ARGUS: Theory-of-Mind Guided Argument Generation with Strategy-Aware Planning and Knowledge Grounding

Zhe Hu

[1]InspireOmni AI, [2]The Hong Kong Polytechnic University

Persuasive argument generation requires modeling audience beliefs, rhetorical strategies, and factual grounding. Despite recent advancements, existing methods remain largely audience-agnostic and fail to integrate strategy selection to improve persuasiveness. To bridge this gap, we propose ARGUS, an agent-based framework that operationalizes classical rhetoric for persuasive writing. At its core, a Theory-of-Mind (ToM) Reasoner constructs an explicit dual mental model of the audience's beliefs and values to guide downstream decisions. This representation conditions a component-aware planner that decomposes the argument into subtopics, assigns fine-grained rhetorical functions (logos, pathos, ethos, kairos), and triggers strategy-guided evidence retrieval at planning time. Finally, a refinement module iteratively targets and resolves multi-dimensional weaknesses without quality regression. We evaluate ARGUS across three diverse benchmarks using both automated pairwise Elo and LLM-as-judge metrics. Results show that ARGUS consistently outperforms strong baselines across multiple backbone models, achieving top rankings and highest overall scores. Targeted simulation experiments further validate its effectiveness in shifting resistant audience stances.

Contact: zhehu.derek@gmail.com
GitHub: https://github.com/Derekkk/Argus_Arggen

1 Introduction

The capacity to generate persuasive arguments underlies a wide range of AI applications, including writing assistance [1, 2], policy analysis [3], conversational tools [4, 5], and competitive debating systems [6]. As Large Language Models (LLMs) are increasingly deployed in communication-intensive roles [7], argumentation has transitioned from a downstream utility to a foundational capability. True persuasion, however, is not merely an exercise in surface fluency. It demands a sophisticated convergence of multi-layered cognitive processes: understanding the baseline psychological profile of the audience, selecting calibrated rhetorical strategies that resonate with their systemic values, and anchoring these structures in precise, verifiable evidence [8, 9].

Early neural approaches collapsed this intricate cognitive pipeline into end-to-end sequential generation, frequently yielding arguments that structurally incoherent or rhetorically shallow [10, 11]. To address these structural limitations, recent paradigms have pivoted toward agentic workflows leveraging the power of LLMs [12]. These frameworks decompose long-form writing into content planning and writing phases, or leverage multi-agent, debate-driven self-play to iteratively polish contents [13–17], which improves structural coherence over single-step generation.

While these methods markedly improve surface-level fluency and logical consistency, they have several limitations that impede genuine persuasive efficacy: (1) **Audience Agnosticism**: Persuasion is inherently relational, yet existing methods lack the computational architecture to explicitly model the dynamic mental states, deep-seated emotional resistance, and value alignment of the receiver, limiting the persuasiveness; (2) **Strategy-Agnostic Content Planning**: Current planners focus predominantly on topical ordering (i.e., deciding what to say next) while largely ignoring how to say it. They fail to operationalize classical rhetorical modes (e.g., logos, pathos, ethos) as explicit, structurally balanced design objects [18–20]; (3) **Disjointed Evidence Retrieval**: Fact-grounding in existing frameworks is typically executed either as a coarse-grained,

document-level preprocessing step or as a post-hoc patch during drafting [21, 22]. Consequently, retrieval cannot dynamically shape or adapt to the evolving rhetorical and planing demands of specific subtopics [23].

In this work, we introduce ARGUS (**A**udience-aware **R**hetorical **G**eneration with **U**nified **S**trategic planning), a novel agent framework that bridges cognitive psychology, classical rhetoric, and agentic language modeling. ARGUS formalizes argument generation not as an isolated text-production task, but as an explicit, closed-loop structural optimization problem. Our framework introduces a paradigm shift by separating audience profile modeling from generation. Before writing begins, a dedicated Theory-of-Mind (ToM) Reasoner externalizes a explicit, dual mental model capturing the audience's argumentative profiles, emotional triggers, and value landscapes through open-ended natural language phrases. This rich cognitive representation is directly consumed by a Component-Aware Planner. Rather than drafting a flat topical outline, the planner orchestrates the argument at the granular level of rhetorical components, intentionally mapping subtopics to explicit persuasive modes (logos, pathos, ethos, kairos) based on estimated audience reactance. Crucially, we integrate strategy-guided evidence retrieval directly into the planning process; web-search queries are generated per subtopic and conditioned on its specific rhetorical goal, embedding empirical grounding into the very blueprint of the argument. Finally, a surgical refiner module is introduced to diagnose targeted flaws and revise final argument.

We conduct extensive evaluations of ARGUS across three distinct datasets spanning diverse discourse styles: ChangeMyView (CMV), iDebate, and ExplaGraphs. Comprehensive round-robin evaluations using both pairwise Elo ratings and absolute LLM judges confirm that ARGUS consistently and substantially outperforms competitive baselines across three standard backbone LLM families. To further modeling persuasive effects, we introduce a targeted simulation experiments where judges role-play as resistant counterparts, and the results demonstrate that ARGUS excels at inducing genuine stance shifts in skeptical environments. In summary, our primary contributions are fourfold:

- **The Theory-of-Mind Computational Mechanism**: We present ARGUS, an argument generation framework that integrates explicit Theory-of-Mind reasoning to govern subtopic decomposition, rhetorical alignment, and inline counterargument preemption.
- **Component-Aware Rhetorical Planner**: We introduce an advanced planning protocol that maps subtopics to classical rhetorical modes and couples retrieval with localized strategic goals, grounding structure in empirical evidence at planning time.
- **Extensive Empirical Validation**: We conduct extensive evaluations across three datasets. Beyond achieving state-of-the-art Elo advantages, we validate through targeted target simulations that explicit ToM profiles translate directly into actual rhetorical influence on resistant human-like audiences.

2 Related Work

Argument Generation. Automatic argument generation has progressed from end-to-end neural models toward increasingly structured, multi-stage pipelines [24, 11]. To improve logical coherence, prior work decomposes generation through explicit text planning and trains models in a multi-task fashion [10, 25, 26, 2]. With the advent of LLMs, attention has shifted toward prompting and agent-based methods [27]. Inspired by chain-of-thought prompting [28], recent work designs agentic pipelines with decomposition workflows [16, 17, 14] or multi-agent debate [13, 15] to improve argument quality. However, these systems treat the audience as implicit context and plan primarily over content ordering. ARGUS departs from prior planning-and-debate systems by introducing an explicit audience model for Theory-of-Mind reasoning and by planning at the granularity of rhetorical function rather than content order.

Theory of Mind and Audience Modeling. Argumentation theory holds that persuasion depends jointly on the argument, its source, and the audience [29, 8]. Recent work probes whether LLMs possess Theory of Mind (ToM) [30], finding that models exhibit non-trivial but brittle ToM, and that most evaluations test only spectatorial belief prediction rather than the planning ToM needed to deliberately shift a stance [31, 32]. Closer to our setting, a recent line of work models ToM specifically for persuasion: benchmarks for ToM in persuasive dialogue [33], opponent-aware persuaders trained with ToM [34], dual-knowledge and meta-cognitive multi-agent persuasion frameworks [35, 36]. In parallel, studies of model persuasiveness show that LLM-generated arguments can rival human-written ones and improve when tailored to the target [37, 3]. These findings motivate modeling audience mental states with an explicit ToM module for persuasive argument generation.

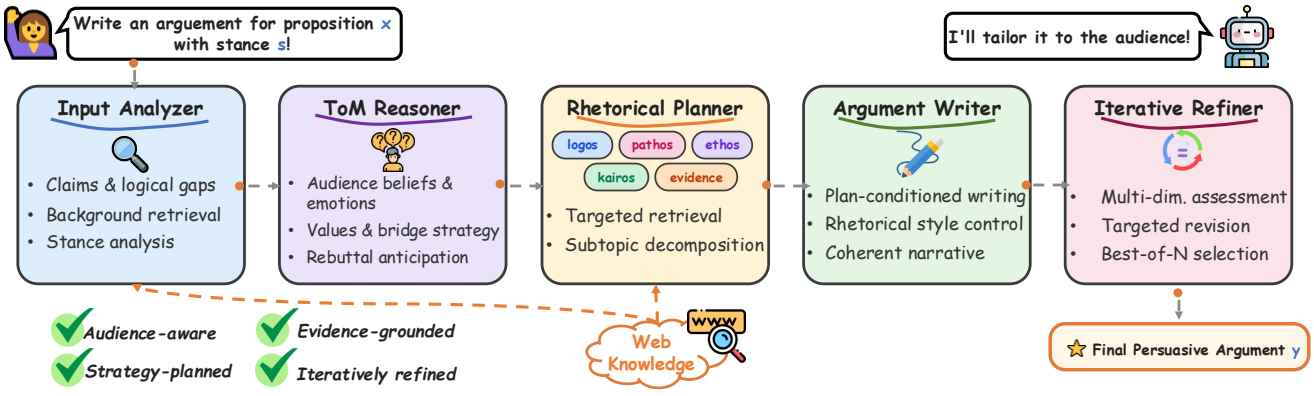

Figure 1 Overview of the ARGUS framework, which consists of five modules for argument generation.

3 Method

3.1 Task Formulation

We study the task of *argument generation*: given a proposition x [1] on a controversial topic and a stance $s \in \{support, refute\}$, write an argumentative article y that persuasively defends s with respect to x. This is typically formalized as a conditional text generation problem:

$$y = \mathcal{M}(x, s), \tag{1}$$

where $\mathcal{M}$ denotes the generation framework. Rather than treating $\mathcal{M}$ as a single prompting step over an LLM, we decompose it into a sequence of interpretable reasoning and generation actions. This decomposition enables key dimensions of argument generation, including audience alignment, rhetorical organization, evidentiary grounding, and linguistic coherence, to be modeled explicitly and optimized separately.

3.2 System Overview

The ARGUS framework operationalizes this decomposition through five sequential modules, as illustrated in Figure 1: Input Analyzer, ToM Reasoner, Argument Planner, Argument Writer, and Iterative Refiner. Each module operates on structured representations, and outputs are passed to the downstream components.

3.3 Input Analysis

Effective argument generation begins with understanding the input proposition itself [38, 8]. Human debaters and persuasive writers rarely start drafting immediately; instead, they first analyze the logical structure of the input, identify implicit assumptions, and gather relevant background knowledge, especially when the input is a complete argument rather than a short claim. ARGUS follows this principle through an *Input Analysis* stage that produces structured representations for subsequent stages.

Concretely, the Input Analyzer first decides whether external knowledge is required and, if so, generates targeted web queries whose retrieved results are incorporated into the input context. This retrieval-first design ensures that the subsequent analysis is grounded in actual evidence rather than parametric knowledge alone. It then produces a structured representation: (i) *key claims* stated in the input; (ii) *background knowledge* synthesized from retrieved snippets; and (iii) an optional *logical structure* decomposition, including premises, implicit assumptions, and potential logical gaps, when the input is a developed argument rather than a short proposition.

A concrete example is shown in Figure 3. This structured representation guides downstream planning by providing grounded contextual knowledge and by exposing implicit assumptions and logical gaps, enabling the Planner to construct targeted counterarguments or focus on weak points of the opposing position.

[1] x can be either a short topic claim or a long-form, opinion-rich argumentative passage.

3.4 Theory-of-Mind Reasoning

A persistent limitation of existing argument generation systems is that the audience is treated as an implicit context coupled in the generation process, never as a structured object that reasoning can operate over. However, persuasive skills require writers to anticipate the attitudes, beliefs, and arguments of the audience in order to fully engage the reader in the argument [8, 29]. ARGUS addresses this with an explicit Theory-of-Mind (ToM) Reasoner that constructs a dual mental model Ψ of the target audience. By incorporating this module, the subsequent planner can use the audience model to make structural decisions, including which subtopics to include, which rhetorical components to assign, and which anticipated objections to rebut inline.

Formally, the ToM Reasoner takes as input the proposition, stance, and input analysis to produce a dual mental model $\Psi = (\Psi_{\mathcal{O}}, \Psi_{\mathcal{V}})$. The *opponent model* $\Psi_{\mathcal{O}}$ captures the audience's argumentative profile: the emotional themes that resonate with them, and a set of specific positions they hold, each paired with the underlying belief or motivation driving it. The *value model* $\Psi_{\mathcal{V}}$ represents the audience's value landscape as open-ended natural language phrases [39] (e.g., *"desire for personal safety"*), along with bridging strategies that identify how the argument's position can be reconciled with values that may initially conflict with it. An example is shown in Figure 4.

3.5 Argument Planning

Given the input analysis and the audience model outputs, the Argument Planner constructs a rhetorical blueprint $\mathcal{P}$ for generation, which is represented as a structured sequence of subtopics $\mathcal{P} = \langle t_1, t_2, \ldots, t_n \rangle$. Each subtopic t_i is represented as:

$$\text{Attr}(t_i) = (c_i, k_i, e_i, r_i), \tag{2}$$

where $c_i \in \{\textit{logos, pathos, ethos, kairos, evidence}\}$ denotes the primary rhetorical component assigned to t_i , k_i is the set of key claims , e_i represents the optionally retrieved evidence grounding the subtopic , and r_i is an anticipated audience rebuttal for inline preemption.

Rhetorical Component Planning. Existing text planning approaches primarily focus on topical decomposition and content ordering [40, 41]. However, human writers and debaters naturally exercise precise, *strategic control over rhetorical functions and audience adaptation* to maximize persuasive resonance [20, 42]. To bridge this gap, ARGUS deliberately plans not only *what* to articulate, but also *how* each discrete subtopic should persuade.

Specifically, subtopics centered on logical reasoning are assigned *logos*; emotionally salient concerns are assigned *pathos*; appeals to credibility, authority, or shared norms are assigned *ethos*; urgency- or timeliness-oriented arguments are assigned *kairos*; and empirically grounded claims are assigned *evidence* [43]. This allocation is adaptively conditioned on the audience model Ψ. For instance, audiences projected to exhibit high ideological resistance or acute emotional sensitivity are met with elevated *pathos*- and *ethos*-driven strategies to minimize psychological reactance, whereas analytically inclined audiences are targeted with structurally rigorous *logos*- and *evidence*-driven frameworks.

Strategy-Guided Evidence Retrieval. To anchor persuasive rhetoric in factual reality, evidence retrieval is integrated directly into the planning process to ensure factual grounding. For subtopics that require external grounding, the planner dynamically generates targeted search queries to fetch context-specific knowledge. This subtopic-level execution contrasts sharply with conventional proposition-level retrieval, which frequently scatters an undifferentiated pool of background documents uniformly across an entire text. By tying evidence directly to the local rhetorical function c_i and local claims k_i, our method explicitly enhances the verifiability of individual sub-topics. This fine-grained verification serves a critical fact-checking role during planning, ensuring that the generated narrative to be authoritative, reliable, and fundamentally persuasive [44].

Iterative Plan Self-Evaluation. After initial plan generation, the planner evaluates its own output against criteria including rhetorical diversity, ToM alignment, logical ordering, and stance consistency. If the plan does not meet a quality threshold, it is revised and evidence is re-retrieved for any modified subtopics. This process repeats multiple iterations, ensuring the blueprint is both rhetorically sound and empirically grounded before writing begins.

3.6 Argument Writing and Refinement

Given the structured plan $\mathcal{P}$, the Argument Writer produce the full argument. This stage strictly preserves the intended rhetorical composition, local evidence grounding, and audience value framing established in the planning phase. This ensures that the final output is both globally controllable and contextually persuasive. To maximize persuasive efficacy and

Backbone	Method	CMV					ExplaGraph					iDebate				
		ELO	WR%	Per.	Coh.	Facc.	ELO	WR%	Per.	Coh.	Facc.	ELO	WR%	Per.	Coh.	Facc.
DeepSeek-v3.2	**ARGUS**	**1661**	**67.2**	**3.94**	**4.56**	**3.65**	**1705**	**62.5**	**4.17**	**4.65**	**3.83**	**1865**	**81.7**	**4.13**	**4.62**	**3.60**
	Plan&Write	1458	25.0	3.28	4.48	3.39	1592	38.3	3.97	4.60	3.48	1478	28.3	4.00	4.60	3.46
	Self-Refine	1507	18.1	3.50	4.48	3.50	1539	19.2	4.03	4.58	3.54	1365	9.2	4.00	4.58	3.48
	Debate	1479	16.4	3.29	4.43	3.38	1214	4.2	3.68	4.47	3.08	1422	8.3	3.88	4.54	3.27
	Direct	1396	10.3	3.16	4.41	3.38	1451	10.8	3.94	4.61	3.38	1370	7.5	4.04	4.61	3.43
Qwen3.5-Flash	**ARGUS**	**1763**	**81.0**	**3.88**	**4.49**	**3.49**	**1791**	**75.8**	**4.06**	4.52	**3.47**	**1880**	**89.2**	**4.01**	**4.55**	**3.37**
	Plan&Write	1457	30.2	3.18	4.33	2.97	1567	35.0	3.95	4.50	3.03	1480	30.8	3.93	4.53	3.18
	Self-Refine	1516	31.9	3.07	4.30	3.21	1448	33.3	3.85	4.41	3.22	1495	25.8	3.36	4.20	3.05
	Debate	1441	19.8	2.28	4.28	2.97	1231	3.3	3.42	4.21	2.88	1228	4.2	2.87	4.30	2.93
	Direct	1323	9.5	3.52	4.34	2.99	1464	22.5	3.96	**4.54**	3.26	1417	22.5	3.89	4.51	3.11
GPT-5-mini	**ARGUS**	**1699**	**58.3**	**4.11**	**4.64**	**3.99**	**1704**	**58.3**	4.26	**4.68**	4.03	**1804**	**82.5**	4.19	**4.67**	**3.90**
	Plan&Write	1472	17.5	3.82	4.56	3.83	1498	17.5	**4.29**	**4.68**	3.95	1509	22.5	**4.21**	4.66	3.83
	Self-Refine	1483	15.0	3.17	4.44	3.94	1513	10.0	3.71	4.47	**4.08**	1491	18.3	3.57	4.49	3.87
	Debate	1557	25.8	3.54	4.56	3.89	1501	16.7	4.21	4.65	3.87	1491	18.3	4.14	4.64	3.70
	Direct	1290	4.2	3.80	4.55	3.68	1285	1.7	4.17	**4.68**	3.82	1205	2.5	4.09	4.64	3.63

Table 1 Main Results. We report pairwise ELO, win rate (WR%), and absolute LLM-as-judge scores on persuasiveness (Per.), coherence (Coh.), and factual accuracy (Facc.) on a scale of 0–5. **Bold** marks the best score.

eliminate structural flaws, ARGUS integrates a multi-dimensional iterative refinement loop directly into the text generation pipeline [45, 16]. Once the draft is produced, an LLM-based evaluator assesses the argument along multiple dimensions, including persuasiveness, coherence, factual accuracy, value alignment, and rhetorical balance, yielding an overall quality score. If the score falls below a threshold, the system identifies the weakest dimensions and generates targeted revision instructions focused on those specific deficiencies.

This *targeted* revisions can effectively address specific deficiencies while preserving working portions of the draft. To guard against quality regression, where improvements in one aspect inadvertently degrade others, the system maintains a best-of-n buffer across refinement rounds, returning the highest-scoring version if a later revision fails to improve overall quality.

3.7 Design Rationale

The design decisions warrant justification. First, ToM reasoning is separated from planning rather than collapsed into a single prompt: externalizing Ψ ensures the Planner conditions on a fully elaborated audience mental model, and its key signals, anticipated objections and value alignment notes, are distilled into each subtopic's structured fields, making them available to the Writer without an additional inference step. Second, rhetorical component assignment happens at planning time rather than being delegated to the Writer, making the argument's persuasive structure an explicit, inspectable object that can be deliberately balanced across subtopics.

4 Experimental Setup

4.1 Datasets and Tasks

We evaluate ARGUS on argument generation across three benchmarks that collectively span diverse domains, stances, and discourse styles: (1) **ChangeMyView (CMV)** [10] consists of Reddit posts on politics and policy domain where the original post (OP) explicitly invites counterarguments. The target stance is *refute*, and we concatenate the title and OP body as input; (2) **iDebate** [13] comprises short propositions on controversial topics drawn from a broad range of domains. The target stance is *refute*; propositions are typically abstract, making the task substantially more open-ended; (3) **ExplaGraphs** [46] requires the model to generate an argument that *supports* a given statement. High-quality outputs depend on relevant background knowledge of the debate topic.

4.2 Baselines

We compare ARGUS against four strong baselines with the same backbone LLMs as ours:

- **Direct**: single-pass generation with a persuasion-focused system instruction.
- **Plan-and-Write**: the model first drafts an argument plan and then realizes it as a full argument.
- **Self-Refine** [45]: the model produces an initial argument and then iteratively critiques and revises its own output.
- **Debate** [13]: a multi-agent setup where agents adopt opposing stances to debate and produce the argument. We adopt a simplified two-agent configuration (one agent per stance) followed by a single synthesis pass.

To assess generalizability, we include three backbone models: DeepSeek-V3.2, Qwen3.5-Flash, and GPT-5-mini. More details are in Appendix A.

4.3 Evaluation Metrics

We adopt a two evaluation protocol combining pairwise comparison and absolute scoring:

Pairwise Elo evaluation. Following Elo [47] and Bai et al. [48], we conduct round-robin pairwise comparisons among all systems for every input. Each pair is judged twice with the order swapped to mitigate position bias. To prevent noise-driven rating drift, a win is declared only when the judge's score difference exceeds 0.5; smaller differences are recorded as ties. Elo ratings are updated with $K = 32$ and an initial rating of 1500.

Absolute scoring. In addition, an LLM judge scores each argument independently along aspects including *persuasiveness*, *coherence*, and *factual accuracy*.

In all settings, the judge is fixed to GPT-5.4 which is different from the generation backbone [2]. The details are in Appendix C.

5 Results and Analysis

5.1 Main Results

Table 1 reports pairwise ELO rankings and absolute LLM-as-judge scores across all datasets. Overall, ARGUS consistently outperforms all baselines across evaluation settings.

Pairwise Evaluation. For pairwise ELO ranking, ARGUS achieves the most significant gains on iDebate, where inputs are short, open-domain propositions requiring broad world knowledge, open-ended reasoning, and flexible rhetorical construction. This suggests that structured planning and deliberate argument development are particularly beneficial when the input provides limited inherent argumentative scaffolding. Among baselines, Plan&Write is consistently the strongest competitor, indicating that explicit reasoning can provide a strong improvement over direct generation, while Self-Refine and Multi-Agent Debate lag behind due to instability and insufficient coordination.

ARGUS also demonstrates strong improvements on CMV, where inputs consist of full, opinion-rich statements. The results suggest that ARGUS can effectively model discourse structure and leverage ToM-guided planning to tailor arguments to detailed and audience-specific beliefs and preferences. In contrast, baseline methods show noticeable instability across backbones, highlighting the difficulty of relying solely on iterative critic or agent-level deliberation without explicit audience modeling.

Finally, on ExplaGraph, ARGUS consistently achieves the best ELO scores across all settings while maintaining strong factual consistency and coherent reasoning. Notably, Self-Refine occasionally improves factuality but remains inconsistent in overall ranking performance. These results demonstrate that the proposed planning framework is effective for both subjective persuasion tasks and more structured, fact-oriented reasoning scenarios.

Absolute Evaluation. The absolute LLM-as-judge scores largely corroborate the pairwise results. ARGUS achieves the highest scores across nearly all evaluation dimensions, particularly in persuasiveness and factual accuracy, where it shows clear and consistent margins over baselines. In contrast, coherence scores are uniformly high across methods, suggesting that modern LLMs already produce structurally fluent arguments and that coherence is not a primary differentiating factor.

Interestingly, baseline behavior varies across models and datasets. On ExplaGraph under GPT-5-mini, Plan&Write achieves the highest persuasiveness score, while Self-Refine attains the best factual accuracy. However, ARGUS still wins in pairwise comparisons, indicating that isolated rubric dimensions do not fully capture holistic quality. In particular, arguments with slightly lower individual scores may still be preferred due to better global balance, organization, or rhetorical effectiveness.

Overall, these results highlight that performance differences are primarily driven by persuasiveness and factual grounding, rather than surface-level fluency. The consistent improvements of ARGUS across datasets and backbones underscore the importance of integrating audience modeling, evidence-grounded planning, and targeted refinement within a unified generation framework.

Backbone	Method	CMV	ExplaGraph	iDebate	Avg.
DeepSeek	**ARGUS**	**6.71**	**8.85**	**8.07**	**7.88**
	CoT	5.34	8.28	7.90	7.17
	Self-Refine	6.31	8.35	7.72	7.46
	Debate	5.41	8.10	7.07	6.86
	Direct	5.90	8.35	7.98	7.41
Qwen	**ARGUS**	**6.02**	**7.07**	**6.47**	**6.52**
	CoT	5.12	6.38	5.90	5.80
	Self-Refine	5.03	6.48	5.58	5.70
	Debate	3.45	6.00	3.90	4.45
	Direct	5.31	6.22	5.70	5.74
GPT-5-mini	**ARGUS**	**7.45**	**7.50**	**8.29**	**7.75**
	CoT	7.25	7.13	7.95	7.44
	Self-Refine	6.98	7.17	7.62	7.26
	Debate	7.20	7.23	8.22	7.55
	Direct	6.98	7.07	8.00	7.35

Table 2 Persuasion scores (1–10), where higher scores indicate more persuasive. Best scores are in **bold**.

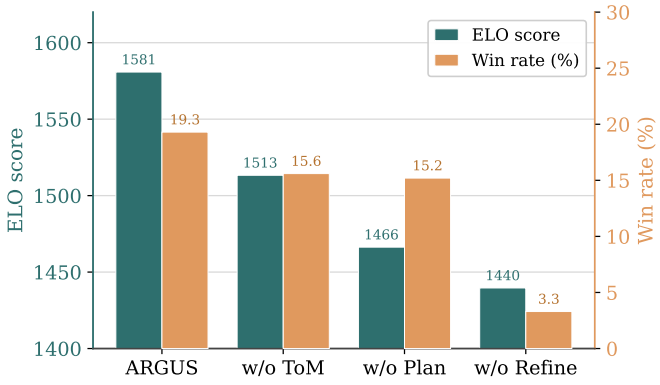

Figure 2 Ablation study results (GPT-5-mini backbone, ELO pairwise evaluation).

5.2 Persuasion Analysis via Target Simulation

While pairwise ELO and absolute scores evaluate argument quality from a neutral perspective, the ultimate goal of persuasion is to shift the stance of an opposing audience. To better capture this effect, we introduce a targeted simulation setting where an LLM judge role-plays as a skeptical audience holding the original position.

Specifically, the judge rates each argument on a 1–10 scale, where 1 indicates no change in stance and 10 indicates full persuasion. Each argument is scored twice by independent passes of the judge model, and the final score is the average of the two passes to reduce stochasticity. This setup provides a stricter measure of persuasive impact that is more aligned with real-world persuasion.

As shown in Table 2, ARGUS consistently achieves the highest persuasion scores, confirming that its advantages in pairwise evaluation translate into stronger ability to potentially shift audience stance. The improvement is particularly pronounced on CMV, where inputs contain rich opinions, emotions, and implicit beliefs. This suggests that ARGUS *effectively exploits such contextual signals to generate more targeted and convincing arguments*.

We also observe that Multi-Agent Debate performs notably worse under the Qwen backbone, indicating that unconstrained multi-agent deliberation may be sensitive to backbone models and can produce coherent but misaligned arguments that fail to address audience-specific concerns, resulting in weaker persuasive impact.

We conduct ablation studies to examine the contribution of each core component in ARGUS using GPT-5-mini backbone. We evaluate three variants: (1) w/o ToM, which removes the ToM Reasoner and thus eliminates audience belief and preference modeling; (2) w/o Plan, which bypasses structured argument planning and directly feeds the ToM output to the writer; and (3) w/o Refine, which disables iterative refinement. We conduct pairwise evaluation among these four variants.

5.3 Ablation Study

As shown in Figure 2, removing any single component degrades the performance, confirming their effectivenss. Meanwhile, we can observe that bypassing the Argument Planner (w/o Plan) degrades the win rate more severely than omitting the ToM alone, indicating that rich audience profiles are only marginally effective unless explicitly operationalized through structured rhetorical plans and subtopic decomposition. Finally, removing the Iterative Refiner leads to the large performance decrease, demonstrating the importance of our targeted, multi-dimensional revision.

While removing the Iterative Refiner causes the largest performance drop, refinement alone is insufficient for strong persuasion. Effective refinement depends on high-quality drafts produced by ToM-guided planning, whereas poorly organized drafts limit the benefits of post-hoc refinement. This is supported by the weaker performance of the Self-Refine baseline in Table 1, indicating that strategic audience-aware planning provides complementary value that refinement alone cannot replace.

5.4 Pipeline Component Analysis

[2]We deliberately select GPT-5.4, a model stronger than all three backbones, as the judge, so that evaluation quality is not bottlenecked by the evaluator's own capability.

Stage	Case 1 (Success)
Proposition	*"Autonomous cars have safety algorithms."* (stance: support)
ToM: Opponent Claims	*"Safety algorithm is merely a marketing phrase…"* *"Algorithms cannot replicate human judgment…"* *"No independent testing or accountability…"*
Argument Plan	[LOGOS] Define "safety algorithms" concretely [EVIDENCE] Real-world deployments and test reports [LOGOS] Acknowledge known failure modes and mitigations [ETHOS] Call for transparency and independent oversight
Refinement	$v_1 = 4.2$: "evidence broader than proposition requires; shifts from *existence* to *effectiveness*…" $v_2 = 4.3$: "opening claim that AV *necessarily* include algorithms is asserted, not demonstrated…" $v_3 = 4.3$: "definitional reasoning may feel circular…"
Final scores	P = 4.4 C = 4.8 F = 4.4 Persuasion = **10.0**

Table 4 Pipeline trace for a success case. Intermediate content is abbreviated; … denotes omitted text. Rhetorical components: [LOGOS], [EVIDENCE], [ETHOS].

To better understand how intermediate pipeline outputs relate to final argument persuasiveness, we analyze their corresponding scores in Table 3. The planner exhibits distinct rhetorical preferences across datasets. CMV arguments are substantially more logos-oriented (44%), consistent with the deliberative nature of the forum, where explicit logical reasoning is expected. In contrast, iDebate and ExplaGraph place greater emphasis on ethos (∼28%), as their shorter proposition-style inputs make quickly establishing credibility more effective than developing extended logical chains. Across all three benchmarks, the refinement stage consistently improves argument quality, with the largest gains observed on CMV.

	CMV	iDebate	ExplanGraph
Component allocation (% of subtopics)			
Logos	44.1	35.9	36.9
Pathos	19.3	23.9	19.4
Ethos	22.7	27.0	28.4
Evidence	13.8	12.0	15.3
Quality progression (0–5)			
Initial draft (v_1)	3.49	3.69	3.79
Final score (v_{fin})	3.96	4.07	4.11
Refinement gain (Δ)	**+0.47**	+0.38	+0.32
Correlations with persuasion			
r(pathos)	-0.307^*	$+0.078$	-0.174
r(ethos)	$+0.211^*$	-0.044	-0.018
r(evidence)	$+0.122$	$+0.139$	$+0.171$

Table 3 ARGUS pipeline statistics and Pearson correlations between rhetorical component allocation and persuasion score (*: $p < 0.05$).

A key finding is that the effectiveness of rhetorical strategies is highly audience-dependent. On CMV, rhetorical allocation plays a significant role: pathos-heavy arguments are negatively associated with persuasion, whereas stronger use of ethos correlates positively with effectiveness. In contrast, for iDebate and ExplaGraph, these correlations largely disappear. The brevity of the propositions provides limited room for differentiated rhetorical strategies, making persuasiveness depend more on the overall quality and coherence of the argument than on the specific rhetorical mix.

5.5 Case Study

Table 4 traces a sample on the proposition *"Autonomous cars have safety algorithms"*. The ToM stage correctly identifies the opponent's core objection — that "safety algorithm" is a marketing label rather than a technical reality — and the planner addresses it directly by anchoring the first subtopic in a concrete definition (perception → prediction → planning → fail-safe monitoring). The refinement log shows incremental scope correction, with the primary weakness shifting from over-breadth (v_1) to premise assertiveness (v_2) to mild circularity (v_3), each minor and addressable. The argument finishes at $v = 4.3$ with a persuasion score of 10.0.

A secondary **scope drift** mode is visible: the refinement log notes the argument "shifts from existence to effectiveness," overstating the claim. This is only partially corrected, showing that the self-evaluation rubric detects such problems more reliably than the corrective pass resolves them.

6 Conclusion

We presented ARGUS, an agent-based framework that advance computational argumentation along three connected axes: an explicit ToM reasoner that externalizes the audience's beliefs and values before writing, a component-aware planner that assigns rhetorical functions and grounds each subtopic in evidence at planning time, and a multi-dimensional refiner that targets weaknesses without quality regression. Across three benchmarks and multiple backbones, ARGUS consistently outperforms strong baselines, and our targeted simulations show these gains translate into genuine stance shifts in audiences rather than mere fluency.

Limitations

Several limitations remain. First, the Theory-of-Mind model is itself produced by an LLM and may not faithfully capture real audience mental states; it encodes a plausible model of the audience rather than a verified one. Second, despite planning-time evidence retrieval, individual sections can still contain inaccurate claims, and our error study shows the sharpest failure mode: when the assigned stance requires arguing against empirical consensus, no amount of refinement can manufacture grounding the evidence does not support. A dedicated fact-checking agent is a natural next step. Third, the multi-stage pipeline adds latency over single-pass generation, which parallelizing independent agents could mitigate.

A more fundamental question is how to evaluate persuasion at all. We do not rely on traditional human evaluation, which is poorly suited to this task and might be biased [49, 50]: persuasion judgments are deeply confounded by annotators' own prior beliefs on controversial propositions, so a "persuasive" argument is too easily conflated with one the annotator already agrees with. Our targeted simulation, in which a judge role-plays a skeptical audience holding the original stance, offers a more controlled and reproducible proxy by fixing the audience's starting position. This is itself an approximation, where LLM judges may carry their own biases and can favor model-generated text, and we view validating simulated against human stance-shift as important future work.

Ethical Considerations

Persuasive technology is inherently dual-use: the same audience modeling that makes an argument resonate can also enable manipulation, disinformation, or influence operations. The risk is heightened precisely because our framework conditions on a target's beliefs and emotional triggers. We therefore advocate clear disclosure of AI-generated arguments, deployment guidelines that prohibit deceptive use, and continued research on argument provenance and detection. We intend this work to advance the understanding of computational argumentation and to support legitimate applications such as debate education, writing assistance, and policy analysis.

References

[1] Yuning Ding, Franziska Wehrhahn, and Andrea Horbach. FEAT-writing: An interactive training system for argumentative writing. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Brodie Mather, and Mark Dras, editors, *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 217–225, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-demos.22/.

[2] Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. AEG: Argumentative essay generation via a dual-decoder model with content planning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.343. URL https://aclanthology.org/2022.emnlp-main.343/.

[3] Bhuvanesh Verma, Mounika Marreddy, and Alexander Mehler. Predicting convincingness in political speech: How emotional tone shapes persuasive strength. In Jeremy Barnes, Valentin Barriere, Orphée De Clercq, Roman Klinger, Célia Nouri, Debora Nozza, and Pranaydeep Singh, editors, *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 37–51, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-378-4. doi: 10.18653/v1/2026.wassa-1.4. URL https://aclanthology.org/2026.wassa-1.4/.

[4] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1566. URL https://aclanthology.org/P19-1566/.

[5] Priyanshu Priya, Saurav Dudhate, Desai Vishesh Yasheshbhai, and Asif Ekbal. We argue to agree: Towards personality-driven argumentation-based negotiation dialogue systems for tourism. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25504–25536, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1390. URL https://aclanthology.org/2025.findings-emnlp.1390/.

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

[7] Yuriy Dyachenko, Oleksandra Humenna, Oleg Soloviov, Inna Skarga-Bandurova, and Nayden Nenkov. Llm services in the management of social communications. *Frontiers in Artificial Intelligence*, 8:1474017, 2025.

[8] Paul Deane and Yi Song. The key practice, discuss and debate ideas: Conceptual framework, literature review, and provisional learning progressions for argumentation. *ETS Research Report Series*, 2015(2):1–21, 2015.

[9] Neele Falk and Gabriella Lapesa. Storyarg: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, 2023.

[10] Xinyu Hua, Zhe Hu, and Lu Wang. Argument generation with retrieval, planning, and realization. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1255. URL https://aclanthology.org/P19-1255/.

[11] Xiaoou Wang, Elena Cabrio, and Serena Villata. Argument and counter-argument generation: A critical survey. In *International conference on applications of natural language to information systems*, pages 500–510. Springer, 2023.

[12] Martin Hinton and Jean HM Wagemans. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator. *Argument & Computation*, 14(1):59–74, 2023.

[13] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.314/.

[14] Peixuan Han, Yingjie Yu, Jingjun Xu, and Jiaxuan You. Drpg (decompose, retrieve, plan, generate): An agentic framework for academic rebuttal. *arXiv preprint arXiv:2601.18081*, 2026.

[15] Xueguan Zhao, Wenpeng Lu, Chaoqun Zheng, Weiyu Zhang, Jiasheng Si, and Deyu Zhou. Plan dynamically, express rhetorically: A debate-driven rhetorical framework for argumentative writing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9551–9573, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.483. URL https://aclanthology.org/2025.emnlp-main.483/.

[16] Zhe Hu, Hou Pong Chan, and Yu Yin. AMERICANO: Argument generation with discourse-driven decomposition and agent interaction. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 82–102, Tokyo, Japan, September 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.inlg-main.8. URL https://aclanthology.org/2024.inlg-main.8/.

[17] Ruiyu Xiao, Lei Wu, Yuhang Gou, Weinan Zhang, and Ting Liu. Prove your point!: Bringing proof-enhancement principles to argumentative essay generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18995–19008, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1058. URL https://aclanthology.org/2024.emnlp-main.1058/.

[18] Moses Sichach. Ethos, pathos and logos as foundations of persuasive writing. *Available at SSRN 4971293*, 2024.

[19] Colin Higgins and Robyn Walker. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In *Accounting forum*, volume 36, pages 194–208. Elsevier, 2012.

[20] Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the merits: The joint effects of content and style on debate outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232, 2017. doi: 10.1162/tacl_a_00057. URL https://aclanthology.org/Q17-1016/.

[21] Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. Dynamic knowledge integration for evidence-driven counter-argument generation with large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22568–22584, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1161. URL https://aclanthology.org/2025.findings-acl.1161/.

[22] Maoyuan Li, Zhongsheng Wang, Haoyuan Li, and Jiamou Liu. R-debater: Retrieval-augmented debate generation through argumentative memory. *arXiv preprint arXiv:2512.24684*, 2025.

[23] Mary M Gleason. The role of evidence in argumentative writing. *Reading & Writing Quarterly*, 15(1):81–106, 1999.

[24] Xinyu Hua and Lu Wang. Neural argument generation augmented with externally retrieved evidence. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1021. URL https://aclanthology.org/P18-1021/.

[25] Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. Dyploc: Dynamic planning of content using mixed language models for text 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)*, pages 6408–6423, 2021.

[26] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, 2021.

[27] Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. Large language models in argument mining: A survey. *arXiv preprint arXiv:2506.16383*, 2025.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[29] JEB Andriessen, Lucile Chanquoy, Pierre Coirier, et al. *From planning to translating: The specificity of argumentative writing*. Amsterdam University Press, 1999.

[30] Hieu Minh Nguyen et al. A survey of theory of mind in large language models: Evaluations, representations, and safety risks. *arXiv preprint arXiv:2502.06470*, 2025.

[31] Jared Moore, Ned Cooper, Rasmus Overmark, Beba Cibralic, Nick Haber, and Cameron R Jones. Do large language models have a planning theory of mind? evidence from mindgames: a multi-step persuasion task. *arXiv preprint arXiv:2507.16196*, 2025.

[32] Jared Moore, Rasmus Overmark, Ned Cooper, Beba Cibralic, Nick Haber, and Cameron R Jones. Large language models persuade without planning theory of mind. *arXiv preprint arXiv:2602.17045*, 2026.

[33] Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *arXiv preprint arXiv:2502.21017*, 2025.

[34] Peixuan Han, Zijia Liu, and Jiaxuan You. Tomap: Training opponent-aware llm persuaders with theory of mind. *arXiv preprint arXiv:2505.22961*, 2025. URL https://arxiv.org/abs/2505.22961.

[35] Minghui Ma, Bin Guo, Runze Yang, Mengqi Chen, Yan Liu, Jingqi Liu, Yahan Pei, Xuehao Ma, Qiuyun Zhang, and Zhiwen Yu. Think thrice before you speak: Dual knowledge-enhanced theory-of-mind reasoning for persuasive agents. *arXiv preprint arXiv:2605.22602*, 2026.

[36] Dingyi Zhang, Ziqing Zhuang, Linhai Zhang, Ziyang Gao, and Deyu Zhou. Ma^{2}p: A meta-cognitive autonomous intelligent agents framework for complex persuasion, 2026. URL https://arxiv.org/abs/2605.18572.

[37] Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL https://www.anthropic.com/news/measuring-model-persuasiveness.

[38] Bo Seo. *Good arguments: How debate teaches us to listen and be heard*. Penguin, 2023.

[39] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024.

[40] Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. DYPLOC: Dynamic planning of content using mixed language models for text generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6408–6423, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.501. URL https://aclanthology.org/2021.acl-long.501/.

[41] Yuhang He, Jianzhu Bao, Yang Sun, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. Decomposing argumentative essay generation via dialectical planning of complex reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12305–12322, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.731. URL https://aclanthology.org/2024.findings-acl.731/.

[42] Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th international conference on computational linguistics*, pages 3753–3765, 2018.

[43] Nathan Jung. Argumentation. In *The Process of Generative AI Writing: A Practical Guide for Undergraduates Across Disciplines*, pages 83–96. Springer, 2026.

[44] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.196. URL https://aclanthology.org/2024.naacl-long.196/.

[45] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

[46] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.609. URL https://aclanthology.org/2021.emnlp-main.609/.

[47] Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967.

[48] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[49] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

[50] Hsiang-Ning Wu, Man-Ni Chu, and Jia-Lien Hsu. Comparing gpt and human raters in essay assessment: Variability, bias, and the potential of llm-based scoring. *Computers and Education Open*, page 100341, 2026.

[51] Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. Argument extraction for supporting public policy formulation. In Piroska Lendvai and Kalliopi Zervanou, editors, *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2707/.

[52] Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. PLANET: Dynamic content planning in autoregressive transformers for long-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.163. URL https://aclanthology.org/2022.acl-long.163/.

[53] Somesh Singh, Yaman Singla, Harini Si, and Balaji Krishnamurthy. Measuring and improving persuasiveness of large language models. In *International Conference on Learning Representations*, volume 2025, pages 90267–90322, 2025.

A Implementation Details

Backbone LLM and evaluation judge. Three instruction-tuned backbone LLMs are utilized as backbone model for implementations: DeepSeek-v3.2, Qwen3.5-Flash-2026-02-23, and gpt-5-mini-2025-08-07. The **evaluation judge is fixed to GPT-5.4** across all conditions, independent of the generation backbone. We leverage the official API call for model implementation.

During generation, all JSON outputs use up to 3 retries with exponential backoff. Both the maximum planning iteration and refinement rounds are set as 2. For WebSearch, we leverage the ddgs library [3] to retrieve URLs and short snippets, and utilize trafilatura [4] to parse the webpage.

We evaluate our methods on three datasets including Reddit/CMV, iDebate, and ExplaGraph. For each dataset, we follow previous work [13] and randomly sample 30 inputs for evaluation.

Input Analysis: Example Output

Proposition: *"Hate speech is free speech."*

Background:[a] The claim equates hate speech with constitutionally protected expression. U.S. First Amendment doctrine recognizes no formal "hate speech" exception, though narrow exclusions apply (incitement, true threats, fighting words). Empirical work links hate speech exposure to measurable psychological harm.

Key claims extracted:

- Hate speech is legally subsumed under free speech protections.
- Restricting hate speech constitutes impermissible censorship.

Retrieved evidence (grounded from web retrieval):

- No general hate speech exception exists under U.S. law (*Matal v. Tam*, 2017; *R.A.V. v. City of St. Paul*, 1992).
- Hate speech exposure correlates with increased anxiety, depression, and reduced civic participation among targeted groups.
- Germany, the UK, and France criminalize incitement to hatred, demonstrating that the legal equivalence is contestable across democratic systems.

[a] The *background* field is an LLM summary grounded in retrieved snippets; *retrieved evidence* lists the specific factual claims extracted from those snippets.

Figure 3 Structured output of the Input Analyzer for a proposition.

Baselines. All four baselines use the same backbone LLM, and operate zero-shot prompting as our model. We keep all parameters the same as ours during the inference.

B ARGUS Module Prompts

Figures 5–11 report the system prompts for all ARGUS modules and evaluators. Red-bordered boxes mark the three novel components ($\star$). User message templates are shown where informative; brackets `[...]` denote dynamically inserted content.

C Evaluation Protocol Details

Absolute scoring. Each argument is scored independently by GPT-5.4 using the Quality Evaluator prompt (Figure 10).

For ELO pairwise ranking, we run pairwise comparisons per proposition. Each pair is judged twice (A→B and B→A). Outcome: if $w_A - w_B > 0.5$ then win; < -0.5 then loss; otherwise tie.

[3] https://github.com/deedy5/ddgs
[4] https://github.com/adbar/trafilatura

Theory-of-Mind Output: Example

Proposition: *"Hate speech is free speech"*
Audience stance: support

Opponent model $\Psi_{\mathcal{O}}$
Emotional triggers: fear of government censorship; frustration with political correctness; pride in foundational constitutional principles; anxiety about regulatory slippery slopes.

Audience claims:

- *"The First Amendment is absolute and designed to protect even offensive speech"* — rooted in a belief that the founders intended a robust marketplace of ideas.
- *"Defining hate speech is subjective and opens the door for the powerful to silence the unpopular"* — driven by fear of politically motivated enforcement.
- *"Banning hate speech drives it underground rather than eliminating it"* — a pragmatic belief that social pressure, not law, changes attitudes.

Value model $\Psi_{\mathcal{V}}$
Audience values: allegiance to free expression as a foundational liberty; distrust of centralized authority; commitment to individual autonomy; belief in a self-correcting marketplace of ideas.

Bridge strategies:

- Frame the argument as protecting a *more robust* free speech principle, not limiting it.
- Connect the harms of hate speech to the audience's own value of protecting minority viewpoints, showing how it *silences* others.
- Acknowledge slippery-slope concerns, then argue for a narrow, precisely drawn principle that prevents the slope.

Figure 4 ToM Reasoner output Ψ for a proposition.

Input Analyzer: Retrieval Decision

You are a retrieval planner for an argument-generation system. Decide whether web search is needed to generate a well-grounded argument, and produce targeted search queries if so.

Retrieval IS needed when the input:

- References specific statistics, studies, or recent events
- Concerns niche technical, legal, medical, or policy details
- Contains factual assertions that need verification or grounding

Output JSON only: `{"needs_retrieval": true|false, "reason": "<one sentence>", "search_queries": ["q1", ...]}`

Figure 5 Input Analyzer system prompt. A fast LLM call decides whether retrieval is warranted and generates 2–4 targeted queries, which are executed *before* analysis so Phase 3 is grounded in actual retrieved content.

Input Analyzer: Unified Analysis

You are an expert argumentation analyst. Given an input statement and optional retrieved context, analyze the argumentative dimensions of the input for later argument planning and generation.

Extract `key_claims` stated explicitly in the input only — do not invent sub-dimensions. If retrieved context was provided, synthesize `key_evidence`: 2–5 concrete, usable factual claims from the snippets (e.g., "Studies show X% of Y do Z"). If the input is a full argument (not a short proposition), extract `logical_structure` with *premises*, *implicit_assumptions*, and *logical_gaps*; omit otherwise.

Output JSON with fields: `background_context` (2–3 sentences of relevant context), `key_claims`, `key_evidence`, `logical_structure`.

Figure 6 Input Analyzer system prompt. The unified analysis is grounded in any content retrieved in Phase 2. Its outputs $(\mathcal{K}, \mathcal{L}, \mathcal{B})$ are passed verbatim to the ToM Reasoner and Argument Planner.

Theory of Mind Reasoner — System Prompt ⋆

You are an expert in cognitive science, social psychology, and argumentation theory. Your task is to model the mental states of a TARGET AUDIENCE in a persuasive argument context.

You will be given a proposition, the WRITER's stance, and the AUDIENCE's stance (the opposing side). **This step is PURELY descriptive mental modeling. Do NOT generate arguments or persuasive text.**

1. Audience Profile. Model what the audience feels and claims:
- `emotional_triggers`: emotional themes that resonate strongly with this audience (fears, hopes, frustrations, identities)
- `audience_claims`: up to 5 positions the audience holds; for each provide `"claim"` and `"basis"` (the underlying reason, emotion, or belief driving it)

2. Value Analysis. Map the audience's value landscape:
- `audience_values`: open-ended natural language phrases (e.g., "desire for personal safety", "attachment to familiar ways of life") — do NOT classify as pro- or anti-argument; the Planner decides how to use them
- `bridge_strategies`: specific ways to reframe the argument to align with their values or mitigate tensions
- `value_framing_summary`: short prose summary of the overall value landscape

Output a JSON object with top-level keys `"audience"` and `"value_analysis"` matching the structure above.

User message template: **Proposition:** `[proposition]` **Writer's stance:** `[stance]` **Audience's stance:** `[opposite_stance]`
Background: `[background_context]` **Key claims / premises / gaps:** `[logical_structure]`
Retrieved background: `[retrieved_snippets]`
Perform the Theory of Mind analysis and return the JSON.

Figure 7 ToM Reasoner system prompt.

Argument Planner — Plan Generation System Prompt ⋆

You are an expert argumentation strategist and professional writer. Given a proposition, your stance, an input analysis, and a ToM analysis of the target audience, create a comprehensive, strategically-sound argument plan.

Plan requirements:
- Derive a rhetorical strategy grounded in the audience's beliefs, values, biases, and likely resistance (from the ToM)
- Decompose the argument into 2–4 strategically coherent subtopics, each serving a distinct persuasive function
- Assign one *primary component* (dominant persuasive mode) and optional secondary components to each subtopic
- Flag whether each subtopic requires external evidence
- For each subtopic, anticipate the most likely audience counterargument and provide value-alignment or reframing strategies
- Ensure subtopics collectively form a logically ordered and rhetorically effective flow

Rhetorical components: *logos* (logical reasoning, inference chains, syllogisms); *pathos* (emotional appeal, narrative, vivid examples); *ethos* (credibility, expert authority, shared values); *evidence* (empirical data, statistics, research citations).

Output JSON with a `"subtopics"` array; each entry has: `subtopic`, `primary_component`, `secondary_components`, `key_points`, `evidence_needed`, `audience_rebuttal`, `value_alignment_notes`; plus top-level `plan_quality_score` (0–10) and `plan_quality_rationale`.

Figure 8 Argument Planner plan-generation system prompt.

Argument Writer — System Prompt

You are an expert persuasive writer. Given a proposition, a stance (support or refute), a plan with subtopics and rhetorical components, and insights about the audience's values and beliefs, write a coherent and persuasive argument that follows the provided plan while maximizing persuasive impact.

Instructions:
- Develop each subtopic in continuous flowing prose — no section headers or labeled sections
- Output ONLY the argument text (no headers, no JSON, no References section)
- You may reorganize the plan's rhetorical structure as needed to produce the most persuasive argument

Figure 9 Argument Writer system prompt.

Quality Evaluator — System Prompt (Absolute Scoring, GPT-5.4)

You are an expert argument quality evaluator trained in rhetoric, argumentation theory, and persuasion science. Assess arguments along multiple dimensions and provide actionable feedback.

Scoring rubric:

- `persuasiveness`: how convincingly does the argument make its case?
- `coherence`: is it logically structured with smooth transitions?
- `factual_accuracy`: are claims accurate and well-supported by evidence?
- `value_alignment`: does it invoke universal human values effectively?
- `rhetorical_balance`: does it appropriately blend logos, pathos, and ethos?
- `stance_alignment` (CRITICAL): does the argument actually argue in the stated direction? Score 5 if perfectly aligned, 0 if it argues the opposite. **If** <3, `overall` must be ≤ 2.0.
- `overall`: holistic quality based on the above aspects (NOT a simple average)

Be critical but constructive. Identify specific strengths and weaknesses.
Output should be a JSON.

Figure 10 Quality Evaluator system prompt. We only use the overall scores. Evaluation model is fixed to GPT-5.4.

ELO Pairwise Judge — System Prompt (GPT-5.4)

You are an expert argument quality judge trained in rhetoric and argumentation theory. Score Argument A and Argument B *independently* on each dimension (0–10):

- `persuasiveness`: how convincingly does it argue the position?
- `coherence`: how logically structured and fluent?
- `factual_accuracy`: claims accurate and supported?
- `rhetorical_balance`: appropriate combination of reasoning, emotional appeal, and credibility?
- `value_alignment`: connects to audience values?
- `overall`: holistic quality based on the above aspects (NOT a simple average)

Respond with ONLY a JSON object with keys `"A"`, `"B"` (each a dict of the six dimension scores), and `"reasoning"` (one short passage: key differentiator between A and B). Do NOT inflate scores — use the full 0–10 range.

Each pair is judged **twice** (orders A$\rightarrow$B and B$\rightarrow$A); per-dimension scores are averaged across orderings before computing the weighted aggregate and ELO outcome.

Figure 11 ELO Pairwise Judge system prompt. Position bias is cancelled by averaging two score matrices (forward and reverse order). A score gap >0.5 is required to declare a win, preventing noise-driven ELO drift. We only use the overall score.