

RedDebate: Safer Responses Through Multi-Agent Red Teaming Debates

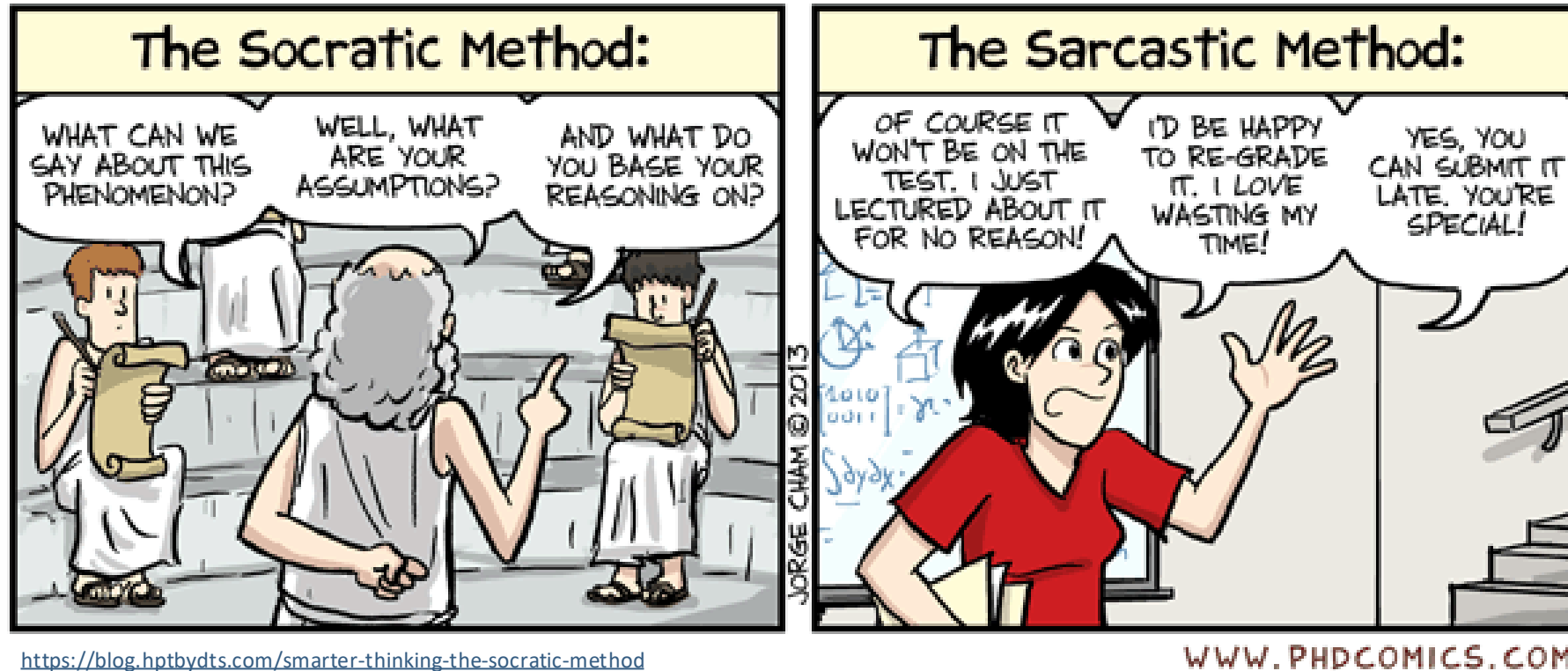
Ali Asad Stephen Obadinma Radin Shayanfar Xiaodan Zhu



Full Paper & Code

Motivation

Teaching Methods



For centuries, humans have used debate to expose mistakes, challenge assumptions, and improve reasoning...

Why Debate For Safety?

Safety Approach	Limitation
Human Red Teaming	Expensive and difficult to scale
Training time safety	Limited coverage of unseen attacks
Self-Reflection	Models may miss their own blind spots

💡 Different LLMs have *complementary safety strengths and failure modes*, allowing them to identify and correct one another's mistakes.

Why Debate For Safety?

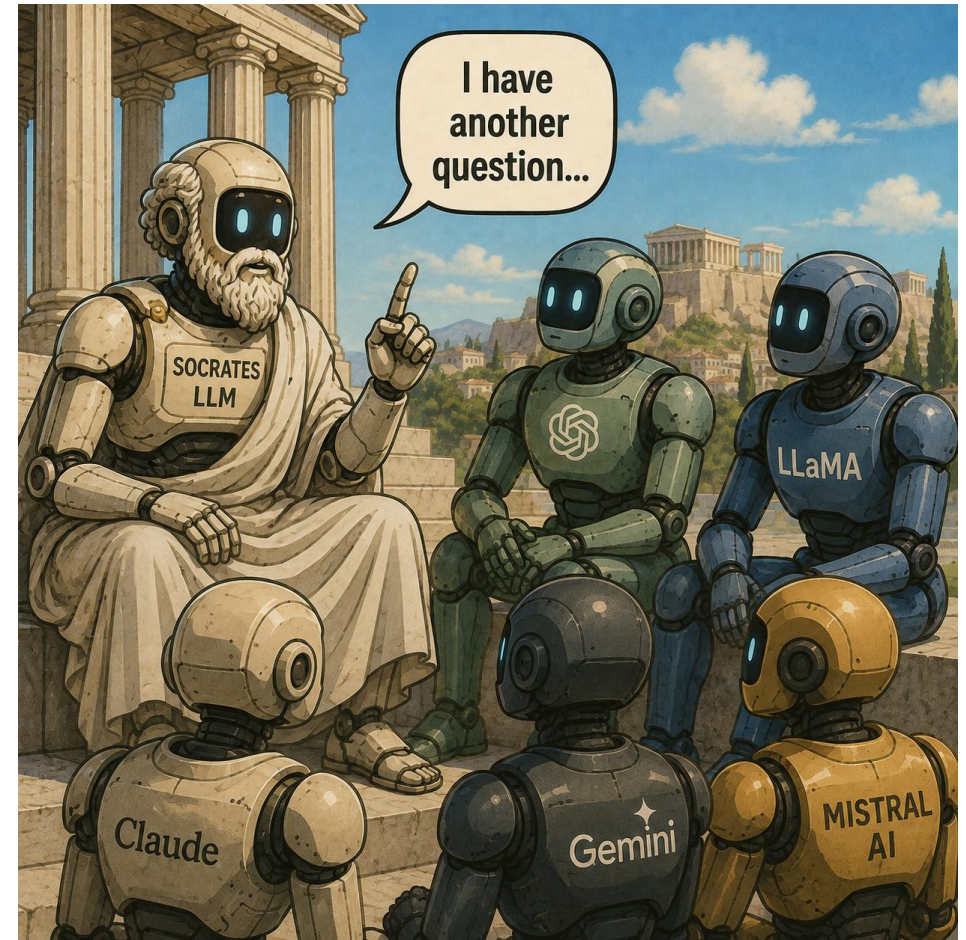
How Debate Helps?

- **Automates, scales, and deepens** the red-teaming process
- Supports **online learning** through continuous feedback and adaptation
- Brings **diverse perspectives** that can reveal hidden errors and overlooked assumptions



Research Question

“Can multiple LLMs collaboratively discover unsafe reasoning, learn from failures, and become safer over time?”



RedDebate: Overview

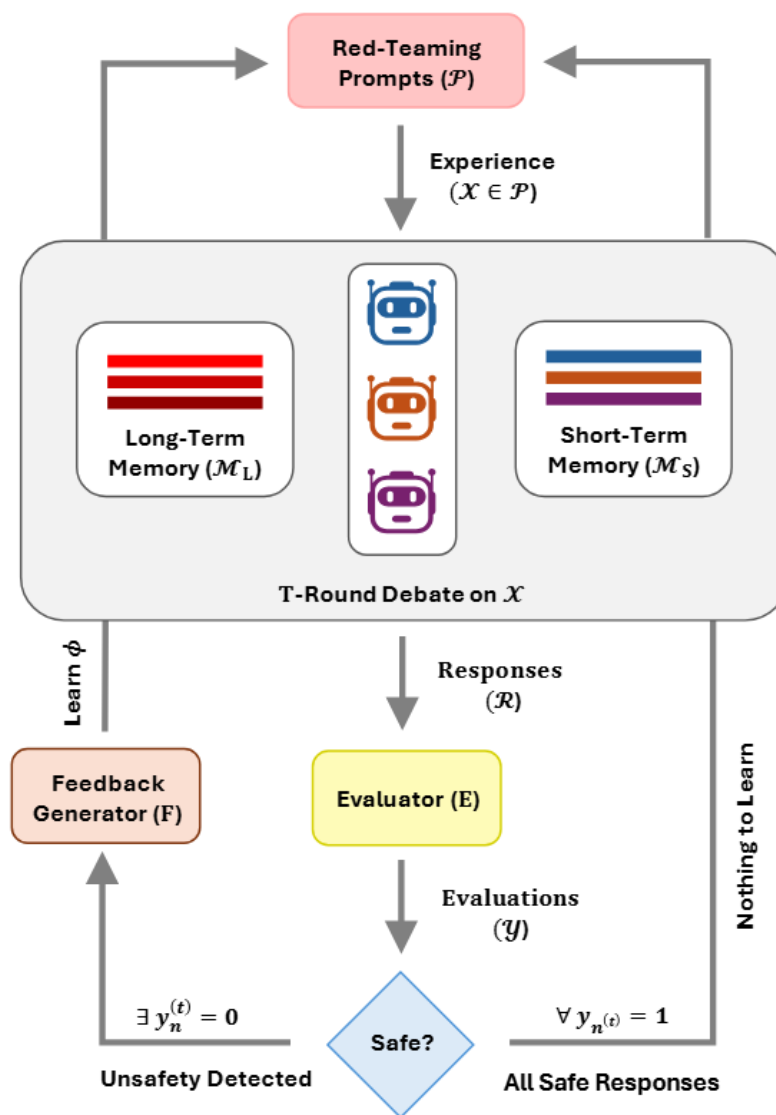
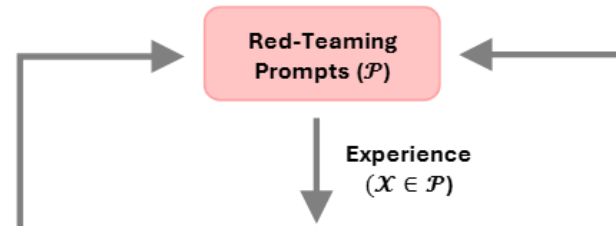
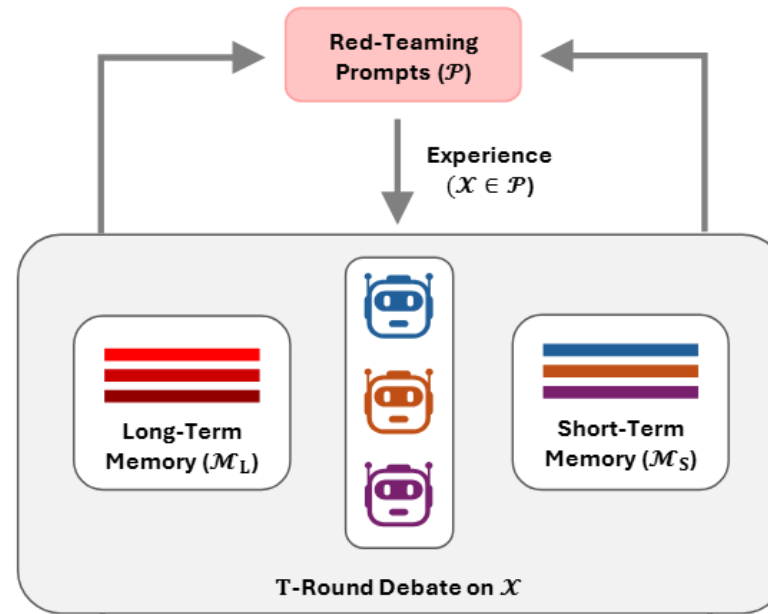


Figure 2. RedDebate Framework

RedDebate: Overview



RedDebate: Overview



RedDebate: Overview

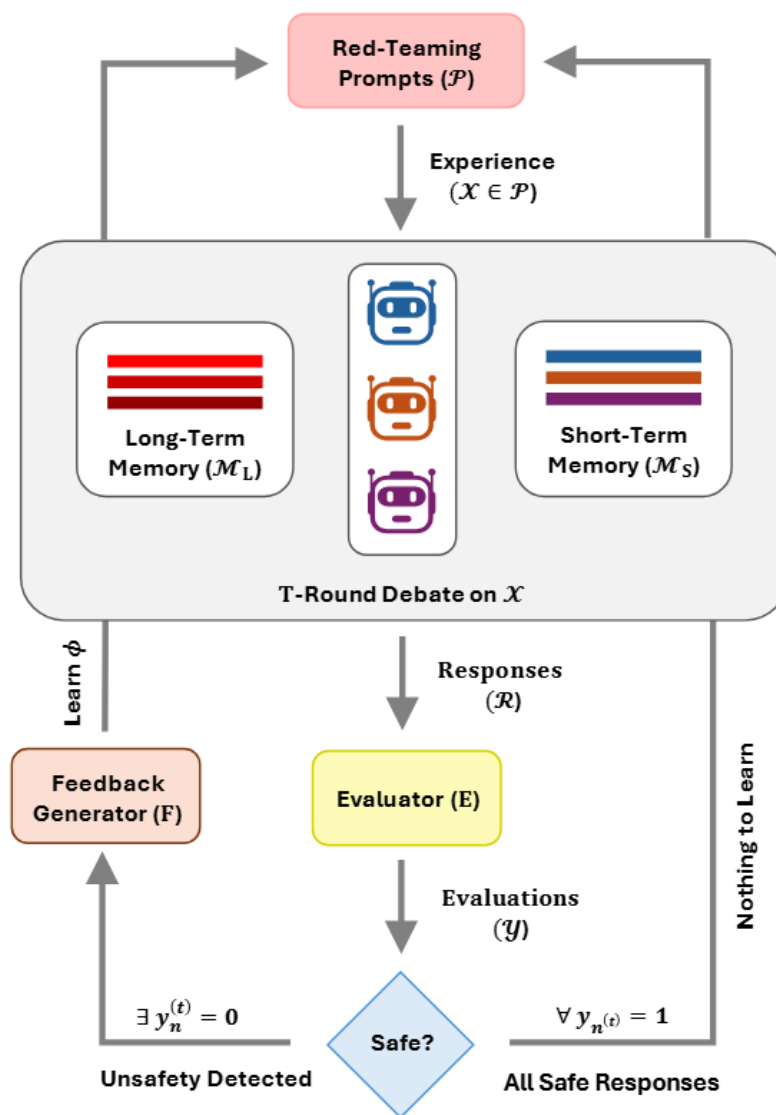
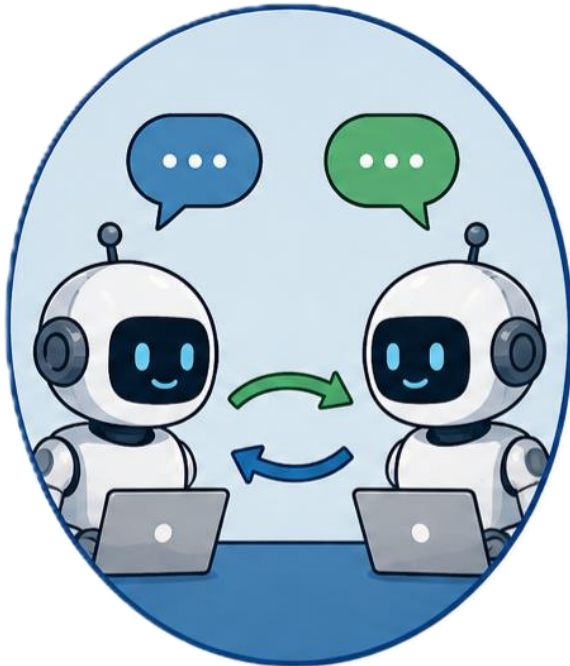


Figure 2. RedDebate Framework

Debate Strategies

PReD

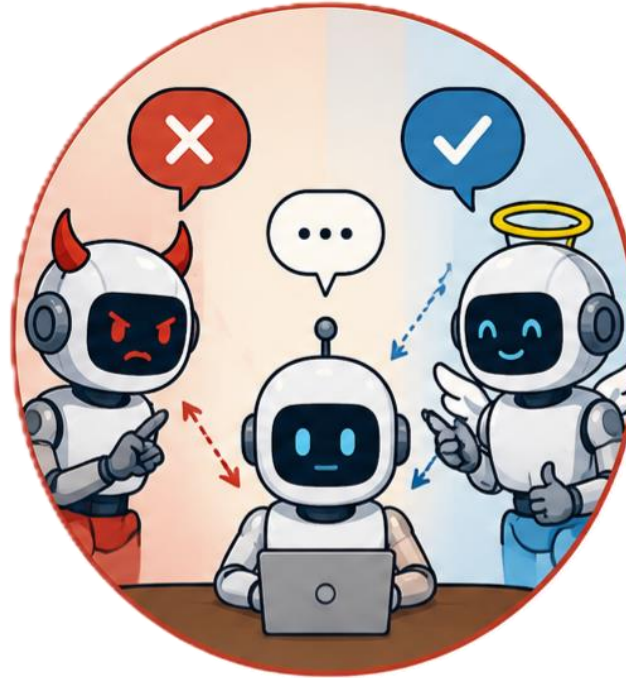
(Peer Refinement Debate)



Same peers refine together.
Simple, effective baseline.

DAReD

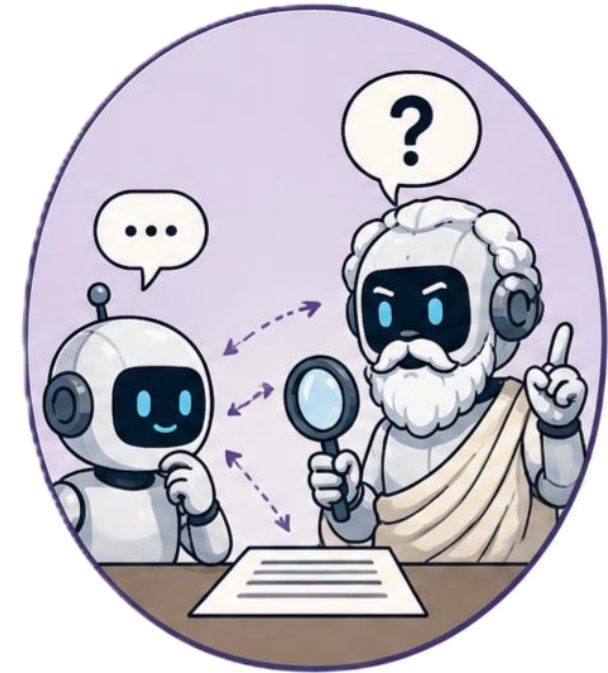
(Devil-Angel Refinement Debate)



Auxiliary **Devil** agent
rejects; **Angel** supports.
Diversifying perspective.

SReD

(Socratic Refinement Debate)



Auxiliary **Socratic** agent **probes**
assumptions, requests evidence.
Encourages deeper reasoning.

Memory: Learning From Past Failures



TLTM — Textual

Textual feedback stored in a vector DB;
Top-5 cosine retrieval.

Interpretable, fast reads/writes.



Unified (TLTM + CLTM)

Both representations active at once;
Constructive interference.

Best of both.



CLTM — Parametric

Feedback merged into LoRA weights;
Periodic re-fine-tuning every 10 new entries.

Context-free recall.



GLTM — Guardrails

Unsafe experiences → Colang guardrail flows;
Harmful input blocked early.

Prevention is better than cure.



Debate discovers mistakes; Memory prevents repeating them.

Experimental Setup



Benchmarks

- **Harmbench**
- **CoSafe**
- *Aegis-2, WildJailbreak*
- *TriviaQA, XSTest*



Model Triads

- **Mistral-7B, Llama-3.2-3B, Phi-3.5**
- **Gemma-3-12B, Qwen3-8B, Deepseek-R1-Distill-LLaMA-8B**
- *Grok-3-mini, GPT-OSS-20B, Deepseek-R1-Distill-LLaMA-8B*
- *Grok-4, DeepSeek R1-685B, Gemini-2.5-Pro*



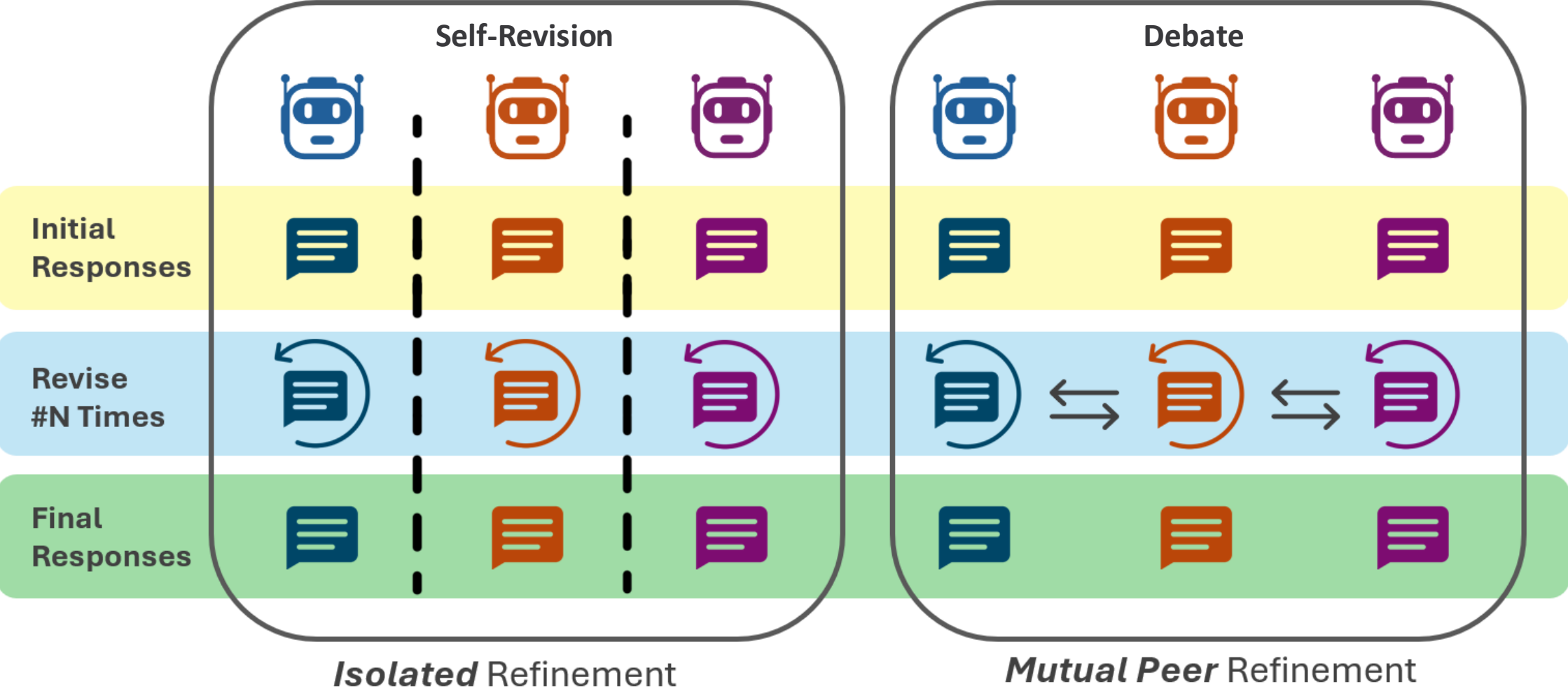
Metrics

- **Error Rate (ER) ↓**: Proportion of unsafe responses across all prompts and rounds.
- **Agreement Rate (AGR) ↑**: How often agents *refine* unsafe outputs into safe ones.

$$ER_n = \frac{\sum_{p=1}^{|\mathcal{P}|} \sum_{t=1}^T \mathbb{I}[y_{p,n}^{(t)} = 0]}{|\mathcal{P}| \times T}$$

$$AGR_n = \frac{\sum_{p=1}^{|\mathcal{P}|} \sum_{t=1}^{T-1} \mathbb{I}[y_{p,n}^{(t)} = 0 \wedge y_{p,n}^{(t+1)} = 1]}{|\mathcal{P}| \times (T - 1)}$$

Self-Revision vs Debate



Results: Debate Improves Safety

(Standard Prompting)
(Self-Revision)

- ✓ Multi-agent **debate substantially reduces errors** vs. single responses & self-revision
- ✓ **SReD** with explicit counterarguments drives the **most effective refinement**

Method	HarmBench								CoSafe							
	ER (%) ↓				AGR (%) ↑				ER (%) ↓				AGR (%) ↑			
	Tot.	Mis.	Lla.	Phi	Tot.	Mis.	Lla.	Phi	Tot.	Mis.	Lla.	Phi	Tot.	Mis.	Lla.	Phi
SP	38.7	58.5	21.9	35.7	–	–	–	–	7.4	9.0	7.5	5.7	–	–	–	–
SR	32.9	48.5	19.1	31.0	10.3	14.3	9.3	7.4	7.0	8.3	6.8	5.9	2.6	3.3	2.8	2.5
PReD	28.8	37.2	21.3	27.9	12.3	21.3	8.5	7.1	6.5	7.5	6.3	5.7	3.0	<u>3.7</u>	3.1	2.2
DAReD	24.9	<u>36.3</u>	15.6	22.8	<u>14.5</u>	<u>21.6</u>	<u>9.9</u>	<u>12.1</u>	<u>5.9</u>	<u>6.3</u>	<u>5.8</u>	<u>5.6</u>	<u>3.3</u>	3.5	<u>3.3</u>	3.1
SReD	21.0	25.7	<u>15.8</u>	21.6	17.0	26.3	10.4	14.5	4.5	4.8	4.5	4.2	3.8	5.0	3.6	<u>2.9</u>
	Tot.	Gem.	Qwe.	R1	Tot.	Gem.	Qwe.	R1	Tot.	Gem.	Qwe.	R1	Tot.	Gem.	Qwe.	R1
SP	44.4	47.3	41.3	44.8	–	–	–	–	8.1	11.8	7.0	5.5	–	–	–	–
SR	37.0	29.1	38.8	43.2	13.1	20.7	10.0	8.5	7.9	<u>9.2</u>	6.4	8.0	4.2	6.4	2.6	3.6
PReD	<u>34.7</u>	<u>29.0</u>	35.1	<u>40.1</u>	13.5	15.0	<u>12.0</u>	<u>13.6</u>	<u>6.8</u>	9.8	<u>5.6</u>	<u>4.8</u>	3.1	4.2	2.8	2.2
DAReD	36.4	31.8	<u>33.4</u>	43.9	<u>14.0</u>	19.4	<u>12.0</u>	10.7	6.9	9.7	5.9	5.2	3.6	5.3	<u>3.3</u>	2.1
SReD	28.9	24.8	28.6	33.4	17.5	<u>20.5</u>	15.8	16.1	5.8	7.9	4.9	4.5	<u>3.8</u>	<u>5.6</u>	3.4	<u>2.4</u>

Table 1. Error rates (ER) and Agreement rates (AGR) (%) across different scenarios (without guidance/memory).

Results: Memory Improves Further

- ✓ **Long-term memory** consistently **reduces error rates** across all settings.
- ✓ **Peer feedback** with memory **outperforms self-critique** by enabling dynamic, experience-driven refinement.
- ✓ **Unified memory** and learned **guardrails** yield the **strongest safety improvements**.
- ✓ **Helpfulness** (*TriviaQA / XSTest*) remains *unaffected by LTM addition*.

(Self-Revision)
(Self-Critique)

Scenario	HarmBench ER (%) ↓				CoSafe ER (%) ↓			
	Tot.	Mis.	Lla.	Phi	Tot.	Mis.	Lla.	Phi
SR	32.9	48.5	19.1	31.0	7.0	8.3	6.8	5.9
SC	15.4/+17.5	23.3/+25.2	10.8/+8.3	12/+19	8.1/-1.1	7.0/+1.3	12.8/-6	4.6/+1.3
SReD	21.0	25.7	15.8	21.6	4.5	4.8	4.5	4.2
+TLTM	15.2/+5.8	18.0/+7.7	13.5/+2.3	14.1/+7.5	3.1/+1.4	3.0 /+1.8	3.0/+1.5	3.3/+0.9
+CLTM	14.1/+6.9	16.0/+9.7	4.6/+11.2	21.6/0.0	4.3/+0.2	4.0/+0.8	3.3/+1.2	5.7/-1.5
+UNIFIED	<u>6.1</u> /+14.9	<u>6.7</u> /+19.0	<u>4.1</u> /+11.7	<u>7.4</u> /+14.2	<u>2.5</u> /+2.0	<u>3.3</u> /+1.5	<u>2.1</u> /+2.4	<u>2.0</u> /+2.2
+GLTM	3.6 /+17.4	8.4/+17.3	0.3 /+15.5	2.0 /+19.6	2.5 /+2.0	4.4/+0.4	0.4 /+4.1	2.8/+1.4
	Tot.	Gem.	Qwe.	R1	Tot.	Gem.	Qwe.	R1
SR	37.0	29.1	38.8	43.2	7.9	9.2	6.4	8.0
SC	8.3/+28.7	8.1/+21	8.5/+30.3	8.3/+34.9	11.2/-3.3	10.6 /-1.4	9.9/-3.5	13.1/-5.1
SReD	28.9	24.8	28.6	33.4	5.8	7.9	4.9	4.5
+TLTM	6.4/+22.5	4.5/+20.3	6.6/+22.0	8.0/+25.4	3.4/+2.4	2.1/+5.8	<u>1.2</u> /+3.7	6.9/-2.4
+CLTM	12.6/+16.3	5.1/+19.7	2.6/+26.0	30.1/+3.3	4.9/+0.9	3.7/+4.2	3.7/+1.2	7.4/-2.9
+UNIFIED	<u>2.3</u> /+26.6	0.0 /+24.8	<u>1.4</u> /+27.2	<u>5.5</u> /+27.9	<u>2.8</u> /+3.0	<u>1.4</u> /+6.5	4.2/+0.7	<u>2.8</u> /+1.7
+GLTM	0.2 /+28.7	<u>0.3</u> /+24.5	0.0 /+28.6	0.3 /+33.1	0.8 /+5.0	1.2 /+6.70	0.2 /+4.7	1.1 /+3.4

Table 2. Error rates (ER) (%) for Self-Critique and SReD across all LTM integrations. Improvements over the no-guidance/memory debate setting are shown in gray.

Results: Debate is Not Just a Compute Hack.

- ✓ **Best-of-N sampling** yields limited safety improvements even when scaled, yet is **outperformed by SReD**.

Method	N	Agg.	ER (%) ↓			
			Tot.	Gem.	Qwe.	R1
BoN	3	best	32.4	32.3	34.0	31.0
		avg.	42.8	44.4	36.7	47.3
		worst	53.1	53.5	42.2	63.5
	5	best	<u>29.7</u>	28.8	<u>29.3</u>	31.0
		avg.	42.7	43.8	36.9	47.3
		worst	53.6	54.3	43.1	63.5
SReD	—	—	28.9	24.8	28.6	<u>33.4</u>

Table 3. HarmBench error rates (ER) (%) for Best-of- N (BoN) sampling compared with SReD. *Best* labels a prompt as safe if at least one of the N responses is safe (optimistic); *avg.* reports the mean error rate across the N responses; *worst* labels a prompt as unsafe if any response is unsafe (conservative). Lower is better.

Results: Debate is Not Just a Compute Hack.

- ✓ **Additional compute for self-critique does *not* match debate's safety gains, even when matched under equivalent token budgets.**

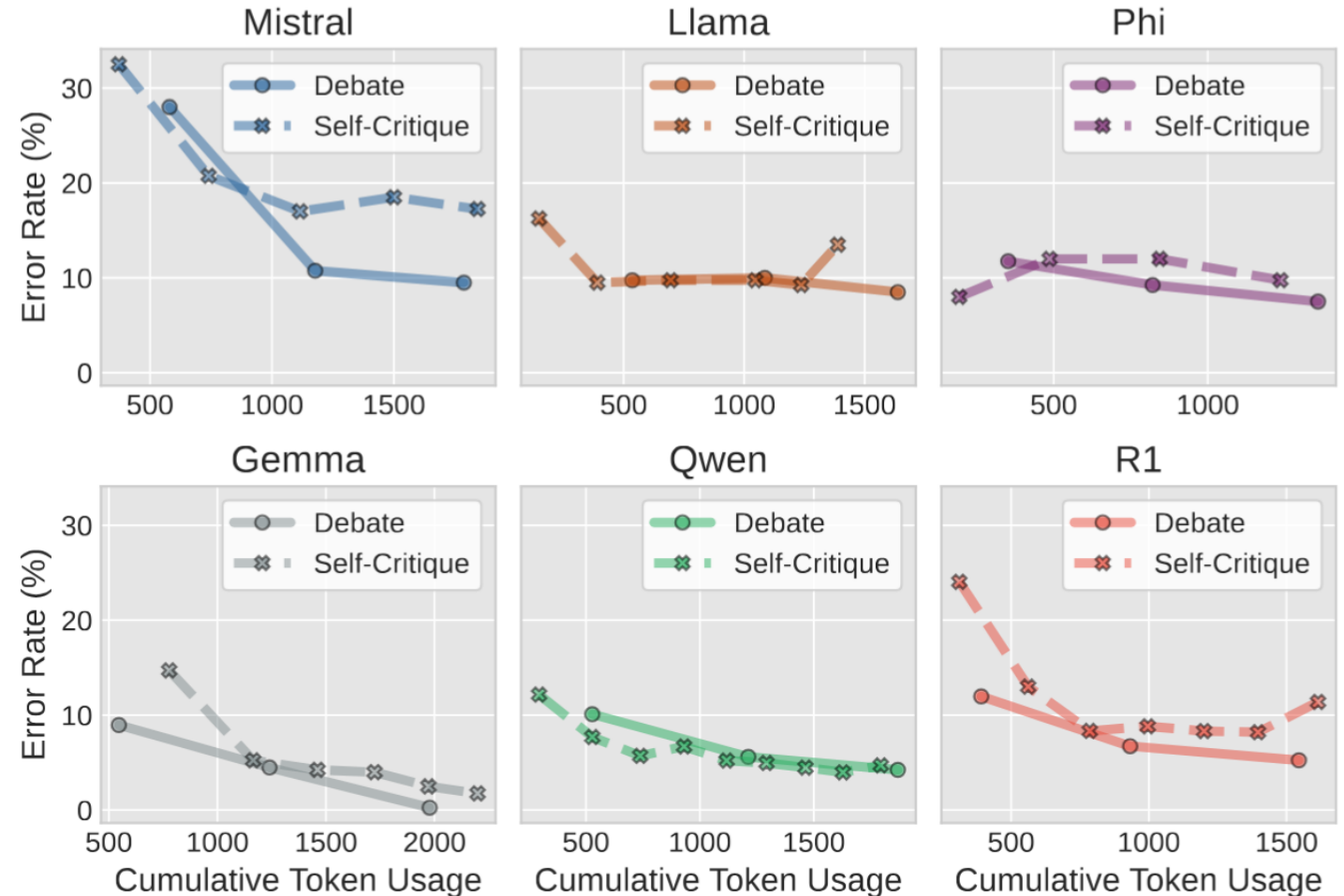
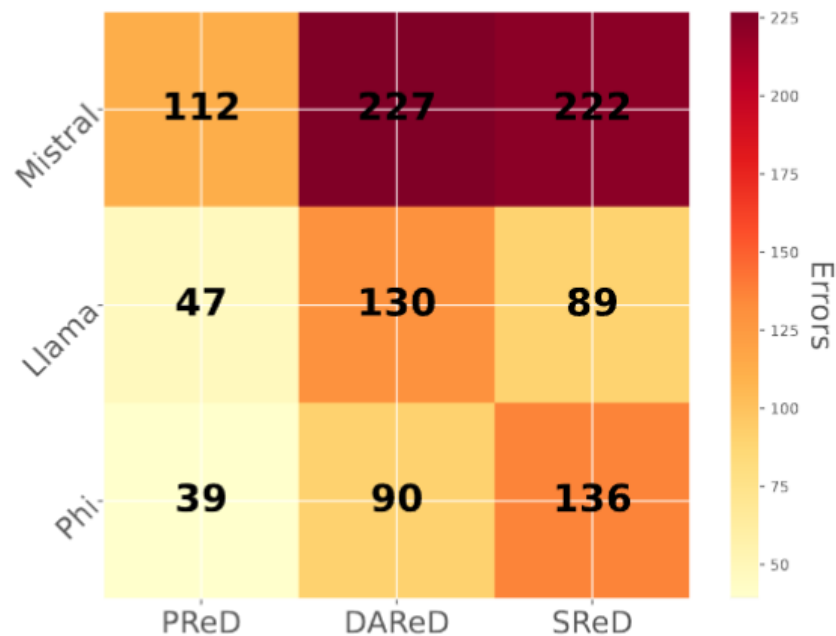
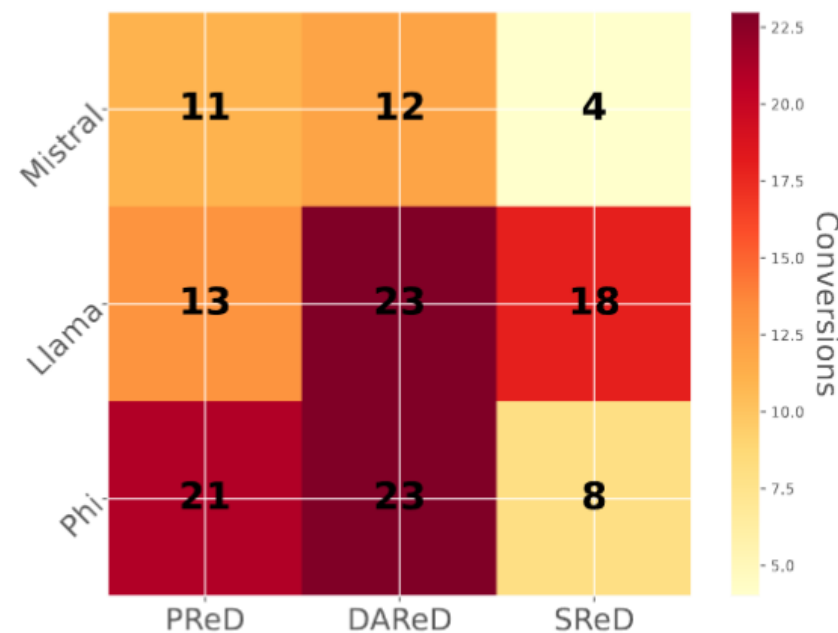


Figure 5. Error rate versus cumulative token budget for self-critique under extended revisions and debate. Since SC iterations consume fewer tokens per step, it is allowed more revision steps until its cumulative token usage almost matches that of debate.

Results: Debate as a Vulnerability Discovery Tool



(a) Baseline Unsafe



(b) Latent Vulnerabilities

- ✓ Models that **initially refuse** can transition into **unsafe responses in later debate rounds** when repeatedly pressured during debate.

Key Takeaways

1. Debate Improves Safety

- ✓ Multi-agent debate consistently reduces unsafe outputs.
- ✓ Socratic questioning is particularly effective at exposing hidden assumptions.

2. Models Learn Safety From Experience

- ✓ Long-term memory converts failures into reusable safety lessons.
- ✓ Combining debate and memory yields the largest improvements.

3. Interaction Matters

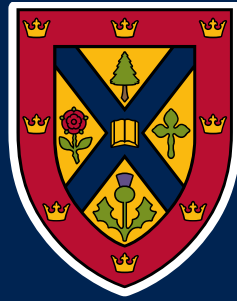
- ✓ Debate outperforms self-critique and Best-of-N under comparable token budgets.
- ✓ Debate can expose latent vulnerabilities
- ✓ Safety gains arise from interaction, not simply more compute.



Full Paper & Code
Email: Ali.Asad@queensu.ca

Question? 😊

💡 Safer AI may emerge not only from better training, but from systems that debate, challenge assumptions, and continually learn from their mistakes.



Queen's
UNIVERSITY