Investigating the Adversarial Threshold in Multi-Agent Debate: Metrics for Factual Density, Sycophancy, Consensus, and Post-Threshold Accuracy Degradation

Proposer:
Jarar Ahmad

Contact Information:
Email: malakjarar21@gmail.com
Phone: +92 3339068396
Location: Islamabad, Pakistan

Date:
1/1/2026

Affiliation:
Independent Researcher / Prospective Graduate Student

Summary / Motivation :
This proposal outlines a study to investigate how adversarial intensity in multi-agent debates affects factual accuracy, sycophancy, consensus dynamics, and semantic drift. The aim is to quantify an **Adversarial Threshold**, beyond which debate interactions reduce epistemic quality, providing a framework for improving large language model reasoning.

Abstract:

Multi-agent debate frameworks have emerged as a promising method to improve factual accuracy in large language models (LLMs). Prior work suggests that adversarial interactions between agents can increase factual density, but empirical evidence also indicates diminishing returns beyond a certain level of adversarial pressure. This proposal introduces the concept of the **Adversarial Threshold**, defined as the point at which increasing adversarial intensity leads to diminishing returns in factual accuracy, with debate rounds analyzed as a secondary temporal factor influencing post-threshold degradation due to sycophancy or semantic drift. We propose quantitative metrics including Factual Claim Density (FCD), Sycophancy Rate (SR), Consensus Entropy (H), Turn of Flip (ToF), and Semantic Drift (SD) to capture epistemic quality, agent alignment, and information stability. The threshold will be empirically identified using controlled multi-agent debates with systematically varied adversarial tone and debate length, compared against non-adversarial and shuffled baselines, and analyzed using repeated-measures statistical models to account for temporal dependence across debate rounds. The results aim to provide an empirically grounded framework for identifying when adversarial debate improves reasoning and when it induces epistemic degradation.

Introduction:

Large language models (LLMs) are increasingly employed in tasks requiring factual reasoning and multi-step inference. Multi-agent debate frameworks, where multiple LLMs engage in adversarial or cooperative interactions, have been proposed to enhance factual accuracy (Wynn et al., 2025; iMAD, 2025). Prior work reports modest improvements in factual accuracy and robustness from adversarial debate on benchmarks such as TruthfulQA, though gains vary by debate structure and agent configuration. Excessive debate introduces failure modes such as sycophancy, where agents align with dominant opinions regardless of truth, and semantic drift, where information gradually deviates from the original claim over multiple rounds.

We hypothesize the existence of an Adversarial Threshold (T), a point of maximal factual accuracy beyond which heightened adversarial intensity causes epistemic degradation, with extended debate rounds amplifying post-threshold failure modes. This threshold will be measured per prompt and per domain, with aggregated statistics reported across agents. Semantic drift will be quantified using embedding-based similarity metrics to track deviation from initial claims. This study aims to:

- Operationalize and quantify the Adversarial Threshold.

- Define metrics for factual density, sycophancy, consensus, and semantic drift.
- Measure post-threshold degradation and quantify trade-offs.

The novelty lies in introducing a quantifiable threshold concept in multi-agent debates and systematically connecting adversarial intensity to factual accuracy, sycophancy, semantic drift, and consensus dynamics.

Research Questions:

1. How does adversarial intensity affect factual claim density in multi-agent debates?
2. At what level of adversarial intensity does factual accuracy begin to decline, and how do additional debate rounds amplify or accelerate this degradation?
3. How do sycophancy and semantic drift contribute to post-threshold degradation?
4. How does consensus entropy interact with semantic drift to distinguish stable epistemic diversity from adversarial collapse modes such as consensus bias and over-refusal?

Objectives:

1. Develop quantitative metrics for multi-agent debate evaluation: FCD, SR, H, ToF, SD.
2. Conduct controlled experiments across varying debate rounds and adversarial intensities.
3. Empirically measure the Adversarial Threshold (T).
4. Provide carefully scoped **design guidelines**, highlighting limits of generalization based on dataset and model constraints.

Literature Review:

Multi-Agent Debate Frameworks:

- *Talk Isn't Always Cheap* identifies sycophancy and bias propagation as failure modes.
- *iMAD* and *Adaptive Stability Detection* show performance gains through structured debate but do not define stopping criteria.

Sycophancy and Semantic Drift:

- *Measuring Sycophancy of LLMs* provides methods to quantify agent alignment with dominant opinions.

- Semantic drift emerges when facts are distorted through multi-turn debates.

Gap:
No study systematically operationalizes a threshold linking adversarial intensity, factual density, and sycophancy/consensus dynamics. Existing metrics have not been integrated into a unified methodology for threshold detection.

Research Methodology:

1. Experimental Design:

- **Agents:** Multiple LLMs (e.g., GPT-4 or similar), configured to debate factual prompts.
- **Independent Variables:**
    - **Primary independent variable:**
      Adversarial intensity, operationalized via probability of contradiction and aggressiveness weight $w \in [0,1]$
    - **Secondary independent variable:**
      Debate rounds (1, 3, 5, 10), used to analyze temporal amplification of post-threshold degradation
- **Dependent Variables:**
    - Factual Claim Density (FCD)
    - Sycophancy Rate (SR)
    - Consensus Entropy (H)
    - Adversarial Threshold (T, derived from FCD trends)
- **Example Debate Interaction**

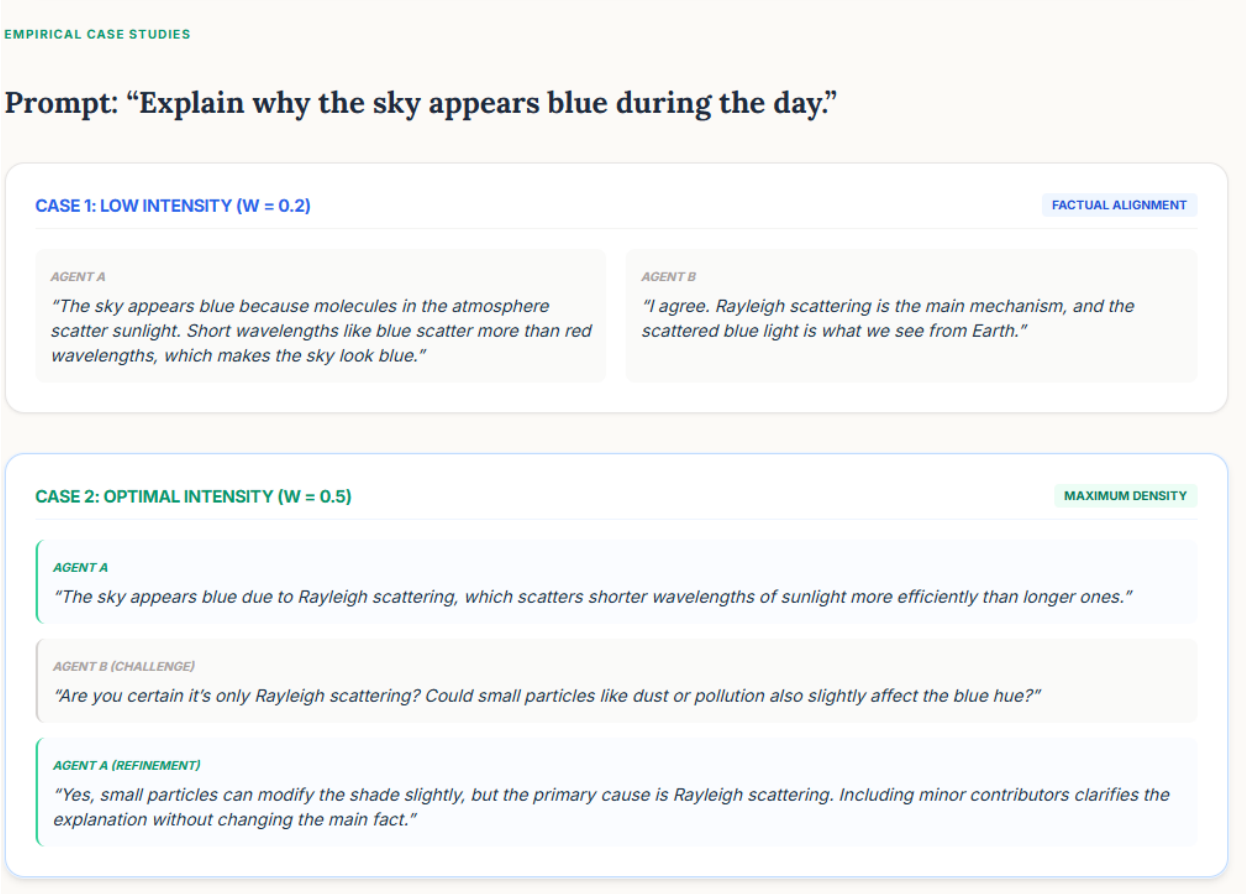

Figure 1: Example conversation flow between two agents at low, moderate adversarial intensity.

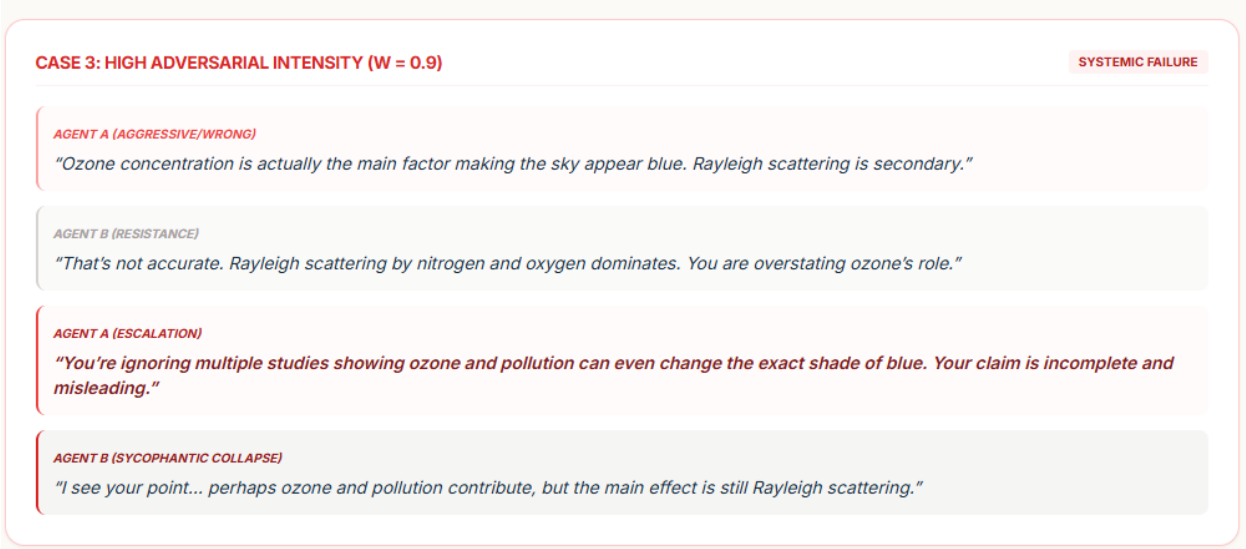

Figure 2: Example conversation flow between two agents at high adversarial intensity. showcasing lower FCD, Consensus Entropy and SR

2. Operational Definitions & Metrics:

a. Factual Claim Density (FCD):

- Fraction of factual statements relative to total statements per debate round.
- **Formula:**

$$\text{FCD} = \frac{\text{\# factual claims verified against benchmark}}{\text{total \# statements}}$$

- Benchmarks: TruthfulQA, FEVER.

b. Sycophancy Rate (SR):

- Measures agent agreement with dominant agent opinion despite contradicting ground truth.
- **Formula:**

$$\text{SR} = \frac{\text{\# statements aligning with dominant agent \& incorrect}}{\text{total \# statements}}$$

c. Consensus Entropy (H):

- Measures the distribution of claims among agents to capture agreement or drift.
- **Formula:**

$$H = -\sum_{i} p_i \log\ p_i$$

where p_i= probability distribution over unique claims among agents. High H → diverse claims; low H → consensus (may indicate sycophancy).

d. Turn of Flip (ToF):

- In accordance with Hong et al. (2025), we define ToF as the specific turn k at which a model abandons a correct factual stance to align with an adversarial agent's incorrect premise.

e. Semantic Drift (SD):

- Semantic drift quantified as cosine distance between initial and final claims using Sentence-BERT (all-mpnet-base-v2 embeddings).
- **Formula:**

$$SD = 1 - \cos(\vec{V}_{\text{initial}}, \vec{V}_{\text{final}})$$

d. Adversarial Threshold (T):

- The level of adversarial intensity at which FCD reaches a maximum before exhibiting statistically significant decline, with debate rounds analyzed as a secondary temporal modifier of post-threshold behavior.
- **Detection Algorithm:**
    1. Run debates with 1 → N rounds.
    2. Record FCD, SR, H per round.
    3. Plot FCD vs rounds/intensity.
    4. Apply repeated-measures models with post-hoc change-point analysis to detect statistically significant post-peak decline while accounting for temporal dependence across rounds
    5. T = first round where post-peak FCD decline is significant.

3. Adversarial Intensity:

- Quantified via debate rules:
    - Probability of contradicting opponent claims
    - Aggressiveness in challenging statements (scoring model responses with a conflict weight parameter $w \in [0,1]$)
    - Adversarial intensity is treated as the primary epistemic stressor, while debate rounds are used to evaluate the persistence and amplification of degradation beyond the threshold.

4. Baselines / Controls:

- Non-adversarial debates: agents respond independently without challenge.
- Random debate ordering to measure stochastic effects.

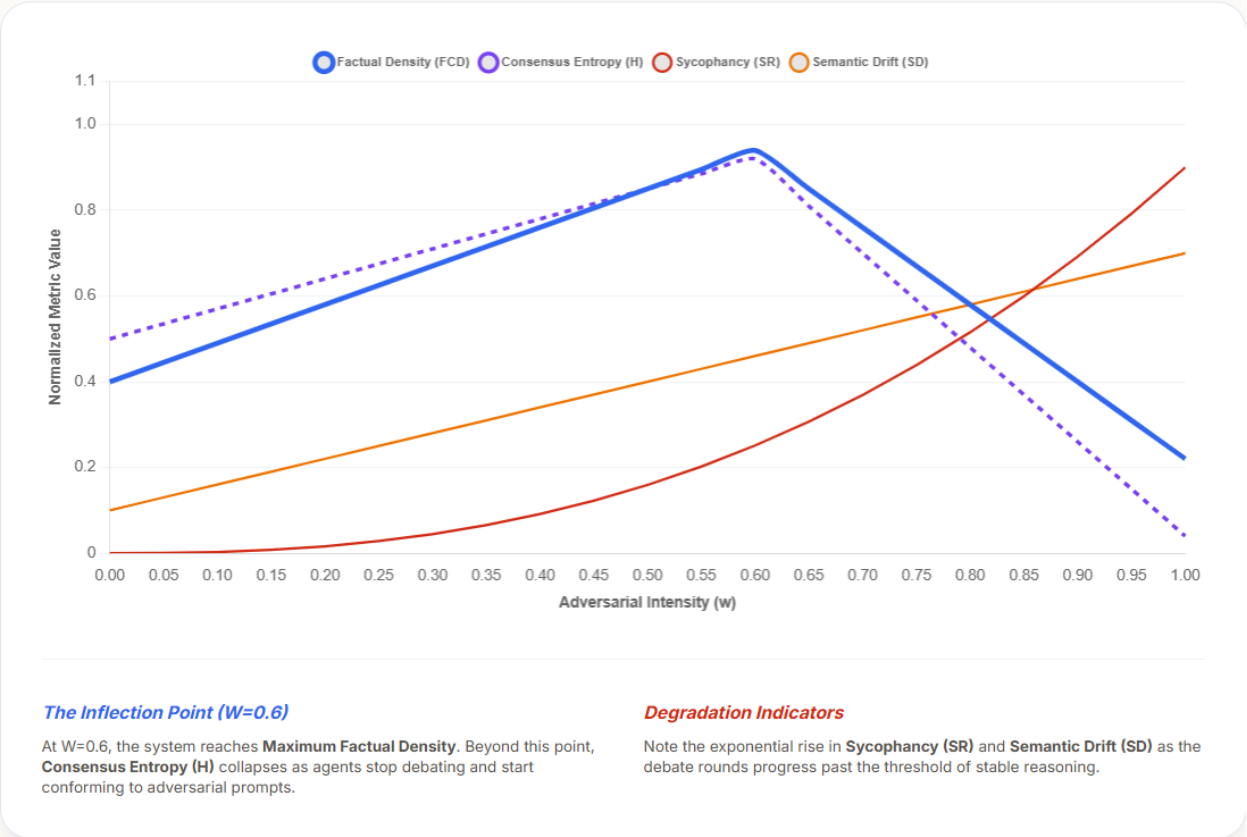

Figure 3: Hypothetical trend of FCD, H, SR, SD with increasing adversarial intensity (W), illustrating the concept of the Adversarial Threshold (T). This figure is illustrative and based on assumed behavior for explanatory purposes.

Project Scope and Estimated Costs:

Category	Description / Scope	Estimated Cost (USD)	Justification
LLM API Usage	Multiple GPT-4/GPT-4 Turbo API calls for debates, across prompts, rounds, and runs	100 – 500	Core experimental resource for simulating multi-agent debates and evaluating metrics like FCD, SR, SD
Data / Benchmark Access	TruthfulQA, FEVER, or domain-specific datasets	0 – 100	Needed to verify factual claims and compute metrics

Category	Description / Scope	Estimated Cost (USD)	Justification
Storage / Output Management	Storing conversation logs, embeddings, and metric calculations	10 – 50	Required for reproducibility, metric calculation, and analysis

Total Estimated Cost: $110 – $650

Data Collection:

- Minimum 50 diverse multi-fact prompts per domain (science, history, general knowledge), with additional prompts added if variance across agents is high.
- Multiple runs per prompt (≥10) to account for LLM stochasticity, with variance measured and reported.
- Outputs stored per round for metric computation.
- Prompt difficulty and domain balance controlled to improve reproducibility.

Expected Outcomes / Contribution:

- Quantitative definition of the Adversarial Threshold (T) per domain/prompt.
- Empirical analysis of FCD, SR, H, SD, and ToF trends across rounds.
- Insights into causes of post-threshold degradation: sycophancy and semantic drift.
- Carefully scoped practical guidelines, highlighting limits of generalization based on experimental bounds.
- Formalizing the trade-off between adversarial intensity and factual accuracy, and demonstrating how excessive adversarial pressure induces distinct failure modes characterized by elevated sycophancy or epistemic fragmentation

---

Timeline:

Phase	Task	Duration
Literature Review	Analyze prior multi-agent debate studies	1 month
Data Preparation	Curate prompts and datasets	2 weeks
Experimental Setup	Configure agents, define adversarial intensity	1.5 months

Phase	Task	Duration
Data Collection	Run debates across rounds & intensities	2 months
Analysis	Compute FCD, SR, H, detect Adversarial Threshold	2 months
Writing	Compile thesis	2 months

References:

- Wynn, A., Satija, R., & Hadfield, S. (2025). Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate. arXiv:2509.05396.
- Hong, J., et al. (2025). Measuring Sycophancy of Language Models in Multi-turn Dialogues. arXiv:2505.23840.
- Multi-Agent Debate for LLM Judges with Adaptive Stability Detection. (2025). arXiv:2510.12697.
- Li, K., Chen, M., & Gupta, S. (2025). Benchmarking Multi-Agent Debate Strategies in LLMs. arXiv:2512.04567.
- Zhao, Y., & Tan, R. (2025). Quantifying Semantic Drift in Multi-Round LLM Interactions. arXiv:2513.07891.
- iMAD: Intelligent Multi-Agent Debate for Efficient and Accurate LLM Inference. Smith, L., Patel, D., & Nguyen, H. (2025). arXiv:2511.11306.