R-Debater: Retrieval-Augmented Debate Generation through Argumentative Memory

Maoyuan Li*
Wuhan College of Communication
Wuhan, China
mli302@aucklanduni.ac.nz

Zhongsheng Wang*
University of Auckland
New Zealand
zhongsheng.wang@auckland.ac.nz

Haoyuan Li
University of Auckland
Auckland, New Zealand
hli962@aucklanduni.ac.nz

Jiamou Liu
University of Auckland
Auckland, New Zealand
jiamou.liu@auckland.ac.nz

ABSTRACT

We present R-Debater, an agentic framework for generating multi-turn debates grounded in argumentative memory. Drawing on principles from rhetoric and memory studies, the framework conceptualizes debate as a dynamic process of retrieving and adapting prior arguments to maintain stance consistency, respond to opposing claims, and support assertions with evidence. Specifically, R-Debater integrates a debate knowledge base for retrieving case-like evidence and prior debate moves with a role-based agent that composes coherent utterances across turns. We evaluate R-Debater on two tasks using ORCHID debates: next-utterance generation (assessed by InspireScore) and multi-turn adversarial simulation (evaluated by Debatrix). Our framework outperforms strong LLM baselines in both settings. Human evaluation with 20 experienced debaters further confirms its consistency and evidence use, demonstrating that retrieval grounding combined with structured planning yields more faithful, stance-aligned, and coherent debates. Code and supplementary materials are available at https://github.com/Maoyuan-li/R-debater.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Discourse, dialogue and pragmatics**; **Knowledge representation and reasoning**; *Machine learning*; *Intelligent agents*; *Language resources*; • **Information systems** → *Information extraction.*

KEYWORDS

Computational Argumentation, Multi-Turn Debate Generation, Retrieval-Augmented Generation, Agentic AI

ACM Reference Format:
Maoyuan Li*, Zhongsheng Wang*, Haoyuan Li, and Jiamou Liu. 2026. R-Debater: Retrieval-Augmented Debate Generation through Argumentative Memory. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/XWXH6253

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/XWXH6253

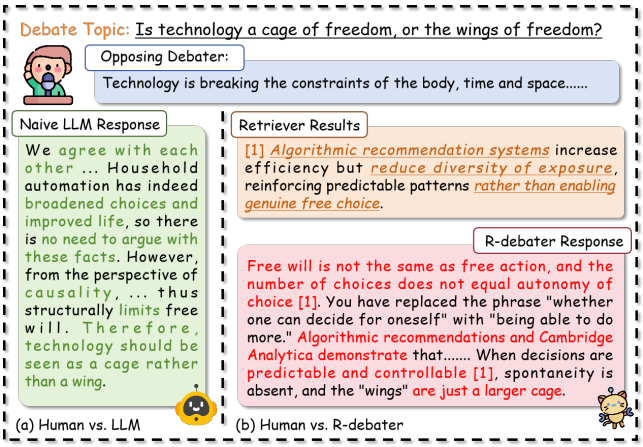

Figure 1: Humans debate by integrating external knowledge and debate strategies, while LLMs often fail to connect retrieval with argumentation. Our task aims to bridge this gap by enhancing LLMs' ability to leverage retrieved debate history and strategies more effectively.

1 INTRODUCTION

Competitive debate distills public reasoning into a structured, adversarial setting in which speakers must plan multi-turn arguments, maintain a stance, and ground claims in verifiable evidence. Prior work, from end-to-end debating systems (e.g., *Project Debater* [38]) to a mature literature on argument mining [23], has advanced the field by developing pipelines for discovering and organizing arguments. Meanwhile, Large Language Models (LLMs) have markedly improved open-ended dialogue and text generation [22, 34], but when faced with competitive debate, their outputs often remain fluent yet shallow: They are insufficiently grounded, weak on stance fidelity, and brittle over many turns. Therefore, one should pivot to the capacities that make argumentation stances faithful and durable across turns.

Building on this insight, we treat argumentation not only as a computational task but as a cognitive and rhetorical process

* Equal contribution.

grounded in human memory and discourse. As Vitale notes, public argumentation operates through a *rhetorical-argumentative memory* that recycles and reformulates prior persuasive strategies in new situations [45]. Likewise, Aleida Assmann in her acclaimed theory of cultural memory conceptualizes collective memory as a dialogic, dynamic reconstruction of past discourses rather than a static archive [4]. These perspectives motivate our central premise: *effective debate generation should retrieve and recontextualize argumentative memory, producing statements that are grounded in prior reasoning and robust rhetorical patterns.*

Retrieval-augmented generation (RAG) is a natural mechanism for recalling prior cases and evidence [26, 37], and has proved useful in QA and open-domain assistants [21]. However, mainstream RAG stacks optimize for short factual responses and struggle with debate's structured, adversarial recontextualization, where models must reconcile retrieved material with evolving discourse and opponent claims. Recent analyses also show a persistent "tug-of-war" between a model's internal priors and retrieved evidence, yielding faithfulness and coverage errors without explicit control [50, 54]. Concurrently, Agentic AI frameworks provide machinery for planning and tool use [27, 51], yet they rarely encode argumentative schemes [48] or dialogic memory needed to preserve stance and reasoning coherence across turns [56].

This paper addresses this gap by conceptualizing argumentative memory in computational terms. Specifically, we treat retrieval as the mechanism for recalling prior discourse and role-based planning as the mechanism for reframing it into new argumentative settings. Tackling this question raises three central challenges: First, generated statements must incorporate appropriate argumentative schemes and logical structures to be persuasive and coherent. Second, the system must maintain stance fidelity throughout the debate, ensuring that rebuttals are grounded in opponents' claims while avoiding hallucinations or sycophantic tendencies. Finally, the model must retrieve high-quality, context-relevant debate materials and strategically integrate them into statement generation rather than reproducing them verbatim.

To address these challenges, we propose **R-Debater**, a framework that integrates retrieval-augmented reasoning with role-based planning for generating debate statements. Unlike prior systems, R-Debater leverages the entire debate history to generate stance-specific and rhetorically coherent utterances, bridging the gap between rhetorical theory and generative modeling.

We conduct extensive experiments on the ORCHID [59] dataset, which comprises over 1,000 formal debates across multiple domains. Evaluation demonstrates that R-Debater achieves substantial gains in both logical fidelity and rhetorical persuasiveness over strong LLM and RAG baselines. In particular, it delivers near-perfect logical coherence and markedly higher factual grounding, while human experts prefer its generated debates in over **75%** of pairwise comparisons. These results confirm that R-Debater not only improves measurable debate quality but also aligns closely with expert judgements, validating its reliability and interpretability.

Overall, our main contributions are as follows:

- We present, to the best of our knowledge, the first systematic investigation of debate statement generation with LLMs under realistic multi-turn settings.
- We propose **R-Debater**, a novel framework that integrates argumentation structures, debate strategies, stance fidelity mechanisms, and retrieval-based case reasoning to address the core challenges of debate generation.
- Extensive empirical validation on debate datasets demonstrates that R-Debater significantly outperforms strong LLM and Naive RAG baselines in terms of factual accuracy, stance consistency, and cross-turn coherence.

2 RELATED WORK

2.1 Computational Argumentation

Argumentation research has long-standing roots in symbolic reasoning and formal logic [11, 12, 40–43, 47], with early computational approaches offering transparency and rigor but relying on brittle, rule-based knowledge bases that limit scalability to open-domain debates [7, 9, 39, 60]. These symbolic methods also often lack expressiveness in realistic discourse.

To address these limitations, subsequent work in computational argumentation has leveraged advances in natural language processing (NLP) [6, 15, 24, 44]. Retrieval-based pipelines such as CANDELA [16, 17] adopt a retrieval-planning-realization process, aggregating evidence from multiple sources to improve the factual grounding and specificity. Other studies enhance neural models with argumentation knowledge graphs [2], which encode structured relationships between claims, premises, and stances, thereby improving both credibility and topical relevance. Beyond single-turn generation, hierarchical dialogue frameworks [35, 36] integrate topic analysis, stance retrieval, and user feedback to produce more adaptive multi-turn interactions. These advances substantially enhance fluency, evidence utilization, and contextual relevance. However, most existing approaches still treat each debate as an isolated reasoning task rather than a reactivation of prior argumentative patterns. They focus on structural modeling but lack cross-turn continuity, making it difficult to capture genuine debate dynamics.

Recent studies highlight that LLMs possess strong capabilities in long-context understanding and coherent discourse generation [3, 10, 20, 31]. These advances open the door to debate systems that move beyond turn-local reasoning and model the holistic argumentative flow. Building on this line, R-Debater aims to integrate external knowledge with full-dialogue modeling to produce strategically grounded and rhetorically coherent argumentative responses.

2.2 Retrieval-Augmented and Agentic AI

LLMs have demonstrated strong long-text generation capabilities across diverse NLP tasks [22, 34, 55, 58], yet they remain prone to hallucination and lack the factual grounding required for credible debate-oriented applications [18]. Retrieval-augmented generation (RAG) mitigates these issues by grounding outputs in external knowledge sources [14, 25], and RAG-based dialogue systems [28, 29, 33] have shown improved factuality and coherence. However, these systems primarily target informational QA or open-domain chit-chat [8, 57], failing to address the complex requirements of stance construction and rebuttal across multiple debate turns. In this sense, RAG can be regarded as an early computational analogue of *rhetorical memory*, capable of retrieving and reorganizing prior

discourses to support new argumentative goals, yet still lacking the dynamic recontextualization central to real debates.

Parallel to RAG, research on Agentic AI explores decomposing complex tasks into interacting role-based agents. Frameworks such as AutoGen [52] and Agent4Debate [13] demonstrate that specialized agents can coordinate, critique, or compete to achieve more robust and strategically informed outputs [8, 57], but the absence of explicit evidence grounding often limits their argumentative credibility. Viewed through the lens of dialogic memory, such systems simulate interactional structure yet fail to capture the cumulative recall of argumentation history.

In contrast, R-Debater builds on both paradigms by explicitly operationalizing *argumentative memory* through retrieval-augmented and role-based mechanisms. It dynamically retrieves debate-specific evidence, assigns strategic roles, and recontextualizes past argumentative moves to generate text that is both stance-specific and globally coherent in rhetorical and strategic senses, thereby achieving a deeper integration of evidence retrieval and rhetorical reasoning within complex multi-turn debates.

3 TASK FORMULATION

Dialogue-based argumentation. A debate is a structured, adversarial dialogue in which two parties alternate turns to advance their own stance while challenging the opponent. Unlike open-domain dialogues, a debate imposes strict constraints: utterances must be *stance-consistent* and *argumentatively relevant*, either reinforcing one's position or rebutting opposing claims.

Formally, we represent a *(dialogue-based) argumentation* A as a pair $A = (T, H)$, where $T = \{t_1, t_2\}$ denotes the two opposing stances (pro and con) on a given topic, and the ordered sequence

$$H = (u_1, u_2, \ldots, u_n)$$

represents an n-turn argument history (where $n \in \mathbb{N}$ is even). Here, u_j denotes the utterance at turn j, which is a natural-language text. We assume the pro side speaks first, so the set $P_n = (u_j)_{j\equiv 1 \pmod 2} = (u_1, u_3, \ldots, u_{n-1})$ contains all utterances produced by the pro side, while $C_n = (u_j)_{j\equiv 0 \pmod 2} = (u_2, u_4, \ldots, u_n)$ containing utterances produced by the con side.

We introduce a single-utterance predicate $\phi(u, s)$ that returns `True` if the utterance u either explicitly supports the stance s or rebuts the opposing stance. We require that in the argument history H, every utterance u_i either satisfies $\phi(u_i, t_1)$ if i is odd or satisfies $\phi(u_i, t_2)$ if i is even, but not both.

Debate and argumentative memory. From the perspective of rhetorical and dialogic memory [45], each utterance not only presents a local claim but also reconstructs argumentative contexts to sustain stance coherence across turns. We thus view a debate as an iterative reconstruction of prior reasoning contexts rather than a sequence of isolated utterances: At each turn, a speaker retrieves and adapts fragments of earlier arguments, evidence, or rhetorical strategies to sustain coherence and respond to the opponent's reasoning. We refer to this process as *argumentative memory*. We now formulate a process in which argumentative memory is instrumental in the construction of utterances in a debate.

We assume to have a pre-curated set $\mathcal{D}$ of latent memories of previously encountered argumentative patterns. Given an ongoing (partial) dialogue history $H_j = (u_1, \ldots, u_j)$, where $j < n$ and stances T, the debater retrieves a subset of relevant memory items $\mathcal{E}_j$ and synthesizes a stance-consistent next utterance u_{j+1}:

$$\mathcal{E}_j = \mathcal{R}(H_j, \mathcal{D}),\ u_{j+1} = \mathcal{M}(T, H_j, \mathcal{E}_j). \tag{1}$$

where $\mathcal{R}$ denotes the *retrieval* function that recalls an early memory $\mathcal{E}_j$, and $\mathcal{M}$ denotes the *reasoning* function that adapts $\mathcal{E}_j$ to the current debate context to produce the next utterance u_{j+1}.

R-Debater Task Definition. We define debate utterance generation as a *next-utterance prediction* problem that requires stance fidelity and argumentative relevance while leveraging argumentative memory and rhetorical strategies. Suppose that we have an argumentative memory set $\mathcal{D}$. Given stances T, partial history H_j with even $j < n$, and opponent C, the problem seeks a pair of retrieval and reasoning functions (R, M) that allows us (i.e., Pros $\mathcal{P}$) to generate a valid argumentation with the opponent. In other words, applying R and M repeatedly to produce the sequence of utterances $H = (u_1, u_2, \ldots, u_n)$, where each utterance in $C_n = (u_2, u_4, \ldots)$ is produced by the opponent C, and u_{i+1} is produced by $\mathcal{M}(T, H_i, \mathcal{R}(H_i, \mathcal{D}))$ for even $i \geq j$. Then the pair (T, H) meets the definition of an argumentation above, i.e., $\phi(u_{i+1}, s) = \texttt{True}$ for all even $i \geq j$.

4 METHODOLOGY

We propose **R-Debater**, as illustrated in Figure 2, a retrieval- augmented and role-based framework that models debate as the dynamic reconstruction of argumentative memory. It consists of three pipelines: (1) **Database Construction**, which segments debate transcripts into utterance-level units and annotates argumentative information; (2) **Debate Data Retrieval**, which leverages this database to obtain stance-relevant evidence while also summarizing the logical gaps and weaknesses in the opponent's reasoning; and (3) **Debate Generation**, which produces stance-consistent, rhetorically coherent utterances through iterative verification.

4.1 Database Construction

We construct a large-scale debate knowledge base from authentic transcripts collected across diverse public sources. Each transcript is processed by a rule-based parser that segments dialogue into utterance-level units $U = \{u_1, \ldots, u_n\}$, removing moderator turns and meta-commentary, retaining only argumentative content. Each u_i corresponds to a single debater's statement at a particular debate stage. A pre-trained encoder f_{emb} then maps every utterance into a dense representation $\mathbf{e}_i = f_{\text{emb}}(u_i)$, yielding pairs $(u_i, \mathbf{e}_i)$ that capture semantic meaning and stylistic features for subsequent retrieval and reasoning.

Argumentation Schemes. Following Walton et al. [48], an *argumentation scheme* is a stereotypical reasoning pattern that shows how premises support a conclusion and poses critical questions for evaluating validity. These schemes are widely used in computational argumentation to represent and assess the plausibility of natural arguments [5, 46]. As Walton et al. stress in their seminal monograph *Argumentation Schemes* [48], schemes serve as dialogical templates that capture recurring structures of everyday reasoning and provide systematic criteria for evaluating strength. In this work, we employ seven commonly adopted schemes from [48]:

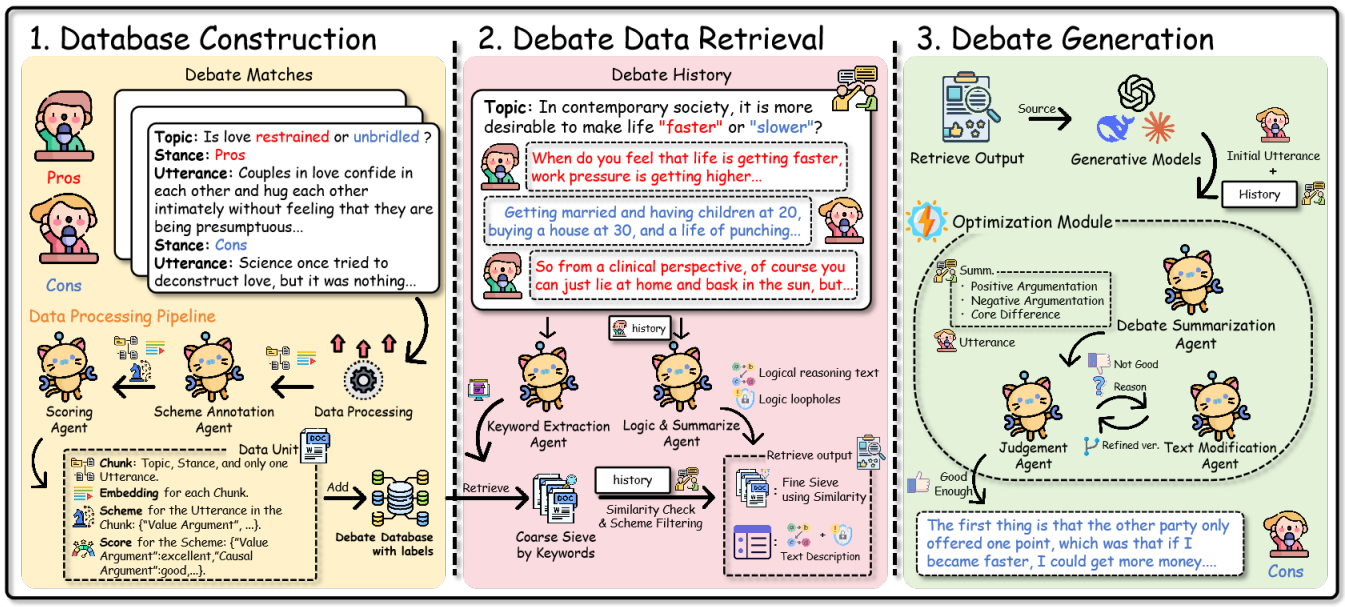

Figure 2: R-Debater consists of three modules: (1) Database Construction, (2) Debate Data Retrieval, and (3) Debate Generation. Together, they enable the dynamic reconstruction of argumentative memory for generating next-turn debates.

example-based argumentation, *expert opinion*, *positive consequence*, *negative consequence*, *causal argumentation*, *analogical argumentation*, and *value-based argumentation*, which together provide the reasoning foundation for our framework.

Scheme Annotation and Scoring. The **Scheme Annotation Agent** uses prompt-based LLM reasoning to assign an argumentation-scheme set $\mathcal{S}_i$ to each utterance. Then, the **Scoring Agent** performs few-shot evaluation with exemplars $\mathcal{X}_{\text{few}} = \{e_{\text{poor}}, e_{\text{general}}, e_{\text{good}}, e_{\text{excellent}}\}$ and outputs *per-scheme* qualitative labels:

$$\text{Score}_i(S) \in \{\text{poor}, \text{general}, \text{good}, \text{excellent}\} \text{ for } S \in \mathcal{S}_i.$$

and $\text{Score}_i(S) = \text{none}$ if $S \notin \mathcal{S}_i$. The scoring criteria follow Walton et al.'s guidelines for evaluating argumentation schemes, emphasizing whether the premises are acceptable, the scheme is applied appropriately, and the critical questions are adequately addressed. Thus, each label reflects the appropriateness and rhetorical effectiveness of the scheme usage in context.

Finally, each database record is represented as $r_i = (u_i, \mathbf{e}_i, \mathcal{S}_i, \text{Score}_i)$, constituting the enriched knowledge base $\mathcal{D} = \{r_i\}_{i=1}^{N}$ that supports retrieval and generation. In subsequent stages, $\mathcal{D}$ provides a foundation for stance-aware evidence selection and supports generation modules in reconstructing contextually coherent, strategically grounded arguments.

4.2 Debate Data Retrieval

Given an ongoing debate and a target stance $t_i \in T$, the retrieval module extracts stance-relevant argumentative context to support the generation of the next utterance. It operates through two complementary components: a *Logic & Summarization Agent*, which identifies logic loopholes in the opponent's arguments, and a *Keyword Extraction Agent*, which derives salient keywords from the current debate history. The extracted keywords are then passed to a basic retriever to retrieve coarse-grained evidence from the knowledge base $\mathcal{D}$, as constructed in Section 4.1. A fine-grained filtering stage then computes semantic similarity between the retrieved candidates and the current debate history to retain the most contextually relevant exemplars.

Logic & Summarization Agent. This component analyzes the accumulated utterances from the opposing side to detect logic loopholes, contradictions, and unsupported assumptions across the entire debate history, rather than at the single-utterance level. Inspired by the SymbCoT [53] paradigm, we design a prompt-based logical decomposition process that guides LLMs to transform the opponent's collective statements into pseudo-first-order predicates and simulate lightweight symbolic reasoning over them.

Formally, given the opponent's utterance sequence defined as $H_n^{\text{opp}} = \{u_1^{\text{opp}}, \ldots, u_t^{\text{opp}}\}$, the agent constructs a set of pseudo–first-order predicates and derives natural-language reasoning chains:

$$\mathcal{P}_{\text{opp}} = \bigcup_{i=1}^{t} f_{\text{pred}}(u_i^{\text{opp}}), \ \tau_{\text{opp}} = f_{\text{infer}}(\mathcal{P}_{\text{opp}}),$$

which summarizes the inferred argumentative logic throughout the opponent's discourse.

The first step, *predicate extraction*, converts each utterance into a symbolic form that captures its structure and causal relations.

We implement f_{pred} as a prompt-based LLM agent that extracts the opponent's major claims into pseudo–first-order predicates (one predicate per line), focusing on causal, contrastive, and quantitative relations while omitting rhetorical filler.

Once the individual predicates are extracted, the next step is to reconstruct them into a coherent reasoning chain that represents the opponent's overall argumentative flow.

Then, f_{infer} linearizes and verbalizes the predicate set into a neutral reasoning chain τ_{opp} that reconstructs the opponent's argumentative flow without producing rebuttals.

Logical inconsistencies or invalid reasoning patterns in τ_{opp} are then identified as $\mathcal{L} = f_{\text{logic}}(\tau_{\text{opp}})$. where $\mathcal{L}$ denotes a collection of natural-language control signals that provide interpretable guidance for generating targeted rebuttals and reinforcing stance fidelity. In this design, only $\mathcal{P}_{\text{opp}}$ corresponds to a pseudo–first-order logical representation, while τ_{opp} and $\mathcal{L}$ remain natural-language constructs, enabling symbolic-style interpretability without explicit theorem proving.

All three functions (f_{pred}, f_{infer}, f_{logic}) are implemented with lightweight prompt templates to ensure consistent outputs that can be reliably reused as control signals.

Finally, f_{logic} acts as a verifier that flags contradictions, unsupported assumptions, circular reasoning, and overgeneralization in τ_{opp}, producing short natural-language control signals $\mathcal{L}$ used to guide targeted rebuttals.

Detailed prompt templates for f_{pred}, f_{infer}, and f_{logic}, together with end-to-end execution examples, annotation guidelines, and intermediate outputs (e.g., predicate sets, reasoning chains, and control signals), are provided in our public repository (see link in the abstract).The repository also includes structured JSON examples illustrating predicate extraction, reasoning-chain construction, and logic-flaw identification in real debate cases.

Keyword Extraction Agent. This component operates in parallel with the Logic & Summarization Agent and focuses on retrieving strategically relevant exemplars from the debate knowledge base $\mathcal{D}$. A prompt-based LLM first extracts salient keywords $\mathcal{K} = \{k_1, \ldots, k_m\}$ from the partial debate history $H_j = \{u_1, \ldots, u_j\}$. These keywords are used in a rule-based coarse retrieval stage to filter database records with explicit keyword matches:

$$\mathcal{E}_{\text{coarse}} = \left\{ r_j \in \mathcal{D} \;\middle|\; \texttt{rule\_match}(r_j, \mathcal{K}) \right\}.$$

Subsequently, a fine-grained re-ranking is performed by encoding the current debate history into a single representation $\mathbf{v}_H = f_{\text{emb}}(H_j)$ and computing cosine similarity against the embeddings of the coarse candidates $\mathbf{e}_j = f_{\text{emb}}(r_j)$:

$$\text{sim}(\mathbf{v}_H, \mathbf{e}_j) = \frac{\mathbf{v}_H \cdot \mathbf{e}_j}{\|\mathbf{v}_H\| \, \|\mathbf{e}_j\|}, \; \mathcal{E} = \texttt{Top-}k\big(\{(r_j, \text{sim}(\mathbf{v}_H, \mathbf{e}_j)) \mid r_j \in \mathcal{E}_{\text{coarse}}\}\big).$$

Each retrieved chunk $r_i \in \mathcal{R}$ is annotated with a set of argumentation schemes $\mathcal{S}_i$ and corresponding quality labels Score_i inherited from database construction. For aggregation, each label is assigned a numeric value (*poor*=1, *general*=2, *good*=3, *excellent*=4), while schemes absent in a chunk receive 0. For each scheme type $S \in \mathcal{S}$, we then compute its average score across the top-k retrieved candidates: $\bar{s}(S) = \frac{1}{k} \sum_{r_i \in \mathcal{E}} \text{Score}_i(S)$. Schemes with $\bar{s}(S) > 2$ (above the *general* threshold) are retained as high-quality argumentative priors, and the corresponding chunks containing these schemes are selected as exemplars. The resulting prior sets $(\mathcal{E}^{\text{prior}}, \mathcal{S}^{\text{prior}})$ jointly form the output of the Keyword Extraction Agent and provide strategy-oriented evidence for subsequent generation.

4.3 Debate Generation

Given the retrieved strategic evidence and logical signals from the previous stage, the model first generates an initial utterance $\tilde{u}^{(0)}$ conditioned on the debate history H_n and the designated stance s:

$$\tilde{u}^{(0)} = M(H_n, \mathcal{L}, \mathcal{E}^{\text{prior}}, \mathcal{S}^{\text{prior}}, s).$$

The *Debate Summarization Agent* then analyzes both the accumulated debate history and the newly generated utterance to produce a structured summary covering the overall debate overview, supporting and opposing arguments, and the core points of divergence. These structured summaries, together with the candidate utterance, are passed to the judgement agent.

The *Judgement Agent* evaluates the candidate utterance under three criteria: stance faithfulness, argumentative relevance, and scheme compliance, and then outputs a binary judgement signal $J^{(t)} \in \{0, 1\}$ along with textual feedback $\rho^{(t)}$ explaining any detected violation. If the judgement is positive ($J^{(t)} = 1$), the utterance is accepted as the final output. Otherwise, the *Text Modification Agent* receives both the utterance and the feedback and produces a revised version: $\tilde{u}^{(t+1)} = f_{\text{modify}}\big(\tilde{u}^{(t)}, \rho^{(t)}\big)$. This process repeats until all evaluation criteria are satisfied, and the utterance u^* is adopted as the debate output for the turn.

5 EXPERIMENTS

5.1 Experimental Setup

Dataset. We conduct all experiments on the ORCHID [59] debate dataset, which contains high-quality transcripts collected from formal debate tournaments. Each record contains the debate topic, participating institutions (e.g., Tongji University), speaker identifiers (e.g., Pro-1, Con-2), explicit stance tags (*pro*/*con*), and detailed utterance texts. To ensure data consistency, we exclude free-debate or informal discussion sessions during preprocessing. On average, each debate contains about ten rounds, with each utterance consisting of 1,000–1,500 Chinese characters. Since the transcripts originate from standardized competitions with explicit stance and role annotations, ORCHID provides consistent quality and minimizes potential bias across topics and domains. From the most recent five years of ORCHID, we extract a total of 1,134 debates, of which 1,000 are used to construct the retrieval corpus. For evaluation, we adopt a curated set of 32 debates covering seven representative domains: *Society & Livelihood*, *Politics & Governance*, *Economy & Development*, *Technology & Future*, *Morality & Values*, *Education & Youth*, and *International Relations*, ensuring topical diversity and no overlap with the retrieval corpus.

Compared Methods. We compare R-Debater against three baselines. (i) **LLM:** a direct prompting approach that generates debate utterances without any external retrieval materials or structured planning. (ii) **Naive RAG:** a retrieval-augmented baseline that simply concatenates retrieved knowledge to the prompt, without filtering or explicit reasoning control. (iii) **Agent4Debate:** a multi-agent coordination framework that assigns distinct roles for analysis, evidence gathering, and rebuttal generation. All baselines are instantiated with **GPT-4o** [19], **DeepSeek-V3** [32], and **Claude-3.7-sonnet** [1] under a zero-shot setting, setting the temperature to 0.2 and without fine-tuning, to ensure fairness and reproducibility.

Evaluation. We adopt two methods: **InspireScore** [49] for utterance-level assessment and **Debatrix** [30] for debate-level evaluation. InspireScore measures utterance quality along three dimensions: Subjective, Logic, and Fact. The *Subjective* dimension

captures persuasiveness and rhetorical appeal, including emotional appeal, clarity of the argument, argument arrangement, and topic elevation. *Logic* assesses reasoning validity and stance consistency, ensuring arguments maintain coherent reasoning chains, avoid contradictions, and align with both the assigned position and debate motion. *Fact* evaluates factual grounding and evidence usage, considering correctness, credibility, and contextual relevance of cited information. We report per-dimension results and the aggregated InspireScore averaged across runs. To complement utterance-level analysis, Debatrix evaluates performance by scoring each utterance across three criteria: **Argument (A)**, **Source (S)**, and **Language (L)**, then aggregates them into an **Overall (O)** score. The final outcome reflects which side achieves superior argument quality, evidence integration, and linguistic fluency, providing a measure of overall debate effectiveness.

Experimental Design. We design three complementary experiments to evaluate R-Debater from different perspectives: micro-level next-utterance generation, macro-level adversarial debate simulation, and internal agent–expert alignment. They assess both external performance and internal reliability of the framework.

Experiment 1: Next-Utterance Generation. We assesses the model's ability to generate the next debate utterance, given a truncated dialogue history and stance. Each model produces one continuation, which is evaluated using InspireScore in terms of linguistic fluency, argumentative soundness, and stance consistency.

Experiment 2: Adversarial Debate Simulation. To assess global debate competence, we conduct simulations in which R-Debater and Agent4Debate engage in multi-turn debates under opposing stances. We initialize each simulation from a balanced stage and allow the debate to proceed until both sides reach a conclusion. The resulting trajectories are evaluated with the Debatrix metric, which measures argument quality, evidence integration, and rhetorical coherence across the debate flow.

Experiment 3: Expert Alignment Evaluation. To validate interpretability and trustworthiness of the framework, we evaluate alignment between internal agents and human experts. Two components are assessed: **Scheme Annotation Agent**, which labels argumentation schemes, and **Scoring Agent**, which rates argumentative quality. For Scheme Annotation Agent, we compute Jaccard Index, Precision, Hamming Loss, Cohen's κ, and Krippendorff's α to capture surface-level and inter-rater consistency. For Scoring Agent, we examine Pearson, Spearman, and Kendall correlations between automatic and expert scores, reflecting absolute, relative alignment.

6 RESULTS AND ANALYSIS

6.1 Single-turn Debate Evaluation

As shown in Table 1, R-Debater consistently surpasses all baselines in single-turn debate performance across different foundation models. It achieves the highest scores in all InspireScore dimensions: *Subjective*, *Logic*, and *Fact*, as well as in the overall composite score. The most notable gains appear in the logical dimension, where R-Debater approaches a near-perfect score ($\approx$1) across all models, indicating a strong ability to maintain reasoning coherence and stance fidelity. Improvements in both the factual and subjective dimensions further demonstrate that integrating retrieval-augmented argumentative memory and multi-agent collaboration enhances both factual grounding and rhetorical persuasiveness, resulting in more convincing and well-structured single-turn arguments.

6.2 Multi-turn Competitive Debate Evaluation

We further assess R-Debater in full multi-turn debate scenarios using the Debatrix evaluation framework. As shown in Table 2, R-Debater consistently surpasses the Agent4Debate baseline across all evaluation dimensions and foundation models. On GPT-4o, it achieves substantial gains in Source Credibility, Language quality, Argument soundness, and Overall performance, while comparable or even greater improvements are observed on DeepSeek-V3. These results demonstrate that R-Debater not only sustains its superiority in single-turn reasoning but also exhibits stronger strategic consistency and cross-turn coherence, highlighting the effectiveness of retrieval-augmented argumentative memory and structured multi-agent reasoning in extended debate settings.

Table 1: Single-turn debate evaluation across models and methods using InspireScore.

Base Model	Method	Subjective	Logic	Fact	InspireScore
GPT-4o	LLM	0.757	0.940	0.226	0.641
	Naive RAG	0.761	0.939	0.609	0.770
	Agent4Debate	0.821	0.913	**0.631**	0.783
	R-Debater (Ours)	**0.842**	**0.997**	0.627	**0.822**
DeepSeek-V3	LLM	0.783	0.877	0.158	0.591
	Naive RAG	0.732	0.927	0.582	0.747
	Agent4Debate	0.837	0.893	0.590	0.773
	R-Debater (Ours)	**0.866**	**0.997**	**0.594**	**0.819**
Claude-3.7-sonnet	LLM	0.780	0.971	0.190	0.647
	Naive RAG	0.772	0.960	0.567	0.766
	Agent4Debate	0.765	0.917	0.618	0.767
	R-Debater (Ours)	**0.789**	**0.985**	**0.701**	**0.830**

Table 2: Multi-turn Competitive Debate Evaluation across models and methods using Debatrix.

Model	Framework	Debatrix			
		S	L	A	O
GPT-4o	R-Debater (ours)	1.13	1.16	1.26	1.23
	Agent4Debate	0.87	0.84	0.74	0.77
Deepseek-V3	R-Debater (ours)	1.25	1.11	1.19	1.25
	Agent4Debate	0.75	0.89	0.81	0.75

6.3 Expert Alignment Evaluation.

Beyond baseline comparisons, we assess the alignment of R-Debater's internal agents with expert judgements.

To measure how faithfully the internal modules of R-Debater approximate expert reasoning, we evaluate both the **Scheme Annotation Agent** and the **Scoring Agent** against gold-standard expert annotations. Table 3 presents the agreement metrics for the annotation task and the correlation results for scoring alignment.

Annotation Agreement with Experts. As shown in Table 3 (a), the Scheme Annotation Agent achieves strong agreement with expert annotations across multiple metrics. The Jaccard Index and Precision demonstrate that the agent captures the majority of expert-labeled schemes, while low Hamming Loss indicates that misclassifications are rare. Furthermore, Cohen's κ (both micro and macro) and Krippendorff's α reveal substantial inter-rater reliability, confirming that the automated annotation process can approximate expert-level scheme identification with high fidelity. These results

Table 3: Evaluation results of the Scheme Annotation Agent and Scoring Agent.

(a) Scheme Annotation Agent vs. expert annotations.

Metric	Value
Jaccard Index (avg)	0.7366
Precision (avg)	0.8225
Hamming Loss (avg)	0.1291
Cohen's κ (micro)	0.7229
Cohen's κ (macro)	0.6964
Krippendorff's α (nominal)	0.7230

(b) Scoring Agent: correlation with expert scores.

Metric	Value
Pearson	0.6356
Spearman	0.6533
Kendall	0.5880

validate that the retrieval-augmented annotation mechanism produces stable and interpretable scheme assignments.

Scoring Alignment with Experts. As shown in Table 3 (b), the Scoring Agent exhibits high correlation with expert scores (Pearson, Spearman, Kendall), preserving both absolute values and relative rankings. This suggests it can serve as a credible proxy for expert evaluation, reducing the reliance on costly human annotation. Such alignment also indicates that the scoring module provides a consistent and scalable way to benchmark debate quality, ensuring that R-Debater's evaluations remain interpretable and trustworthy across large-scale experiments.

Together, these findings indicate that R-Debater's internal modules closely align with expert reasoning, thereby enhancing the framework's interpretability and trustworthiness.

6.4 Case Study

To illustrate the qualitative advantages of R-Debater, we conducted a case study using data recorded from a real debate experiment in Experiment 2, with the topic "Should society encourage Generation Z to 'rectify' workplace culture?" R-Debater and Naive RAG were assigned opposite stances under the same historical debate context and tasked with generating the next utterance for their respective sides. Below is an excerpt for context:

> **Opponent (Con):** "Corporate culture is the foundation of organizational stability, and Gen-Z's questioning of existing systems will only cause internal disorder. Authority and discipline are guarantees of efficiency; we should not overturn the entire system because of a few isolated incidents."

Given this context, the Logic & Summarization Agent identifies a false dichotomy in the opponent's reasoning. Let τ_{opp} denote the set of natural-language reasoning chains extracted from the opponent, the logical control signals are then computed as

$$\mathcal{L} = f_{\text{logic}}(\tau_{\text{opp}}) = \{\text{"False dichotomy: (Questioning} \rightarrow \text{Disorder) is invalid."}\}.$$

This logical signal $\mathcal{L}$ guides generation of a rebuttal emphasizing the necessity of constructive criticism in organizational evolution.

Analysis. The case highlights R-Debater's ability to reconstruct *argumentative memory* across turns. By identifying the false dichotomy between questioning and disorder, R-Debater effectively integrates logical consistency with persuasive coherence. It retrieves historically similar cases from $\mathcal{D}$ that show how constructive criticism can foster positive organizational change, thereby directly countering the opponent's premise. The generated rebuttal remains focused on the topic while providing factual grounding through concrete examples.

Table 4: Qualitative comparison between R-Debater and Naive RAG on logical reasoning and stance consistency.

Module	R-Debater (Pro)	Naive RAG (Con)
Detected Logical Issue ($\mathcal{L}$)	"Conflates constructive questioning with destructive disorder; assumes all systemic challenges stem from isolated incidents."	NA.
Retrieved Evidence ($\mathcal{E}$)	Case: "Tech company (Meituan) innovation spurred by employee feedback programs that improved both culture and productivity." (Argumentation scheme: causal argumentation, score: *good*; exemplification, score: *excellent*)	Generic examples of "corporate hierarchy and discipline" unrelated to the core argument.
Generated Rebuttal (u_{m+1})	"Constructive questioning drives organizational adaptation, not disorder. Many successful companies have integrated employee feedback to improve both culture and efficiency, demonstrating that systemic evolution differs from complete overthrow."	"Encouraging confrontation may undermine teamwork and trust. Rather than opposing authority, fostering open communication would yield more sustainable improvement."

In contrast, Naive RAG fails to engage with the opponent's core reasoning chain, producing surface-level counterpoints that do not address the logical flaw in the original argument. Human evaluation further confirms that R-Debater's outputs exhibit stronger reasoning quality by addressing the opponent's argumentative structure rather than responding to superficial content.

6.5 Ablation Study

We conduct a series of ablation studies under the same setting as Experiment 1, using `gpt-4o-mini` as the base model and InspireScore for evaluation. The results in Table 5 demonstrate the complementary roles of all major components in R-Debater.

Table 5: Ablation results of R-Debater using gpt-4o-mini as the base model. "Opt. Module" / "L. & S. Agent" / "Argu. Schem." remove the Optimization Module, Logic & Summarization Agent, and Argumentation Augmentation Module, respectively.

Method	InspireScore			
	Subjective	Logic	Fact	InspireScore
R-debater	0.841	0.954	0.700	0.831
w/o Opt. Module	0.773	0.917	0.640	0.776
w/o L. & S. Agent	0.735	0.875	0.672	0.761
w/o Argu. Schem.	0.371	0.683	0.531	0.528

In all settings, the overall R-Debater framework is preserved, and only one specific component is removed to isolate its contribution. Removing the *Optimization Module* leads to a notable drop in overall InspireScore (0.831 → 0.776), primarily driven by declines in the subjective dimension. This indicates that the Optimization Module primarily enhances Subjective performance by improving

rhetorical appeal, including clarity, fluency, and emotional resonance, as well as reinforcing the firmness of stance. When the *Logic & Summarization Agent* is excluded, the decrease (0.831 → 0.761) is most pronounced in the logic dimension, confirming that this agent plays a central role in maintaining coherent reasoning chains and reducing logical inconsistencies. The absence of the Argumentation Scheme causes the sharpest degradation (0.831→0.528), even approaching the naive LLM baseline. Without this mechanism, the model struggles to organize or leverage retrieved argumentative memory, even with the support of the Logic & Summarization Agent, resulting in fragmented reasoning and factual drift.

Overall, the ablation pattern reveals a clear division of labor within R-Debater: the Optimization Module enhances rhetorical persuasion, the Logic & Summarization Agent secures reasoning coherence, and the Argumentation Scheme ensures evidence integration. Their synergy underpins the framework's superior performance in debate generation.

6.6 Human Evaluation

Given that the evaluation of debate quality is inherently subjective [38], we complement automatic metrics with a dedicated human assessment phase. To ensure fairness and reproducibility, all evaluations were conducted under strict randomization and anonymization, eliminating bias from model identity or order effects. This evaluation used the same input set as Experiment 1, and annotators assessed the debate statements generated by the four compared models. Our annotator pool consisted of 20 members from the university debate association, each with demonstrated debating experience and critical reasoning skills.

Following the InspireScore framework, the evaluation covered two categories of dimensions. The *subjective* dimensions included emotional appeal, argument clarity, argument arrangement, and topic relevance, while the *objective* dimensions assessed fact authenticity and logical validity. Each dimension was rated on a 1–10 Likert Scale to capture fine-grained differences in argumentative quality. After completing the dimension-level scoring, annotators selected their overall preferred debate statement among the four candidates, providing an additional holistic measure of preference.

Table 6 summarizes the human preference rates. Results indicate that R-Debater achieved the highest selection rate (76.32%), substantially outperforming the Agent4Debate (15.79%),LLM (7.89%), and NaiveRAG (0%). This result underscores both the factual grounding and rhetorical persuasiveness of debates generated by R-Debater.

Table 6: Human preference for debate generation models

Model	LLM	NaiveRAG	Agent4Debate	R-Debater
Choose Rate (%)	7.89	0.00	15.79	76.32

Figure 3 further reports the macro-averaged Likert ratings across the six InspireScore dimensions, providing a complementary perspective to the overall preference rates.

Annotation guidelines and anonymized evaluation templates used in the human study are released.

7 LIMITATIONS

Although R-Debater shows strong performance, several limitations remain. First, the framework inherently depends on pretrained

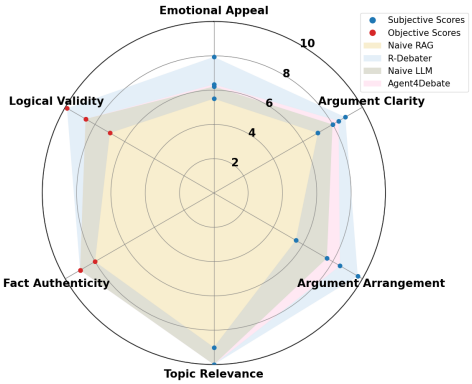

Figure 3: Human evaluation results on the six InspireScore dimensions. Each spoke shows the macro-averaged Likert rating (1-10) for a system; shaded bands indicate 95% bootstrap CIs.

LLMs, which may suffer from outdated knowledge and hallucination issues. While retrieval mitigates these problems to some extent, inconsistencies between retrieved evidence and model priors can still compromise factual reliability. Second, the retrieval process itself introduces additional latency and potential noise, and the multi-agent architecture further increases token usage and computational overhead, limiting the system's scalability for real-time or large-scale applications. Third, the timeliness and coverage of the debate database play a critical role in overall performance. If the repository is not regularly updated or lacks domain diversity, R-Debater may generate arguments that are outdated or topically misaligned. Fourth, the current implementation is primarily trained and evaluated on the ORCHID dataset. Future work includes extending the framework to cross-lingual and cross-domain settings to assess its generalization capacity, particularly in informal or multilingual debate contexts. Finally, metrics such as InspireScore and Debatrix still involve subjective components. Although human evaluation improves reliability, it remains small-scale and subjective, underscoring the need for broader and more systematic studies in future assessments.

8 CONCLUSION

We present R-Debater, a retrieval-augmented debate generation framework that integrates argumentative memory with structured multi-agent planning to generate coherent, stance-consistent, and evidence-grounded arguments across multiple turns. By reconstructing argumentative memory through dynamic retrieval and coordinating specialized agents for reasoning and refinement, the framework mitigates hallucination, reduces stance drift, and maintains logical continuity in long-horizon debates. Empirical results on the ORCHID dataset demonstrate that R-Debater consistently outperforms strong LLM-based baselines in factual accuracy, coherence, and persuasive quality, validating the effectiveness of retrieval-grounded memory reconstruction and coordinated agent reasoning. Future work will extend the framework to broader domains and multilingual settings, while further improving interpretability and efficiency for real-time interactive deployment.

REFERENCES

[1] [n.d.]. Claude 3.7 Sonnet System Card. https://api.semanticscholar.org/CorpusID:276612236

[2] Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. 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)*. 4744–4754.

[3] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems* 37 (2024), 62160–62188.

[4] Aleida Assmann. 2015. *Dialogue as a Trans-disciplinary Concept: Martin Buber's Philosophy of Dialogue and its Contemporary Reception* (1 ed.). De Gruyter. http://www.jstor.org/stable/j.ctvbj7kb3

[5] Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2025. Can Large Language Models Understand Argument Schemes?. In *Findings of the Association for Computational Linguistics: ACL 2025*. 13666–13681.

[6] Federico Castagna, Nadin Kökciyan, Isabel Sassoon, Simon Parsons, and Elizabeth Sklar. 2024. Computational Argumentation-based Chatbots: a Survey. *ArXiv* abs/2401.03454 (2024). https://api.semanticscholar.org/CorpusID:266844589

[7] Günther Charwat, Wolfgang Dvořák, Sarah A Gaggl, Johannes P Wallner, and Stefan Woltran. 2015. Methods for solving reasoning problems in abstract argumentation–a survey. *Artificial intelligence* 220 (2015), 28–63.

[8] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:261530434

[9] Artur S D'Avila Garcez, Dov M Gabbay, and Luis C Lamb. 2005. Value-based argumentation frameworks as neural-symbolic learning systems. *Journal of Logic and Computation* 15, 6 (2005), 1041–1058.

[10] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753* (2024).

[11] Frans H van Eemeren. 2018. *Argumentation theory: A pragma-dialectical perspective*. Springer.

[12] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*. 54–64.

[13] Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2024. M-MAD: Multidimensional Multi-Agent Debate for Advanced Machine Translation Evaluation. *arXiv preprint arXiv:2412.20127* (2024).

[14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv* abs/2312.10997 (2023). https://api.semanticscholar.org/CorpusID:266359151

[15] Camelia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. 2023. Teach me how to argue: A survey on NLP feedback systems in argumentation. In *Proceedings of the 10th Workshop on Argument Mining*. 19–34.

[16] Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. *arXiv preprint arXiv:1906.03717* (2019).

[17] Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. *arXiv preprint arXiv:1805.10254* (2018).

[18] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ArXiv* abs/2311.05232 (2023). https://api.semanticscholar.org/CorpusID:265067168

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[20] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325* (2024).

[21] Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8460–8478. https://doi.org/10.18653/v1/2022.acl-long.579 Introduces the internet-grounded pipeline underlying BlenderBot 2.0.

[22] Pranjal Kumar. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review* 57, 10 (2024), 260.

[23] John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics* 45, 4 (2019), 765–818. https://doi.org/10.1162/coli_a_00364

[24] John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45, 4 (2020), 765–818.

[25] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv* abs/2005.11401 (2020). https://api.semanticscholar.org/CorpusID:218869575

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[27] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.

[28] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. *ArXiv* abs/2407.16833 (2024). https://api.semanticscholar.org/CorpusID:271404721

[29] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:235352646

[30] Jingcong Liang, Rong Ye, Meng Han, Ruofei Lai, Xinyu Zhang, Xuanjing Huang, and Zhongyu Wei. 2024. Debatrix: Multi-dimensional debate judge with iterative chronological analysis based on llm. *arXiv preprint arXiv:2403.08010* (2024).

[31] Bin Lin, Chen Zhang, Tao Peng, Hanyu Zhao, Wencong Xiao, Minmin Sun, Anmin Liu, Zhipeng Zhang, Lanbo Li, Xiafei Qiu, et al. 2024. Infinite-llm: Efficient llm service for long context with distattention and distributed kvcache. *arXiv preprint arXiv:2401.02669* (2024).

[32] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[33] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-Augmented Retrieval for Open-Domain Question Answering. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:221802772

[34] Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819* (2024).

[35] Kazuki Sakai, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. 2020. Hierarchical argumentation structure for persuasive argumentative dialogue generation. *IEICE TRANSACTIONS on Information and Systems* 103, 2 (2020), 424–434.

[36] Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. 109–114.

[37] Kartik Sharma, Peeyush Kumar, and Yunqing Li. 2024. Og-rag: Ontology-grounded retrieval-augmented generation for large language models. *arXiv preprint arXiv:2412.15235* (2024).

[38] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature* 591, 7850 (2021), 379–384. https://doi.org/10.1038/s41586-021-03215-w

[39] Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. 2016. Expert stance graphs for computational argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. 119–123.

[40] Frans H Van Eemeren. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

[41] Frans H Van Eemeren and Rob Grootendorst. 1988. Rationale for a pragma-dialectical perspective. *Argumentation* 2 (1988), 271–291.

[42] Frans H Van Eemeren and Rob Grootendorst. 2016. *Argumentation, communication, and fallacies: A pragma-dialectical perspective*. Routledge.

[43] Frans H Van Eemeren and Peter Houtlosser. 2003. The development of the pragma-dialectical approach to argumentation. *Argumentation* 17 (2003), 387–403.

[44] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36 (2021), e5.

[45] María Alejandra Vitale. 2015. Public memory and the contemporary epideictic genre: death notices devoted to Jorge R. Videla. *Res Rhetorica* (2015). https://bazhum.muzhp.pl/media/texts/res-rhetorica/2015-numer-1/res_rhetorica-r2015-t-n1-s54-64.pdf

[46] Elfia Bezou Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2024. EthiX: A Dataset for Argument Scheme Classification in Ethical Debates. In *27TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE (ECAI)*.

[47] Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3753–3765.

[48] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes.* Cambridge University Press.
[49] Fuyu Wang, Jiangtong Li, Kun Zhu, and Changjun Jiang. 2025. InspireDebate: Multi-Dimensional Subjective-Objective Evaluation-Guided Reasoning and Optimization for Debating. *arXiv preprint arXiv:2506.18102* (2025).
[50] Kevin Wu, Eric Wu, and James Zou. 2024. How faithful are RAG models? Quantifying the tug-of-war between RAG and LLMs' internal prior. *arXiv preprint arXiv:2404.10198* (2024). https://doi.org/10.48550/arXiv.2404.10198
[51] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.
[52] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *ArXiv* abs/2308.08155 (2023). https://api.semanticscholar.org/CorpusID:260925901
[53] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357* (2024).
[54] Wan Zhang and Jing Zhang. 2025. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. *Mathematics* 13, 5 (2025), 856. https://doi.org/10.3390/math13050856
[55] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901* (2024).
[56] Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. Can LLMs beat humans in debating? A dynamic multi-agent framework for competitive debate. *arXiv preprint arXiv:2408.04472* (2024).
[57] Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. 2024. LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:273532096
[58] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A Survey of Large Language Models. *ArXiv* abs/2303.18223 (2023). https://api.semanticscholar.org/CorpusID:257900969
[59] Xiutian Zhao, Ke Wang, and Wei Peng. 2024. ORCHID: A Chinese debate corpus for target-independent stance detection and argumentative dialogue summarization. *arXiv preprint arXiv:2410.13667* (2024).
[60] Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*. 55–62.