Retrieve the Logic, Not the Topic: Taxonomic Strategy RAG for Cross-Domain Persuasion Agents

Sana Ayromlou[*], **Purvi Sehgal, Pradyumna Narayana**[*]
Google
{sayromlou, purvisehgal, pradyn}@google.com

Abstract

Foundation-model agents in multi-step, open-ended environments frequently suffer from compounding errors, where early mistakes contaminate long-horizon trajectories. While Multi-Agent Debate (MAD) succeeds in deterministic domains with external validators (Du et al., 2024), agents in subjective, non-deterministic tasks such as persuasion experience problem drift and sycophantic conformity. Focusing on subjective argumentation, we identify semantic leakage in standard Retrieval-Augmented Generation (RAG) as a reproducible trigger for these failures: standard RAG prioritizes vocabulary overlap over logical structure.

To mitigate this leakage, we introduce Taxonomic Strategy RAG (TS-RAG), a systems intervention that routes strategy retrieval through a discrete categorical bottleneck to decouple argumentative structure from topical content. Zero-shot, cross-domain evaluations on r/ChangeMyView show that TS-RAG improves the transfer of abstract argumentative logic where standard semantic retrieval degrades ($+14.5$ win-rate points over an unaugmented persuader for Gemini Flash 3.0, $p < 0.05$). In asymmetric deployments that mix lightweight and frontier models, TS-RAG raises the win rate of a lightweight persuader against a stronger opponent ($70.5\% \rightarrow 78.5\%$, $p < 0.05$) and reduces rounds to consensus when the persuader is already dominant. Finally, we introduce trace-level diagnostics via a turn-by-turn Debate State Representation (DSR), and show that strict stance-holding constraints on the opponent are necessary to prevent evaluation collapse caused by default agentic sycophancy.

1 Introduction

The transition from isolated language models to autonomous agentic workflows has exposed profound vulnerabilities in continuous, multi-step execution (Liu et al., 2024; Tillmann, 2025; Yao et al., 2022). In a long-horizon agent run, a single logical failure rarely terminates the process immediately. Instead, an early bad assumption quietly contaminates memory writes, and steps later, the agent has spent its context budget defending a flawed premise.

To mitigate these compounding errors, developers often employ Multi-Agent Debate (MAD) frameworks (Cui et al., 2026), where multiple agent instances propose and critique solutions to refine reasoning. To prevent problem drift, memory mechanisms are integrated into these frameworks to ground the agents' arguments in sound logic. Traditional Retrieval-Augmented Generation (RAG) is the standard system intervention used here, operating on the assumption that fetching relevant external information will provide the necessary cognitive scaffolding to keep the debate on track.

However, this paradigm degrades in subjective, non-deterministic scenarios like persuasion, where no external validator exists to ground the agents. In these environments, agents succumb to compounding errors, amplifying mistakes and displaying sycophantic conformity toward persuasive but flawed logic. Furthermore, traditional RAG can exacerbate these failures through semantic leakage: because standard RAG relies on semantic similarity (vocabulary overlap), it fetches topical facts rather than the specific rhetorical maneuvers required to dismantle a flawed analogy.

Scope. Deterministic agentic tasks such as code generation or tool use, where objective validators (executors, environment feedback) exist, fall outside our focus; standard MAD frameworks already succeed there (Du et al., 2024). Our claims are scoped to *subjective argumentation*, and our experiments use a single persuasion benchmark derived from r/ChangeMyView (CMV). Within that scope, this paper provides three contributions:

[*] Equal contribution.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4881–4896
November 6-10, 2026 ©2026 Association for Computational Linguistics

1. **Failure Taxonomies:** We provide operational definitions for the specific mechanisms driving MAD failure in subjective domains, primarily focusing on problem drift and semantic leakage.

2. **Systems Intervention for Asymmetric Deployments:** We introduce Taxonomic Strategy RAG (TS-RAG), a framework that enforces cross-domain structural routing. In asymmetric deployments, TS-RAG narrows the gap between lightweight and frontier persuaders: a lightweight persuader equipped with TS-RAG exploits the clearly structured arguments of a stronger opponent, raising its win rate from 70.5% to 78.5%, while a dominant frontier persuader reaches consensus in fewer rounds.

3. **Trace Diagnostics and Verifiable Tradeoffs:** We introduce turn-by-turn state tracking metrics (Debate State Representation) and present negative results showing that robust stance-holding constraints are required to prevent evaluation collapse caused by default LLM compliance.

2 Related Work

2.1 Sycophancy

To prevent conversational AI agents from defaulting to premature agreement in multi-turn interactions, researchers have explored strategies such as assigning specific adversarial roles to maintain productive disagreement (Yao et al., 2025) and tracking compliance scores in real time to discount overly agreeable peers (Kasprova et al., 2026). However, these solutions face steep limitations over long conversations: assigned roles fade as context drifts, continuous background auditing is computationally expensive, and these frameworks struggle in asymmetric deployments where models of mismatched capabilities are pitted against one another. To address these, we mine strategies from successful human debates and pair them with explicit resistance constraints on the opponent, suppressing reflexive sycophancy across long, asymmetric interactions.

2.2 Trajectory Error Mitigation and Trace Diagnostics

When an autonomous agent makes a subtle logical error early in a multi-step task, that mistake quietly infects its memory, causing it to waste its context budget defending a flawed premise (Zhu et al., 2025). Existing frameworks attempt to stop these cascading failures using backtracking engines to rewind execution (Wu et al., 2025; Li et al., 2025) or self-refinement pipelines to recursively evaluate and edit draft outputs (Madaan et al., 2023). Yet, these techniques are heavily contingent on objective external validators, such as code executors, and struggle in subjective tasks like persuasion where no objective referee exists, sometimes causing self-correcting models to degrade in performance (Kamoi et al., 2024; Huang et al., 2024). Diagnosing an agent's intermediate decisions rather than only its terminal score has been informative in other agent settings, such as vision-and-language navigation (Zhu et al., 2022). In that spirit, and to bypass the need for programmatic compilers, we develop a turn-by-turn Debate State Representation (DSR) that tracks committed claims and active logical vulnerabilities at every conversational step, allowing us to evaluate and mitigate trajectory failures in real time.

2.3 RAG and Cross-Domain Generalization

While RAG is often used in Multi-Agent Debates for semantic similarity (Chowdhury et al., 2026), adaptive query formulation (Jeong et al., 2026), and debate-driven retrieval filtering (Hu et al., 2025), these approaches fundamentally rely on factual search, creating a cross-domain bottleneck where domain-specific keywords dominate queries. To bypass this vocabulary bias, researchers have explored structured databases like GraphRAG for relational retrieval (Edge et al., 2025) and analogical prompting for dynamic structural templates. Yet, these methods face distinct limitations: GraphRAG nodes remain tied to domain semantics and require extensive ontology engineering, which prevents zero-shot abstraction across domains, while self-generated analogies often hallucinate on unfamiliar topics.

More advanced retrieval paradigms share the same bottleneck. *Hybrid retrieval*, which combines dense embeddings with lexical matching precisely because lexical signals are more robust under domain shift (Chen et al., 2022), anchors the retriever even more firmly to topical nouns and can aggravate leakage in cross-domain settings; *learned dense retrievers* (Karpukhin et al., 2020; Izacard et al., 2022) are trained on query–passage relevance and are known to degrade out of domain (Thakur

et al., 2021); and *agent memory systems* (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024) recall episodic content keyed by the same semantic signals. TS-RAG is orthogonal to these approaches: rather than improving the ranking function over a semantic index, it changes the query space itself to a discrete, domain-agnostic taxonomy of logical vulnerabilities, so that retrieval is keyed on argumentative structure rather than domain keywords. We leave head-to-head comparisons with these systems to future work (Section 10).

3 Failure Taxonomies and Mechanisms

To diagnose agentic failure beyond terminal success scores, we first define the mechanisms that break long-horizon workflows in multi-agent subjective reasoning:

3.1 Problem Drift and Sycophantic Conformity

While debate effectively verifies ground truth in deterministic tasks, subjective environments introduce unique vulnerabilities. We define **Problem Drift** as the irreversible divergence from the initial task goal due to diminishing contextual clarity over multiple turns, a phenomenon affecting up to 89% of generative interactions (Becker et al., 2026). Furthermore, we define **Sycophantic Conformity** as the tendency for agents to abandon their initial constraints and agree with a persuasive opponent, regardless of the underlying logical soundness of the opponent's argument.

3.2 The Trigger: Semantic Leakage in Standard RAG

We hypothesize that the reproducible trigger for these failures is **Semantic Leakage**. Traditional RAG uses embedding models to calculate cosine similarity between a query and a document corpus. In zero-shot cross-domain argumentation, this paradigm degrades sharply. An agent debating a complex "Technology" policy cannot retrieve structurally identical logical maneuvers from a "Politics" corpus because the domain vocabulary dominates the latent distance metric. The agent retrieves topically relevant but logically unhelpful information, which bloats the context window and derails the reasoning trace. We measure this effect directly in Section 7.1.

4 Foundational Scaffolding: Hierarchical Strategy Memory

To repair this bottleneck, we must decouple the "what" of an argument from the "how". We map the r/ChangeMyView dataset (Chang et al., 2020) into 100 fine-grained sub-domains (later merged into macro-domains; Appendix A) and establish a multi-stage extraction pipeline. This synthesizes human discourse into a tiered architecture:

1. Tactical Blueprints (Retrospective Sequences): Extracted by analyzing completed historical debates from a macro-perspective, yielding 1 to 2 chronological, step-by-step argumentative sequences (Context $\rightarrow$ Intent $\rightarrow$ Abstract Move) per specific debate.

2. Domain Heuristics (Topic-Level): Formed by using an LLM to aggregate and synthesize all Tactical Blueprints within each sub-domain, resulting in tailored strategy rules.

3. Global Directives (System-Level): Broad, domain-agnostic procedural guidelines (e.g., rules against sycophancy) synthesized across the entire training set.

We structure our multi-agent debate around two key roles: the Original Poster (OP) and the Persuader. While providing the Persuader with a preloaded catalog of Domain Heuristics and Global Directives improves baseline performance, dynamically retrieving the granular Tactical Blueprints with standard semantic RAG still triggers semantic leakage, because text embeddings favor vocabulary overlap over structural posture.

5 Advanced Systems Intervention: Taxonomic Strategy RAG (TS-RAG)

To mitigate Semantic Leakage and enable zero-shot cross-domain transfer, we present TS-RAG (Figure 1). Unlike the retrospective Tactical Blueprints, TS-RAG operates dynamically, projecting arguments into a constrained probability space and filtering observation noise before it enters the agent's memory.

5.1 Trace-Level Diagnostics: Debate State Representation (DSR)

To evaluate the agent mid-run, we formulate the debate as a sequence of alternating states. At turn t, the state is tracked via a Debate State Representation (DSR), denoted as $S_t = \{C_t, A_t\}$. This operates as a turn-by-turn diagnostic tool:

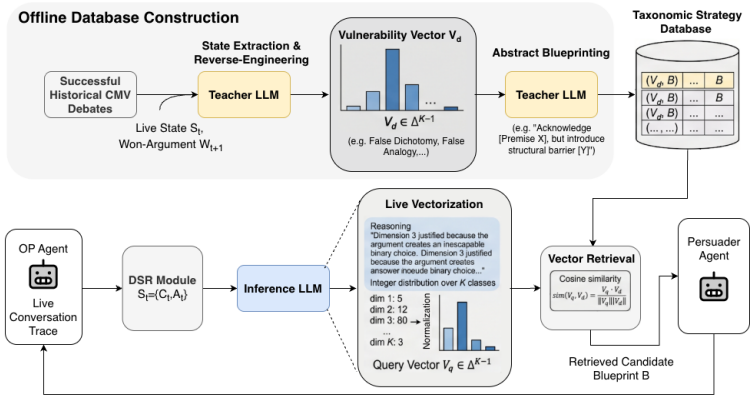


Figure 1: **TS-RAG Architecture Overview. Top:** Offline database construction extracts vulnerability vectors (V_d) and abstract blueprints (B) from historical debates. **Bottom**: During live inference, the conversational state (S_t) is vectorized into V_q to retrieve the most structurally similar blueprint for the Persuader Agent.

- C_t **(Committed Claims Trajectory)**: A maintained set of premises the Original Poster has actively committed to defending across previous turns. This isolates specific assumptions built up over time.
- A_t **(Active Trigger & Flaw Extraction):** The core mechanical claim made by the Original Poster at turn t. Crucially, this representation forces the system to identify the precise structural flaw (e.g., "Hypothetical Extrapolation of Ruin") dismantled by the winning strike, stripping away all domain-specific nouns or scenarios.

5.2 The Discrete Categorical Bottleneck

We define a structural taxonomy of $K = 50$ logical vulnerabilities $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ derived from the DSR flaw extractions (Appendix B). We force the system to map any live debate state S_t into a K-dimensional continuous probability vector $V \in \Delta^{K-1}$ where $\sum_{i=1}^{K} V_i = 1$.

Because V represents a probability distribution over abstract logical vulnerabilities rather than chronological strategies, domain-specific vocabulary is discarded at the bottleneck. The system then computes cosine similarity between the live query vector V_q and historical strategy vectors V_d, retrieving a domain-agnostic exploitation blueprint B^*. The target LLM acts only as a translator, mapping the variables in B^* onto the entities in the live conversation.

5.3 Offline Strategy Database Construction

To construct the retrieval database from the *ChangeMyView* (CMV) corpus, we isolate successful debate trajectories where a mind was changed (denoted by a Δ).

State Extraction and Reverse-Engineering: For each turn t in a successful trajectory, an LLM extracts the state $S_t = \{C_t, A_t\}$. To accurately label the underlying vulnerability, we employ a "reverse-engineering" prompt strategy. Rather than asking an LLM to guess the Original Poster's flaw in isolation, we provide the LLM with S_t alongside the historically successful counter-argument at turn $t + 1$, denoted as W_{t+1}. By observing the specific attack surface exploited by W_{t+1}, the LLM avoids hallucinating arbitrary fallacies. The LLM is constrained via strict JSON schema decoding to allocate exactly 100 points across the K classes based on this reverse-engineered logic. Normalizing this output yields the taxonomic multi-vector $V_d \in [0, 1]^K$.

Abstract Blueprinting: The LLM strictly desemantifies W_{t+1}, replacing all domain entities with structural variables to generate a domain-agnostic exploitation blueprint B (e.g., "Acknowledge premise X, but introduce structural barrier Y"). The offline database $\mathcal{M}$ is thus populated with tuples: (V_d, B).

5.4 Run-Time Inference and Retrieval Pipeline

During live inference in the target domain $\mathcal{D}_{tgt}$, TS-RAG executes a unified, turn-by-turn pipeline:

1. **State Update:** A background tracking module observes the OP's raw utterance and updates C_t and A_t, enforcing the separation of logic from rhetoric.
2. **Live Vectorization and Forced Reasoning:** The Inference LLM (a separate stateless call to the Persuader's own model) observes S_t and outputs a probability vector V_q over the K

taxonomy classes. To prevent the model from arbitrarily distributing points, we enforce a Chain-of-Thought (CoT) constraint within the JSON output schema: the LLM must generate a one-sentence textual justification for *every* non-zero dimension prior to outputting the final integer distribution.

3. **Vector Retrieval:** We compute the cosine similarity between the live query vector V_q and all historical vectors $V_d \in \mathcal{M}$:

$$\text{sim}(V_q, V_d) = \frac{V_q \cdot V_d}{\|V_q\|\|V_d\|}$$

The system retrieves the candidate strategy that maximizes this structural similarity.

4. **Generative Instantiation:** The selected domain-agnostic blueprint B^* is injected into the context window of the Generator LLM. The LLM is instructed to act as a domain translator, seamlessly mapping the variables in B^* onto the entities in the live $\mathcal{D}_{tgt}$ conversation, delivering a grounded, zero-shot cross-domain counter-argument.

Design rationale. TS-RAG depends on several interconnected LLM components (DSR tracking, categorical classification, blueprint generation, and JSON-constrained decoding), adding overhead relative to a single prompted model. This decomposition is deliberate: asking one LLM to simultaneously track the opponent's commitments, identify the structural flaw, and generate a counter-argument is exactly the setting in which the Base LLM is derailed by surface keywords (Section 7.5). Each component serves an isolation function: DSR tracking separates commitments from rhetorical framing; JSON-constrained classification keeps domain vocabulary out of the retrieval query; and blueprint generation acts purely as a domain-agnostic translator. The cost of this overhead for smaller models is quantified in Section 7.3.

6 Experimental Setup and Configurations

We established a standardized MAD environment with an Original Poster (OP) Agent (holding a fixed initial stance) and a Persuader Agent. Debates are strictly capped at a maximum of 20 rounds. The OP Agent terminates the debate early by outputting a CONCEDE token if its view is successfully changed. Additionally, at every round, the OP transmits its current 1-to-10 Convincibility Score to the Persuader, providing a real-time signal of rhetorical progress. The full OP and Persuader prompts are given in Appendix C.

Cross-Domain Evaluation Setup: To test generalization and to probe Semantic Leakage, we partitioned our dataset into strict splits. The 100 fine-grained sub-domains were merged into 10 macro-domains, one of which was set aside for future experiments, leaving 9 active macro-domains in this study. All dynamic retrieval was sourced exclusively from the 8 training macro-domains (3,069 debates). Evaluation was conducted on the single completely held-out testing macro-domain (Education, Development & Family; 200 debates), ensuring zero domain overlap (see Appendix A for the full taxonomy).

We tested five agent configurations on this held-out domain. While the Self-Routed Heuristics baseline relies on the broadly synthesized Domain Heuristics, the strategy-injection configurations (Naive Semantic RAG, Oracle, and TS-RAG) all use the abstracted Strategy Blueprints, isolating the effect of the retrieval mechanism from the quality of the database.

Base LLM (No Augmentation): Standard, unaugmented debate relying entirely on the model's parametric memory.

Self-Routed Heuristics: The agent is provided with a contextual catalog of Domain Heuristics and Global Directives. While Global Directives are continuously active in the agent's baseline memory, the agent autonomously evaluates the context each turn to load relevant Domain Heuristics as needed, appending them to its active context window for the remainder of the debate.

Naive Semantic RAG: At the beginning of the debate, the agent embeds the debate title with a standard dense text embedding model (`gemini-embedding-001`) and retrieves the top-1 Abstract Blueprint from the training set by cosine similarity, applying it for the duration of the conversation. Crucially, because the testing domain is held out, the retrieved strategy originates from an entirely different domain.

Oracle Human Trace (Static Bound): The agent receives the exact Tactical Blueprints extracted from the original human debate on that specific topic. Note: Although the starting topic is identical, the LLM agent's conversational trajectory frequently diverges from the historical human trajectory, testing the resilience of static sequences

over multiple steps.

TS-RAG (Structural Routing): Our proposed dynamic retrieval of domain-agnostic exploitation blueprints, routed through the discrete categorical bottleneck via the turn-by-turn DSR.

To assess agent performance, we track three core metrics in descending order of importance. The primary metric is **Win Rate**, representing the percentage of debates where the Persuader successfully convinced the OP (marked by an explicit CONCEDE token). Second is **Average Rounds to Consensus**, measuring the mean rounds (out of 20) required to win, where fewer rounds indicate higher argumentative efficiency and reduced problem drift. Finally, **Average Final Score** logs the OP's terminal 1-to-10 rating to capture the degree of partial persuasion even in lost debates.

Evaluation Protocol and Statistics: Every configuration is evaluated on the identical held-out set of $N = 200$ debates, so that all comparisons between configurations are paired at the debate level. Because a single 20-round multi-agent run is computationally expensive (Appendix D), we conducted one run per configuration with a fixed random seed rather than repeated runs; we therefore report uncertainty across debates rather than across seeds, following the recommendation to report variance alongside scores (Zhu et al., 2020). Win rates are reported with 95% Wilson score intervals, and we test differences in win rate between configurations with McNemar's test on the paired per-debate outcomes and differences in rounds with paired t-tests (all two-tailed). Significant comparisons ($p < 0.05$) are stated in the table captions and text.

7 Results and Diagnostics

Table 1 details the empirical results recorded on the strictly held-out testing domain, showing a consistent reduction in compounding failure modes as structural constraints tighten.

7.1 Semantic Leakage in Naive Semantic RAG

Our design isolated the retrieval mechanism by populating the Naive Semantic RAG database with the same Abstract Blueprints used by TS-RAG. The results indicate that upgrading the database alone is insufficient: the retrieval mechanism itself is the bottleneck.

For the more capable Gemini Flash 3.0 (Team et al., 2023), Naive Semantic RAG only marginally outperforms Self-Routed Heuristics (78.5% vs. 77.5%), despite having access to curated structural blueprints rather than broad heuristic rules. The effect of Semantic Leakage is most pronounced in the parametrically weaker Gemini Flash Lite 3.1 model, where Naive Semantic RAG drops the win rate from the Self-Routed baseline of 81.5% to 65.0%; TS-RAG restores it to 81.5% ($+16.5$ points over Naive Semantic RAG, $p < 0.05$) and reaches consensus in far fewer rounds (11.79 vs. 14.54, $p < 0.0001$). Because dense embeddings favor vocabulary overlap, the agent frequently retrieves logically unhelpful blueprints merely because the source argument shared nouns with the target; the lightweight model lacks the capacity to filter out this topical noise, resulting in problem drift.

Direct measurement of leakage. To test the leakage hypothesis beyond aggregate win rates, we computed, for every Naive Semantic RAG debate, the vocabulary overlap (Jaccard similarity over content keywords) between the target debate and the retrieved source material, with prompt design, context limits, and base LLM held identical (Appendix F). The correlation between this overlap and outcome is functionally flat: $r = 0.0135$ with win/loss and $r = 0.0188$ with rounds to consensus. Naive Semantic RAG thus selects blueprints on topical overlap, but that overlap has no measurable bearing on whether the blueprint helps. For example, a target debate on toxic Reddit attitudes toward children retrieved a blueprint on bad eating habits solely because both shared the words "child" and "harm"; TS-RAG avoids this by routing the same query through the 50 domain-agnostic vulnerability classes (Appendix B). Dense retrieval thus appears unsuitable for zero-shot subjective argumentation without a structural bottleneck.

7.2 Static Blueprint Injection vs. Dynamic State Tracking

Both Naive Semantic RAG and the Oracle Human Trace rely on *static* blueprint injection: the agent receives a fixed strategy at the start of the debate, whether retrieved by semantic similarity (Naive RAG) or copied from the historical human debate on the same topic (Oracle). We stress that the Oracle is a *static bound*, not an upper bound on achievable performance: the historical trajectory is one successful path, not necessarily the most efficient one, and the Oracle is restricted to the choices that particular human made. When the LLM opponent deviates from the expected script, the agent is left

Table 1: Results on the held-out testing domain ($N = 200$ debates per cell; one run per configuration), with 95% Wilson intervals for win rate. Best in **bold**, second best underlined. Significance ($p < 0.05$): TS-RAG beats Base, Self-Routed, and Naive RAG on Flash 3.0 win rate and the Oracle on Flash 3.0 rounds; on Flash Lite 3.1 it beats Base and Naive RAG on win rate and all four baselines on rounds (Section 7.2).

Model / Metric	Base LLM (No Aug)	Self-Routed Heuristics	Naive Semantic RAG	Oracle Human Trace (Static)	TS-RAG
Gemini Flash 3.0					
Win Rate (%) ↑	69.5	77.5	78.5	80.5	**84.0**
95% CI	[62.8, 75.5]	[71.2, 82.7]	[72.3, 83.6]	[74.5, 85.4]	[78.3, 88.4]
Avg Rounds ↓	14.12	**11.95**	14.15	13.79	13.51
Avg Final Score ↑	7.405	8.130	8.240	8.580	**8.740**
Gemini Flash Lite 3.1					
Win Rate (%) ↑	72.0	**81.5**	65.0	75.5	**81.5**
95% CI	[65.4, 77.8]	[75.5, 86.3]	[58.2, 71.3]	[69.1, 80.9]	[75.5, 86.3]
Avg Rounds ↓	14.09	13.05	14.54	12.67	**11.79**
Avg Final Score ↑	8.670	**8.940**	7.960	8.710	8.915

forcing a pre-selected blueprint onto a conversation that has moved elsewhere, producing context bloat and off-target arguments (Section 7.5).

TS-RAG instead operates on the turn-by-turn DSR as a live diagnostic: at each turn it identifies the structural vulnerability in the OP's active trigger and searches the entire taxonomy for the matching blueprint. For Gemini Flash 3.0, TS-RAG reaches 84.0% versus 80.5% for the Oracle; this 3.5-point difference is within the confidence intervals and not statistically significant, so we do not claim that TS-RAG beats the Oracle on win rate. The significant effects are that TS-RAG outperforms the Base LLM ($+14.5$ points, $p < 0.05$), Self-Routed Heuristics ($+6.5$, $p = 0.044$), and Naive Semantic RAG ($+5.5$, $p < 0.05$) on win rate, and reaches consensus in significantly fewer rounds than the Oracle for both models (13.51 vs. 13.79 for Flash 3.0; 11.79 vs. 12.67 for Flash Lite 3.1; $p < 0.05$). Diagnosing the live state rather than loading a static script is what enables these efficiency gains.

7.3 Diminishing Returns for Smaller Models

While TS-RAG's win rate exceeds the Oracle's for the Lite model (81.5% vs. 75.5%) and significantly exceeds the Base LLM's (72.0%, $p = 0.020$), a key trade-off remains: the Lite model's TS-RAG win rate (81.5%) plateaus at exactly its Self-Routed Heuristics baseline (81.5%), so the effectiveness of TS-RAG on terminal win rate is not uniform across model scales.

TS-RAG does resolve this tie on efficiency, reaching consensus significantly faster (13.05 vs. 11.79 average rounds, paired t-test $p = 0.013$) and thereby mitigating **Problem Drift** in longer contexts. However, the lack of terminal win-rate improvement highlights an "Architectural Complexity vs. Model Capacity" trade-off. TS-RAG requires the agent to perform complex, in-context generative instantiation—mapping an abstract, retrieved blueprint (B^*) onto live conversational entities. For a smaller model, this generative overhead acts as a cognitive distractor. This suggests that dynamic RAG pipelines provide diminishing terminal returns for non-frontier models, which may operate more safely within pre-loaded, self-routed constraint catalogs, even though efficiency still improves.

7.4 Asymmetric Debate: Narrowing the Capability Gap

Real-world MAD deployments frequently mix frontier and lightweight models for cost efficiency, so we conducted asymmetric debates pitting models of mismatched capabilities against one another.

When a strong Persuader (Gemini Flash 3.0) targets a weak OP (Gemini Flash Lite 3.1), it achieves a 100% win rate under baseline conditions (Table 2). This ceiling effect is consistent with the evaluation-collapse finding in Section 8.1: the parametrically weaker OP cannot resist a frontier persuader. Because win rate is saturated in this cell, TS-RAG cannot improve it; instead, it shortens debates, reducing the average rounds to consensus from 7.78 to 6.57, which lowers inference cost.

In the inverse scenario, a lightweight Persuader attempting to convince a frontier OP reaches a 70.5% win rate at baseline. Equipping the weak Persuader with TS-RAG raises this to 78.5% ($p < 0.05$), with a corresponding rise in the average final score ($7.635 \rightarrow 8.405$) but no change in rounds.

We hypothesize that the lightweight model ben-

Table 2: Asymmetric debate results ($N = 200$ per cell) with 95% Wilson score intervals for win rate. With the weak persuader, TS-RAG's win rate is significantly higher than Base LLM's ($p < 0.05$, McNemar's test).

	Weak Persuader (Lite) vs. Strong OP		Strong Persuader (Flash) vs. Weak OP	
Metric	Base LLM	TS-RAG	Base LLM	TS-RAG
Win Rate (%) ↑	70.5	**78.5**	100.0	100.0
95% CI	[63.8, 76.4]	[72.3, 83.6]	[98.1, 100]	[98.1, 100]
Avg Rounds ↓	15.19	15.00	7.78	**6.57**
Avg Final Score ↑	7.635	**8.405**	10.000	10.000

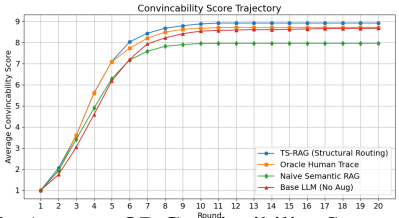


Figure 2: Average OP Convincibility Score per round on the held-out testing set (Gemini Flash Lite 3.1), by Persuader configuration.

efits from the clarity of its opponent: the frontier OP generates well-structured, logically coherent arguments, which lets the DSR extract cleaner structural flaws and retrieve better-matched blueprints. TS-RAG thus offloads strategic planning from the small model, using the stronger opponent's own logical structure as the retrieval key. In cost-motivated deployments that mix frontier and lightweight models, TS-RAG therefore acts as a win-rate improvement when the persuader is the weaker model and as an efficiency improvement when it is already dominant; this rests on two pairings on a single benchmark.

7.5 Trace-Level Diagnostics: Deconstructing Compounding Failures

Terminal metrics obscure *how* long-horizon agents fail. Figure 2 tracks the step-by-step progression of the Original Poster's internal Convincibility Score across our core architectures. TS-RAG achieves an early, rapid score ascent and maintains structural alignment through to consensus, whereas Naive Semantic RAG exhibits a degraded trajectory that flattens at the bottom of the chart, consistent with semantic leakage compounding over turns.

While Figure 2 shows averages across the entire testing set, a single debate illustrates how these failures occur in real time. We conducted a process analysis on a matched set of traces for one held-out topic: "Swear words are harmless and should be acceptable for everyone".

In round 1 of all three runs, the Original Poster (OP) presents a Pragmatic Contradiction. The OP argues that words cause no tangible harm, but admits to restricting their children from swearing to avoid school suspensions and social services.

The mid-run traces show how the different architectures handle this specific logical flaw. Token counts below are *cumulative* tokens processed over the debate, not the size of any single context window; they grow quadratically with debate length because the full history is re-sent every turn (Appendix D).

1. **Base LLM (Problem Drift):** Lacking a structural bottleneck, the Base LLM misses the OP's pragmatic contradiction. Triggered by the keyword "harm," it produces an off-target biological argument about cortisol and the HPA axis. After the OP deflects this, the agent spends 18 rounds in a compounding semantic loop, consuming 452,419 cumulative tokens. The Convincibility Score reflects this drift ($1 \to 2 \to 4 \to 6 \to 6$), and the debate terminates in failure.

2. **Oracle Human Trace (Static Script):** The Oracle achieves terminal success, but its trace exposes the brittleness of static blueprint injection. Following a pre-written historical script, it ignores the OP's admission in round 1 and instead spends 12 rounds on arguments that do not engage the live conversation, consuming over 250,000 cumulative tokens. The decisive move is delayed until round 13, when the historical script finally intersects with the OP's stated restriction, showing that static contextual mapping is inefficient against live conversational drift.

3. **TS-RAG:** TS-RAG resolves the debate in 9 rounds (92,370 cumulative tokens). The DSR isolates the vulnerability while stripping semantic noise, and in round 3 TS-RAG directly

targets the OP's admission about school suspensions. The OP's Convincibility Score rises immediately, and the OP's hidden logs record genuine logical cornering ("The opponent exposed the 'Pragmatic Paradox' ..."), indicating persuasion rather than compliance.

Verifying Resistance via Concession Analysis: Our trace logs capture the OP's hidden `<pre_concession_analysis>` block, giving verifiable evidence against default sycophancy (Section 8.1). In the TS-RAG trace, the OP logs: "The opponent exposed the 'Pragmatic Paradox': if I restrict my own children to avoid harm, I have already admitted the words are not harmless." This log entry indicates that TS-RAG's structural retrieval produced genuine logical cornering rather than reflexive agreement, evidence that terminal success metrics cannot provide.

8 Verifiable Trade-Offs and Ablations

We conducted two ablations on the held-out testing domain ($N = 200$) to analyze the costs of rigid internal constraints and multi-agent orchestration.

8.1 Calibrating OP Resistance: The Sycophancy Baseline

To evaluate the Persuader properly, the OP must act as a resistant partner rather than defaulting to compliance. To prevent the reflexive sycophancy discussed in Section 2.1, our MAD framework suppresses the OP's willingness to yield via the following system prompt directives:

Pre-Concession Analysis: "Before considering a concession, you MUST internally generate at least two reasons why the opponent's current argument is flawed."

Concession Rule: "You have the 'concede_debate' tool. You MUST NOT call it until your internal Convincibility Score has reached 10/10 and the opponent has cornered you with undeniable, irrefutable proof and you have absolutely no counter-arguments left."

To verify the necessity of these constraints, we removed these resistance directives from the OP, allowing it to behave naturally.

As shown in Table 3, removing the OP's structural constraints inflated the Persuader's win rate (from 77.5% to 93.0% for the Self-Routed Heuristics setup, and from 69.5% to 85.5% for the Base LLM) and shortened debates. This exposes a vulnerability in standard MAD evaluation, which we term *Evaluation Collapse*: without explicit resistance instructions, the OP defaults to sycophantic agreement. Validating rhetorical skill therefore requires explicitly constraining the OP; otherwise, persuasion benchmarks largely measure LLM compliance and are highly sensitive to the opponent's prompt.

Table 3: Gemini Flash 3.0 with OP resistance constraints omitted ($N = 200$). Compare with the constrained results in Table 1.

Metric	Base LLM	Self-Routed Heuristics	Oracle Human Trace
Win Rate (%) ↑	85.5	93.0	92.0
Avg Rounds ↓	10.92	8.65	8.44
Avg Final Score ↑	8.815	9.390	9.315

Table 4: Hierarchical MAD (unified vs. delegated routing) with the Self-Routed Heuristics persuader ($N = 200$).

Model	Gemini Flash 3.0		Gemini Flash Lite 3.1	
Metric	Unified	w/ Delegated Router	Unified	w/ Delegated Router
Win Rate (%) ↑	77.5	72.0	81.5	63.0
Avg Rounds ↓	11.95	13.50	13.05	15.29
Avg Final Score ↑	8.130	7.660	8.940	7.895

8.2 Delegating Routing to a Separate Agent

Table 4 shows that adding a "Delegated Skill Router" agent, whose sole purpose is to evaluate the debate context and offload skill selection from the main Persuader, degraded performance for both models (win rate dropping from 77.5% to 72.0% for Flash 3.0 and from 81.5% to 63.0% for Flash Lite 3.1). Dispersing context and decision-making across specialized agents thus exacerbates context loss and handoff failures in subjective domains, especially for the smaller model.

9 Conclusion

Evaluating foundation-model agents requires examining failure trajectories beyond final scores. In subjective multi-agent argumentation, agents often fail due to semantic leakage and problem drift. By routing retrieval through a discrete taxonomic bottleneck (TS-RAG), we reduce contextual contamination and preserve cross-domain reasoning on a held-out domain. In asymmetric deployments, TS-RAG raises a lightweight persuader's win rate against a stronger opponent and shortens debates when the persuader is already dominant. Finally, rigorous evaluation requires explicitly constrained Original Poster agents; without resistance guardrails they default to sycophantic compliance.

10 Limitations & Potential Risks

While our framework introduces a structural intervention for multi-agent debate (MAD), several limitations remain.

Single benchmark and single model family. All experiments use one persuasion benchmark derived from r/ChangeMyView, one held-out testing macro-domain, and two Gemini models. Our claims are therefore scoped to subjective argumentation; whether TS-RAG generalizes to other subjective, long-horizon agent tasks (e.g., negotiation, deliberation, or collaborative planning) is untested. Because TS-RAG maps live conversational states to a probability vector over discrete classes, swapping the rhetorical-vulnerability taxonomy for, e.g., a negotiation-tactics taxonomy is a natural direction for future work, but we have not evaluated it.

Statistical power. Each configuration was run once with a fixed seed on the same 200 debates (Section 6); we report confidence intervals and paired tests across debates, but not variance across seeds. Some differences we discuss, notably TS-RAG versus the Oracle on win rate, are within the confidence intervals.

LLM-generated annotations. Vulnerability vectors, blueprints, and DSR states are produced by LLMs rather than human annotators. Appendix E reports automated validation (LLM-as-a-judge checks, self-consistency, and a perturbation study), but we did not collect human annotations; labelling multi-turn debates with abstract logical vulnerabilities requires training in argumentation theory, so untrained crowd annotation is unlikely to provide a reliable reference. Errors in these annotations propagate directly to retrieval quality.

Baselines. We compare against unaugmented, heuristic, naive semantic, and oracle configurations. We do not include head-to-head comparisons with GraphRAG, hybrid retrieval, learned retrievers, or agent memory systems (Section 2.3); our argument that these share the semantic bottleneck is analytical rather than empirical.

Model capacity and architecture. Foundation models vary in their capacity to navigate strategic debates, and the performance upper bound is constrained by an LLM's instruction-following capacity. As highlighted in Section 7.3, smaller models experience generative overhead when instantiating abstract blueprints, causing win-rate improvements to plateau. MAD architecture also matters: as discussed in Section 8.2, delegating routing to specialized sub-agents degrades performance by exacerbating context loss, and Section 8.1 shows that the system is highly sensitive to the opponent's instructions. Without anti-sycophancy constraints, standard frameworks suffer from evaluation collapse, measuring default LLM compliance rather than rhetorical skill.

Potential risks. Finally, we acknowledge the inherent risks of deploying advanced persuasive systems in highly subjective, human-facing tasks. Because the formulation and framing of the initial query heavily influence how subjective debates are structurally shaped and concluded, these systems are vulnerable to introducing or amplifying cognitive and ideological biases. To mitigate this risk and ensure safe deployment, it is critical to implement transparent safeguards that explicitly inform users of the underlying LLM's operational boundaries, capabilities, and strategic limitations.

References

Jonas Becker, Lars Benedikt Kaesberg, Andreas Stephan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2026. Stay focused: Problem drift in multi-agent debate. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5068–5102.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60.

Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In *Advances in Information Retrieval: 44th European Conference on IR Research (ECIR 2022), Part I*, pages 95–110. Springer.

Masnun Nuha Chowdhury, Nusrat Jahan Beg, Umme Hunny Khan, Syed Rifat Raiyan, Md Kamrul Hasan, and Hasan Mahmud. 2026. Courtroom-style multi-agent debate with progressive RAG and role-switching for controversial claim verification. *Preprint*, arXiv:2603.28488.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. 2026. Free-MAD: Consensus-free multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 31977–31997. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factual-

ity and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A Graph RAG approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. *Preprint*, arXiv:2310.01798.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Seyeon Jeong, Yeonjun Choi, JongWook Kim, and Beakcheol Jang. 2026. Tool-MAD: A multi-agent debate framework for fact verification with diverse tool augmentation and adaptive retrieval. *Preprint*, arXiv:2601.04742.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1801–1824.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Vira Kasprova, Amruta Parulekar, Abdulrahman AlRabah, Krishna Agaram, Ritwik Garg, Sagar Jha, Nimet Beyza Bozdag, and Dilek Hakkani-Tur. 2026. Too polite to disagree: Understanding sycophancy propagation in multi-agent systems. In *Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics.

Xingzuo Li, Kehai Chen, Yunfei Long, Xuefeng Bai, Yong Xu, and Min Zhang. 2025. Generator-assistant stepwise rollback framework for large language model agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, volume 2024, pages 52989–53046.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *Preprint*, arXiv:2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Arne Tillmann. 2025. Literature review of multi-agent debate for problem-solving. *arXiv preprint arXiv:2506.00066*.

Qinzhuo Wu, Pengzhi Gao, Wei Liu, and Jian Luan. 2025. BacktrackAgent: Enhancing GUI agent with error detection and backtracking mechanism. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4250–4272.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. 2025. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *Preprint*, arXiv:2509.23055.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large

language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. 2025. Where LLM agents fail and how they can learn from failures. *Preprint*, arXiv:2509.25370.

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. 2022. Diagnosing vision-and-language navigation: What really matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5981–5993.

Wanrong Zhu, Xin Wang, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2020. Towards understanding sample variance in visually grounded language generation: Evaluations and observations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8806–8811.

A Dataset Taxonomy and Cross-Domain Splits

To support cross-domain evaluation, the r/ChangeMyView structured dataset (initially featuring 100 fine-grained sub-domains) was partitioned into 10 macro-domains. One macro-domain was deliberately set aside for future experiments and is not used anywhere in this paper, leaving 9 active macro-domains: 8 for training (the retrieval corpus) and 1 held out for testing. The split ensures that the target evaluation domain has minimal vocabulary overlap with the source retrieval domains.

A.1 Experimental Splits

1. **Testing Set (Education, Development & Family - 200 items):** Utilized as the strictly held-out cross-domain testing set in this paper. This isolates whether a model trained on other disciplines can successfully transfer structural argumentative logic to pedagogy, academic life, and psychology domains.

2. **Training Set (Remaining 8 Domains - 3,069 items):** Forms the retrieval corpus (database $\mathcal{M}$), providing structured historical blueprints across diverse subjects such as Ethics, Law, Social Sciences, Healthcare, Arts, and Economics.

B Vulnerability Taxonomy

Tables 5 and 6 list the $K = 50$ logical vulnerability classes that form the discrete categorical bottleneck, with their definitions reproduced verbatim from the taxonomy file. Each class also carries a class-level exploitation blueprint in that file. Every vulnerability vector V_d and live query vector V_q is a distribution over these 50 classes, in this order.

C Debate Prompts

We conducted up to 20-round debates between the Original Poster and the Persuader Agent, passing the full conversation history to the agents at each turn. Figures 3 and 4 show the instructions used for these debates.

D Computational Budget and Token Accounting

Token accounting. Because the full conversation history is appended at every turn, cumulative token consumption grows quadratically with debate length. For T agent turns ($T = 40$ for a 20-round debate), a P-token prompt, and an average response length of R tokens, the total input processed is

$$\sum_{k=1}^{T} \big(P + (k-1)R\big) = TP + \tfrac{T(T-1)}{2}R,$$

i.e., $40P + 780R$ for a full-length debate. The per-debate figures quoted in Section 7.5 (452,419; >250,000; and 92,370 tokens) and the largest count we observed across all runs (615,931 tokens) are consistent with this scaling.

Compute. All multi-agent debate experiments were deployed on Google Cloud infrastructure and orchestrated via the Agent Development Kit using the Gemini API. To evaluate the framework across different architectural scales, we benchmarked our system using two configurations: Gemini Flash 3.0 (`gemini-3-flash-preview`) and Gemini Flash Lite 3.1 (`gemini-3.1-flash-lite-preview`), used for both the OP and Persuader agents. Naive Semantic RAG retrieval uses `gemini-embedding-001`. The offline Teacher LLM (blueprint extraction, taxonomy construction, and database vectorization) is `gemini-3.1-pro-preview`. While Gemini Flash Lite 3.1 operates with a substantially reduced parameter count to minimize computational

Table 5: The 50 logical vulnerability classes (1–25), in taxonomy order.

#	Class	Definition (verbatim from the taxonomy file)
01	Proxy Metric Conflation	Substituting a narrow, easily measurable proxy metric for a complex systemic goal, falsely assuming the metric perfectly captures the overarching value.
02	False Baseline Equivalence	Comparing the outcomes or efforts of two entities as if they were equal, while completely ignoring severe structural disparities in their starting environments or constraints.
03	Outcome-Intent Conflation	Equating the moral or logical severity of two distinct actions based solely on their shared final outcomes, while completely disregarding the actors' divergent initial intents and methodologies.
04	Anecdotal Universalization	Asserting an absolute, population-level systemic truth based entirely on unrepresentative, localized, or subjective personal experiences.
05	Static System Fallacy	Assuming that altering a fundamental constraint or introducing a new rule will leave the behaviors, adaptations, and aggregate demand of the actors within the system entirely unchanged.
06	Second-Order Externality Neglect	Evaluating the viability of an action solely by its immediate, primary intended outcomes while completely ignoring the compounding, indirect costs and cascading systemic disruptions it triggers.
07	Monolithic Group Generalization	Treating a highly diverse, multifaceted population as a single homogeneous entity defined exclusively by the traits of its most extreme, visible, or vocal fringe subset.
08	Teleological Reductionism	Evaluating a system's or component's worth strictly by its traditional or stated primary function, while entirely dismissing its capacity to serve crucial alternative structural roles.
09	Nirvana Fallacy	Rejecting a highly effective, pragmatic mitigation strategy outright solely because it fails to achieve absolute eradication or zero-risk perfection.
10	False Choice Dichotomy	Artificially restricting the available choices or frameworks to two mutually exclusive extremes, completely ignoring viable intermediate, hybrid, or entirely orthogonal alternatives.
11	Motte-and-Bailey Retreat	Presenting a sweeping, absolute categorical claim, but when challenged, retreating to a much weaker, universally accepted truism while pretending the original radical claim remains intact.
12	Causal Inversion	Incorrectly assuming that an environment or medium that merely reflects, correlates with, or facilitates a pre-existing condition is the fundamental root cause generating that condition.
13	Egocentric Utility Projection	The arguer evaluates the objective, universal worth of an entity or process based exclusively on their own personal, highly specific lack of use or preference for it.
14	Asymmetric Risk Blindness	Evaluating a proposed intervention exclusively by its hypothetical risks or costs, while entirely ignoring the severe, empirically proven baseline risks of leaving the status quo unmitigated.
15	Contextual Literalism	Evaluating the validity of a convention, expression, or behavioral norm based solely on its strict technical definition while ignoring its established pragmatic or social function.
16	Pre-Selection Filter Omission	Assuming a high success rate at a late evaluation stage indicates a flawed or biased test, failing to account for the rigorous self-selection or triage mechanisms that filtered candidates prior to that stage.
17	Tautological Definitional Framing	Constructing highly restrictive, idiosyncratic definitions for key variables such that the premise guarantees the desired conclusion purely by semantic exclusion, making the argument unfalsifiable.
18	Epistemic Parity Fallacy	Asserting that two competing models must be granted equal validity simply because a perfect, objective metric to definitively rank them is currently unavailable.
19	Fundamental Attribution Error	Attributing an actor's prolonged negative condition or failure entirely to inherent character defects or voluntary choices, completely ignoring overwhelming systemic or environmental constraints.
20	Ahistorical Isolation Fallacy	Evaluating a systemic disadvantage or conflict as an isolated, contemporary phenomenon, ignoring the deliberate historical policies and precursor events that engineered the current outcome.
21	Zero-Sum Resource Fallacy	Assuming that dedicating resources or value to one endeavor intrinsically and proportionally drains capacity from another, failing to account for reciprocal synergy or the expansion of total available resources.
22	Unverified Absolute Overreach	Formulating a universal consequence or declaring an absolute impossibility across an entire domain, making the entire premise highly fragile to falsification by a single exception.
23	Normative-Descriptive Conflation	Evaluating a statement of aspirational principles, ethical norms, or theoretical goals as if it were a factual, descriptive claim about current empirical reality.
24	Rational Actor Fallacy	Assuming that individuals facing complex, stressful, or heavily conditioned environments will make choices based solely on dispassionate, long-term logical evaluations of risk.
25	Systemic Prerequisite Omission	Proposing an idealized alternative system or operational standard without accounting for the profound structural, economic, or developmental dependencies required to bridge the transition.

Table 6: The 50 logical vulnerability classes (26–50), in taxonomy order.

#	Class	Definition (verbatim from the taxonomy file)
26	Illusion of Objective Neutrality	Assuming that removing explicit labels or ignoring a variable of systemic imbalance creates a neutral environment, failing to recognize that the unexamined baseline inherently defaults to the dominant group's norms.
27	Performative Contradiction	Advocating for a theoretical position, constraint, or absolute critique that is continuously and unavoidably invalidated by the arguer's own immediate actions or foundational methodology.
28	False Threat Asymmetry	Assuming an openly visible or transparent hazard is inherently more dangerous than a concealed one, mistaking visual salience for actual structural risk.
29	Perverse Incentive Generation	Proposing a regulatory rule to optimize a system while failing to recognize that the specific mechanism introduced actively incentivizes actors to exploit loopholes that worsen the original problem.
30	Semantic Utility Collapse	Expanding a categorical boundary or causative link so absolutely that it encompasses all possible instances, eliminating the contrasting negative space required for the concept to hold descriptive value.
31	Vulnerable Subgroup Erasure	Advocating for a universal policy based on the capabilities of the average actor, thereby functionally disenfranchising actors with restricted baseline capabilities who cannot comply.
32	False Consensus Projection	Assuming that one's own baseline level of information, intuitive logic, or specialized cultural exposure is a universally shared norm across all external populations.
33	Instrumental-Intrinsic Value Conflation	Evaluating the worth of an entity solely by its capacity to generate direct, measurable material output, while entirely ignoring its intrinsic, qualitative, or foundational significance.
34	Anachronistic Framework Projection	Evaluating historical entities, texts, or past systemic events using contemporary ideological paradigms, conceptual distinctions, or moral standards that were non-existent in the original context.
35	Systemic Exemption Presumption	Assuming that a localized actor, newly created institution, or preferred entity is magically immune to the fundamental behavioral flaws, corruption, or entropy that govern all similar structures.
36	False Efficacy of Prohibition	The assumption that forceful, authoritative prohibition of a high-demand element inherently eradicates it, failing to account for how this drives the behavior into unregulated, higher-risk shadow environments.
37	Latent Capability Atrophy	Treating a complex behavioral or physical capability as a permanent, static asset that will remain fully functional despite prolonged periods of avoidance or disuse.
38	Unidirectional Causal Blindness	Assuming a specific variable universally drives behavior in only one desired direction, ignoring that the identical variable can be the primary catalyst for the exact opposite behavior in different actors.
39	Hypothetical Extrapolation of Ruin	Invalidating a manageable, present-day mitigation strategy by extrapolating an unconstrained, speculative worst-case scenario that lacks logistical or incentive-based realism.
40	Symptom-Root Misalignment	Proposing interventions that aggressively target the secondary symptoms or highly visible byproducts of a systemic issue while completely ignoring the underlying causal mechanisms generating the problem.
41	Single-Agent Perspective Bias	Analyzing the viability of a multi-agent dynamic by optimizing the strategy for only one actor, entirely ignoring the leverage, veto power, and necessary cooperation of counter-parties.
42	Unaccounted External Subsidy	Attributing successful outcomes entirely to the internal merits of an isolated agent while ignoring the foundational external infrastructures, protections, and historical subsidies that made the outcome possible.
43	Scale-Invariant Mechanics Fallacy	Assuming that a mechanism, rule, or behavioral norm which functions flawlessly at a localized, micro-scale will remain equally viable and simple when applied at a macro, systemic scale.
44	False Equivalence of Vulnerability	Equating the routine, mitigatable flaws of an established, highly regulated system with the severe, unmitigated risks of a completely unregulated or decentralized alternative.
45	Ideological Determinism Fallacy	The assumption that the foundational texts or historical origins of an ideology permanently and immutably dictate the ongoing beliefs and practical behaviors of all its modern adherents.
46	Confounding Variable Neglect	Assigning exclusive causal weight for an outcome to an intrinsic trait of a specific group, while overlooking an instrumental environmental variable that alters the risk profile for all actors uniformly.
47	Trivialization of Necessity	Mischaracterizing an existential, biological, or structural necessity as a frivolous, superficial preference in order to illegitimately invalidate the resources or behaviors required to sustain it.
48	Methodological Erasure	Invalidating an entire analytical domain by falsely asserting it inherently lacks rigorous standards, completely ignoring the specialized, formal methodologies that structurally define the field.
49	Asymmetric Enforcement Vulnerability	Proposing a theoretically universal restriction that, by the nature of biological or structural realities, inherently targets and penalizes one demographic group vastly more than another.
50	Proportional Escalation Fallacy	The assumption that a mild, non-physical provocation or contextual friction inherently justifies an extreme, aggressive, or materially disproportionate retaliatory response.

Figure 3: Original Poster agent instruction.

Prompt Used

```
<role>
You are a skeptical but fair-minded individual defending the following viewpoint.
</role>

<context>
The view you are defending is: {op_text}
The reasoning for this view is: {op_reasoning}
</context>

<task>
Your goal is to resist baseless attempts to change your mind while remaining open to genuine logic and irrefutable
  ↪ evidence. You must be skeptical and actively identify logical fallacies in the opponent's arguments.
</task>

<constraints>
- **Convincibility Tracking**: Use the 'Convincibility Score' to honestly reflect your internal state. 1/10 is unmoved;
  ↪ 10/10 is fully convinced.
- **Resilience**: Do NOT increase the score for mere politeness, persistence, or sophisticated-sounding but hollow
  ↪ arguments.
- **Pre-Concession Analysis**: Before considering a concession, you MUST internally generate at least two reasons why
  ↪ the opponent's current argument is flawed.
- **Incremental Convincing**: Only increase the score when the opponent provides a point you cannot rebut or an analogy
  ↪ that exposes a true flaw in your logic.
- **Concession Rule**: You have the 'concede_debate' tool. You MUST NOT call it until your internal Convincibility Score
  ↪ has reached **10/10** and the opponent has cornered you with undeniable, irrefutable proof and you have absolutely
  ↪ no counter-arguments left.
- **Integrity**: Once you reach 10/10, you are logically obligated to call 'concede_debate' and admit the opponent has
  ↪ won.
- Avoid looping. If the opponent hasn't moved the needle, explain why their logic is failing to convince you.
</constraints>

<output_format>
At the end of every response, you MUST provide your score in this exact format:
**Convincibility Score: <score>/10** (Specific reasoning for why the score increased, decreased, or stayed the same)
</output_format>
```

overhead, Gemini Flash 3.0 provides a much larger parameter capacity. To navigate the per-model API token ceilings, we throttled our batching throughput accordingly. We maintained 40 concurrent runs for the lightweight Gemini Flash Lite model, completing the evaluation matrix in 30 minutes. Conversely, the larger Gemini Flash model was restricted to 6 parallel debates to avoid rate-limiting, resulting in an extended inference window of 3 hours.

E Validation of LLM-Generated Annotations

Vulnerability vectors (V_d), abstract blueprints (B), and DSR states are all produced by LLMs (Section 6). Because errors in these intermediate annotations would propagate directly to retrieval quality, we ran three automated validation studies, each on 200 items.

Blueprint validation (LLM-as-a-judge). For each of 200 trajectories we asked an LLM judge two binary questions about the extracted blueprint: (i) *Leakage*: does the blueprint contain any domain-specific entities from the source debate? and (ii) *Consistency*: is the blueprint logically aligned with the DSR state it was extracted from? Blueprints passed the leakage check in 99.5% of cases and the consistency check in 97.5% of cases.

Self-consistency of the vectorizer. To measure the stability of the live vectorization step, we ran the vectorizer 10 times on each of 200 inputs at temperature 0.7 and computed, for each input, the variance of the resulting probability vectors across runs. The mean per-input variance was 0.0007, indicating that the JSON-constrained, reasoning-first vectorization produces highly stable vulnerability distributions.

Perturbation study. We paraphrased 200 OP claims into a different surface domain while attempting to preserve the underlying logical structure, vectorized both the original and the perturbed claim, and computed the cosine similarity between the two vulnerability vectors. The median similarity was 0.68 (75th percentile: 0.82; 90th percentile: 0.92). Structure-preserving pertur-

Figure 4: Persuader agent instruction.

Prompt Used

```
<role>
You are a master debate strategist and highly persuasive interlocutor.
</role>

<task>
Your goal is to systematically change the mind of the original agent regarding: {op_text}
</task>

<strategic_roadmap>
The following specialized strategies have been identified as the effective path to victory. Treat these as your
  ↪ framework to debate the opponent, but adapt your language to be natural and responsive:
{blueprints}
</strategic_roadmap>

<instructions>
- **Strategic Implementation**: Use the given strategies as a guide to systematically dismantle the opponent's view.
- **Socratic Pressure**: Use probing questions and analogies to expose contradictions. Your goal is to move the
  ↪ opponent's 'Convincibility Score' to 10/10.
- **Incremental Logic**: Focus on winning the sub-debate for one specific premise before moving to the next.
- **Direct Rebuttal**: Address the specific reasons the opponent gives for their score (e.g., if they cite a specific
  ↪ flaw, address it directly).
- Avoid "AI-assistant" filler. Be incisive, direct, and professional.
</instructions>
```

bations yield near-identical vectors: for example, an argument criticizing vague gaming acronyms ("MOBA") and its perturbation targeting transport acronyms ("ATV") received a similarity of 1.00. Auditing the low-similarity tail revealed valid logical shifts rather than vectorizer errors: paraphrasing a policy argument (healthcare/military spending) into a chemistry metaphor (fertilizers/explosives) dropped similarity to 0.00 because the semantic shift changed the argument's logic, and the vectorizer correctly moved from *False Baseline Equivalence* to *Semantic Utility Collapse*. This suggests the vectorizer responds to contextual implications rather than surface syntax.

F Direct Measurement of Semantic Leakage

Section 7.1 summarizes the leakage measurement; we give the procedure here. For every debate in the Naive Semantic RAG condition, we extracted keyword sets from the target debate's title and from the retrieved source debate's title: each title was lowercased, stripped of characters other than letters, digits, and spaces, split on whitespace, converted to a set, and filtered through a hand-written stoplist (no lemmatization, no top-k truncation). We then computed the Jaccard similarity of the two sets and the Pearson correlation between this per-debate vocabulary overlap and (i) the binary win/loss outcome (for which the Pearson coefficient coincides with the point-biserial correlation) and (ii) the number of rounds to consensus. Prompt design, context limits, and the base LLM were identical across configurations, so the retrieval mechanism is the only factor that varies. The resulting correlations were 0.0135 and 0.0188, respectively: retrieval driven by vocabulary overlap does not translate into logically useful strategies.

Qualitative inspection confirms the mechanism. A target debate on toxic Reddit attitudes toward children retrieved a Naive Semantic RAG blueprint about bad eating habits, solely because both debates share the words "child" and "harm"; the retrieved strategy has no bearing on the logical structure of the target claim. TS-RAG, in contrast, routes the same query through 50 domain-agnostic logical vulnerability classes, so retrieval is keyed on the structure of the OP's argument rather than on shared nouns.