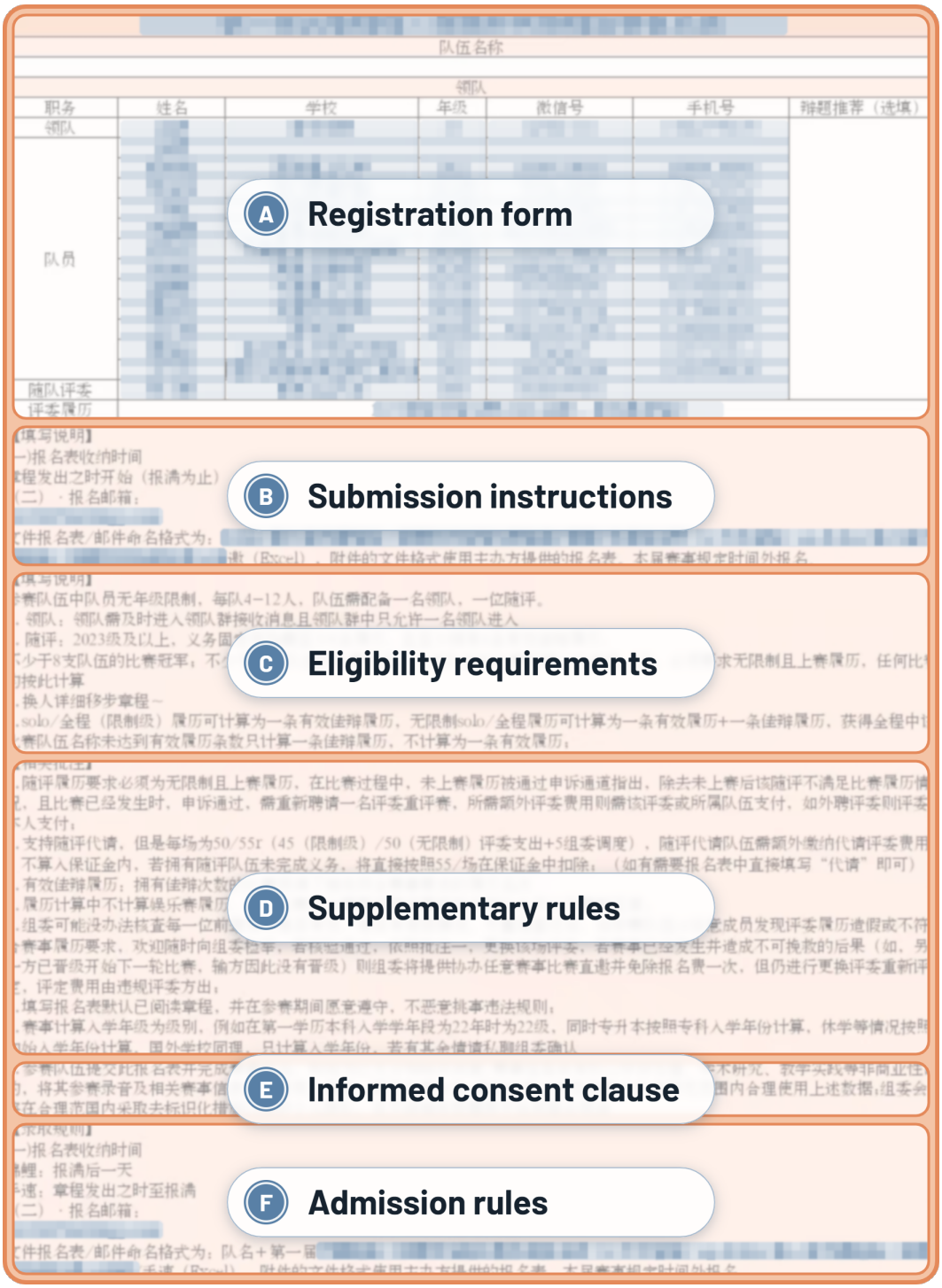

Figure A2: An anonymized illustration of the Competition 5 registration form. Personal and identifying information has been removed or obscured. The highlighted region indicates where the data-use clause appears in the registration form; the figure is included solely to document the presence and placement of this clause.