Probabilistic Interventions for Stabilizing Multi-Agent Debate

Alexis Hu
Harvard University
alexishu@harvard.edu

Mira Yu
Harvard University
mirayu@college.harvard.edu

Charles Zhou
Harvard University
czhou@college.harvard.edu

Abstract

Multi-Agent Debate (MAD) is often proposed as a method for improving the factuality and reasoning of large language models (LLMs). Yet, recent empirical evidence has shown that standard debate protocols frequently fail to outperform strong single-agent Chain-of-Thought (CoT) baselines, potentially amplifying errors through "social drift" and over-correction. We argue that this limitation stems from the temporal structure of existing debate topologies. In particular, batch-style debate—analogous to the U.S. Congress—suffers from a "Filibuster" effect: early hallucinations propagate unchecked through long-context generations, producing obfuscated arguments that are difficult for a critic to unwind post hoc. We introduce a taxonomy of debate *interruption granularities* inspired by comparative governance: (1) American Congress (batch, post-hoc critique), (2) British Parliament (streaming, atomic interruption), and (3) Socratic Dialogue (soft, non-blocking guidance). We formalize the British Parliamentary protocol as a Prover–Estimator system in which a Verifier can halt generation at small token chunks and force recursive correction of low-confidence claims. Evaluating all three protocols on a multi-hop reasoning benchmark, HotpotQA, we find that while single-agent CoT remains the strongest overall baseline (66.1%), atomic interruption substantially improves multi-agent reliability: British Parliament outperforms the standard Congress protocol (63.3% vs. 59.2%, $p < 0.001$). Our results suggest that debate effectiveness does not purely depend on number of agents, but also on the *topology* of interaction: imposing fine-grained verification suppresses error cascades far more effectively than adversarial batch critique.

1 Introduction

Large Language Models (LLMs) remain highly susceptible to hallucination, particularly in long-context reasoning tasks where an early factual error can propagate through the remainder of the model's chain of thought. Multi-Agent Debate (MAD) has been proposed as a mechanism for mitigating such errors by introducing adversarial scrutiny between agents. However, the opinions on MAD are mixed. Some recent theoretical work suggests that debate amplifies correctness [1]. However, other recent empirical and theoretical work casts doubt on its effectiveness: debate agents frequently converge toward incorrect answers, exhibit sycophancy or "social drift," and often fail to outperform strong single-agent baselines [2].

We argue that these failures stem not from the idea of debate itself, but from the *temporal structure* of standard debate protocols. Most existing MAD systems operate in a *batch* or "post-hoc" regime: a Prover generates an entire response before a Critic is allowed to intervene. This structure, analogous

to a U.S. Congressional filibuster, creates room for error and an asymmetry of effort: a minor mistake near the beginning can result in lengthy texts of downstream reasoning that the Critic must then address. The resulting critiques are often unfocused, incomplete, or themselves misguided.

To overcome this structural weakness, we introduce an alternative topology inspired by British Parliamentary procedure. In this model, a Critic may issue a *Point of Information* (POI) (an interruption targeted at a specific sub-claim) as soon as a potentially incorrect premise appears. This converts debate from a post-hoc evaluation problem into a streaming verification process, enforcing recursive correction before an error can cascade. We empirically evaluate this approach on HotpotQA, comparing it against both single-agent Chain-of-Thought (CoT) and existing batch-style MAD baselines.

Our contributions are as follows:

- We formulate a unifying view of debate protocols as governance-inspired state machines, categorized by their *interruption granularity*: Batch (American Congress), Atomic (British Parliament), and Soft (Socratic Dialogue).
- We introduce the first empirical implementation of **atomic, streaming interruption** in MAD, enabling real-time detection and recursive correction of sub-claims as they are generated.
- We show that the British Parliamentary protocol significantly outperforms batch debate ($+4.1\%$ accuracy), supporting our hypothesis that temporal structure and not simply the presence of multiple agents determines whether debate is stabilizing or destabilizing.
- We analyze social drift, demonstrating that soft, non-blocking interventions often destabilize otherwise correct reasoning, whereas targeted interruptions reduce drift by limiting ambiguity.

2 Related Work

Multi-Agent Debate (MAD). Debate-based supervision was originally proposed as a mechanism for improving factuality and transparency in LLM reasoning [3]. Early empirical work reported encouraging results on constrained tasks such as math word problems and short reasoning chains [4]. Motivated by the limitations of single-agent self-reflection—such as the Degeneration-of-Thought phenomenon, where models become unable to generate novel corrections once committed to an incorrect solution—subsequent work proposed multi-agent debate frameworks to encourage divergent reasoning [5].

However, more recent analyses have highlighted fundamental instabilities in debate dynamics. Prior work shows that debate protocols often suffer from *social drift*, in which agents converge toward incorrect but mutually agreeable answers, and that symmetric adversarial setups may amplify rather than suppress hallucinations [2] [6]. Our findings build upon this critique by showing that the timing of the interaction, specifically whether critique is batch or incremental, is a primary driver of these failures.

Prover–Estimator and Recursive Verification. The theoretical foundation for our British Parliamentary protocol derives from work on Prover–Estimator games. Irving et al. [3] propose recursive debate as a defense against obfuscated arguments, but the formulation remained largely conceptual. Brown-Cohen et al. [7] provide a formal analysis showing that debaters must verify sub-claims at sufficiently fine granularity to avoid compounding errors. Our contribution is the empirical realization of this theory: we instantiate atomic, streaming verification in a live LLM system and show that enforcing local consistency significantly mitigates long-range hallucination cascades.

Factuality, Error Cascading, and Hallucination. A growing body of work documents the brittleness of long-chain reasoning in LLMs, where an early incorrect inference can corrupt all dependent steps [8]. Traditional solutions focus on self-consistency, external retrieval, or self-correction mechanisms, but these approaches typically operate after errors have already propagated. Our results align with the view that hallucination is fundamentally a temporal problem and that interventions must occur at the moment an error is introduced, rather than after the full answer is generated.

Dialogue-Based Correction and Socratic Methods. Several works explore softer forms of corrective interaction, such as guided questioning or Socratic prompting [9] [10]. These approaches tend to improve interpretability but offer limited protection against persistent hallucinations, as the Prover remains free to continue an incorrect line of reasoning. Our Socratic protocol evaluates this class of methods directly and confirms their susceptibility to "sycophancy": a tendency of models to over-update in response to ambiguous or leading questions.

Our Positioning. Across these threads, the gap we address is clear: prior work has compared different *roles* in debate (adversarial, cooperative, guided), but not the method of critique (e.g. time structure). We show that interruption granularity, whether verification is batch, soft, or atomic, is the decisive factor governing whether debate suppresses or amplifies error. Our British Parliamentary protocol provides the first practical demonstration that enforcing fine-grained, recursive verification materially improves robustness over standard MAD implementations.

3 Methodology: Governance Topologies as Debate Protocols

We formalize a debate protocol as a state machine:

$$\mathcal{D} = (A, B, \mathcal{Q}, \Gamma),$$

where A is the *Prover*, B is the *Critic/Estimator*, $\mathcal{Q}$ is the query, and Γ defines the transition rules governing message exchange. A debate proceeds as a sequence of states

$$\sigma_0 \xrightarrow{\Gamma} \sigma_1 \xrightarrow{\Gamma} \cdots \xrightarrow{\Gamma} \sigma_T,$$

where each state contains the partial transcript and the current control holder.

We study three topologies differing only in Γ: (1) batch critique, (2) streaming atomic interruption, (3) soft, non-blocking guidance. Naturally, the switching between which agent is speaking is modeled as a transition between debate states.

3.1 Protocol 1: The American Congress (Batch / Post-hoc Critique)

The "American Congress" topology represents the dominant structure used in prior MAD work where one agent fully generates a statement before the next can respond. In this case, the Prover generates a full answer sequence

$$y_{1:T} = (y_1, y_2, \ldots, y_T), \qquad P(y_{1:T} \mid \mathcal{Q}) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, \mathcal{Q}),$$

with no intervention from the Critic. After the sequence terminates, the Critic produces a single global critique, taking into account all of the context from the Prover:

$$c = B(y_{1:T}, \mathcal{Q}).$$

This produces a protocol with the following state transitions:

$$\sigma_0 \xrightarrow{A} \sigma_1 \xrightarrow{A} \cdots \xrightarrow{A} \sigma_T \xrightarrow{B} \sigma_{T+1}.$$

Because error detection occurs only at σ_T, any early hallucination at step k contaminates all downstream reasoning $y_{k+1:T}$. This induces the *filibuster effect*: the Prover may generate long, coherent but incorrect arguments that are expensive for the Critic to address post-hoc.

3.2 Protocol 2: The British Parliament (Streaming and Atomic Interrupts)

This topology modifies Γ to allow real-time fact-checking. The Prover produces a token stream

$$s_t = (t_1, t_2, \ldots, t_t),$$

and the Critic continuously monitors incremental chunks Δs_t. A chunk boundary is detected using the implementation rule where a chunk ends once its length is greater than 120 chars or the length is greater than 40 chars and it ends with punctuation:

$$\text{ChunkBoundary}(\Delta s_t) = (|\Delta s_t| > 120) \ \vee \ (|\Delta s_t| > 40 \wedge \Delta s_t \text{ ends with } \{., ?, !\}).$$

At each boundary, the Critic performs a binary factuality test:

$$\text{Interrupt}(\Delta s_t) = \begin{cases} 0, & \text{chunk appears factually sound,} \\ 1, & \text{suspected factual error.} \end{cases}$$

This binary decision can be interpreted as enforcing an implicit factuality threshold τ, rather than estimating an explicit probability of correctness, which is difficult to calibrate reliably for LLM outputs. Interruption triggers a transition into a recursive correction state:

$$\sigma_t \xrightarrow{\text{POI}} \sigma_t^{\text{corr}} \xrightarrow{A} \sigma_{t+1},$$

where the Critic issues a Point of Information (POI) describing the suspected error, and the Prover must either revise or defend the disputed sub-claim before resuming the main argument. Thus the state machine interleaves generation and verification:

$$\sigma_0 \xrightarrow{A} \sigma_1 \xrightarrow{B\,?} \sigma_1^{\text{corr}} \xrightarrow{A} \sigma_2 \cdots$$

This protocol enforces atomic verification: no premise may propagate until its correctness has been checked. As such, generation is periodically gated at chunk boundaries, where the Critic can issue targeted objections before the Prover proceeds. This aims to implement the anti-obfuscation guarantees of Prover–Estimator debate [7].

3.3 Protocol 3: Socratic Dialogue (Soft, Non-Blocking Guidance)

This topology preserves the Prover's left-to-right generation but equips the Critic with a non-blocking intervention. The Prover produces tokens as usual:

$$P(y_{1:T} \mid \mathcal{Q}) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, \mathcal{Q}).$$

When the Critic detects ambiguity or insufficient justification, it issues a clarifying question:

$$q_{\text{soc}}(t) = B(y_{1:t}, \mathcal{Q}),$$

which the Prover may optionally incorporate:

$$y'_t = A(y_{1:t}, q_{\text{soc}}(t)).$$

The control flow (unlike Protocol 2) never halts or rewinds:

$$\sigma_t \xrightarrow{B} \sigma'_t \xrightarrow{A} \sigma_{t+1}.$$

Socratic Dialogue therefore induces guidance without enforcement. Its lack of hard interruption makes it computationally inexpensive, but it is susceptible to social drift: Provers often over-correct when faced with open-ended questions that implicitly signal doubt.

4 Experimental Setup

4.1 Benchmarks

We evaluate all debate protocols on the HotpotQA distractor benchmark, a multi-hop question-answering dataset where errors in early reasoning steps frequently induce long-range hallucinations. All agents across all protocols use the same backbone model (GPT-4o-mini) to ensure differences arise solely from the debate topology rather than model capacity.

We conducted 27 independent runs for each protocol, each consisting of 100 sampled questions (2,700 total debate trials). A single-agent Chain-of-Thought (CoT) system serves as the baseline control condition.

4.2 Implementation of Granular Interruption

To operationalize the British Parliamentary protocol, we implement a streaming controller that mediates token-level interaction between the Prover and the Critic. Unlike batch-style debate, where both agents exchange complete messages, the Prover's output is handled as an incremental stream:

$$S = (t_1, t_2, \ldots)$$

The Critic observes the stream through a moving buffer B_t consisting of the current unfinished clause or sentence.

An interruption check is triggered only at syntactic chunk boundaries as described in Protocol 2. At each boundary, the Critic performs a lightweight factuality assessment of the buffer. If the probe flags potential error, the system injects an interruption and transfers control, allowing the Critic to issue a targeted clarification or POI. After the Prover resolves the sub-claim, normal streaming resumes. This implementation ensures that atomic interruption occurs only at syntactically meaningful boundaries, minimizing overhead while retaining the core recursive-correction behavior of the theoretical protocol.

4.3 Evaluation Metrics

We report results across three key metrics:

- Accuracy – exact match or containment of the gold answer.
- Token Efficiency – total tokens produced across all agents.
- Stability – standard error over the 27 independent runs.

5 Results

Our experiments reveal a consistent ordering of performance across the three debate topologies and the single-agent baseline. The results challenge the assumption that multi-agent interaction inherently improves reasoning: without structural constraints on when agents may intervene, debate often underperforms even strong single-agent systems.

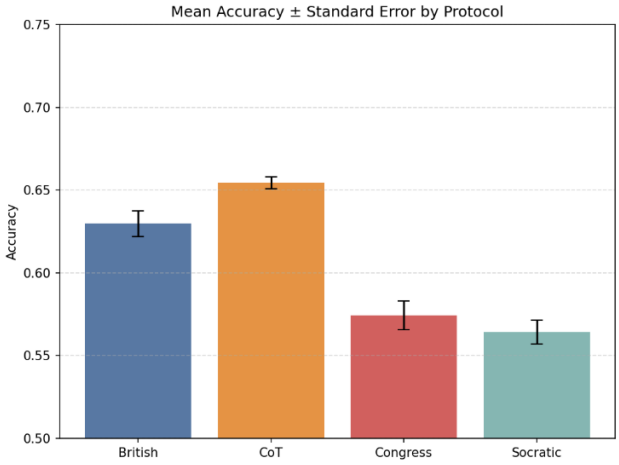

Figure 1: Accuracy across protocols on HotpotQA (100 samples per run). Chain-of-Thought remains the strongest overall, but the British Parliamentary protocol substantially outperforms the American Congress and Socratic variants.

5.1 Statistical Significance and Stability

To assess robustness, we conducted paired statistical tests across the 27 independent runs per protocol (2,700 debate trials total). Despite its higher compute cost, the British Parliamentary protocol

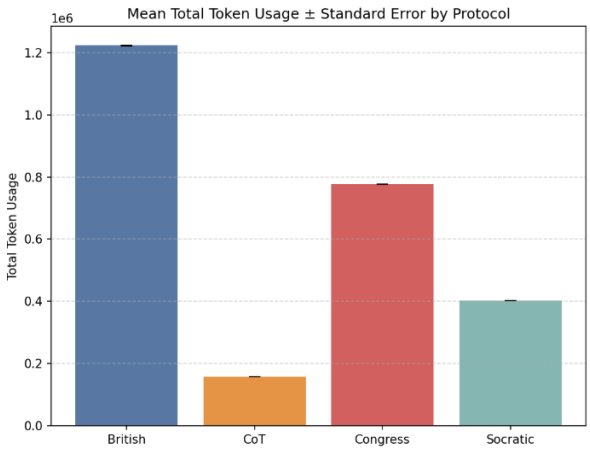

Figure 2: Mean token usage per protocol. Streaming interruption in the British protocol incurs significant overhead relative to batch-style debate.

Table 1: Accuracy and Token Usage Across Protocols (Mean $\pm$ Standard Error).

Protocol	**Accuracy (%)**	**Token Count**
Single CoT (Baseline)	$\mathbf{66.1 \pm 0.3}$	$\mathbf{157{,}635 \pm 100}$
British Parliament	63.3 ± 1.4	$1{,}224{,}000 \pm 900$
American Congress	59.2 ± 0.7	$777{,}115 \pm 1200$
Socratic Dialogue	57.3 ± 1.5	$402{,}055 \pm 120$

significantly outperforms the American Congress protocol:

$$t = 4.70, \quad p < 0.001.$$

Because each trial shares the same question set, the paired design ensures that improvements cannot be attributed to dataset variance. The results provide strong evidence that interruption granularity, not model scale or the number of agents, drives accuracy differences.

Stability differs notably across protocols. While CoT exhibits very low variance ($SE = 0.3$), both Socratic and British protocols introduce higher stochasticity due to adversarial or corrective interactions. The British protocol's variance ($SE = 1.4$) reflects the sensitivity of streaming interruption to borderline sub-claims, which can alter the number and timing of POI events.

5.2 The Cost of Interaction

Debate introduces substantial computational overhead. The British protocol consumes approximately:

$$\frac{1.22\text{M tokens}}{0.158\text{M tokens}} \approx 7.8\times$$

the token budget of CoT. Nevertheless, compared to the Congress protocol, British-style atomic interruption achieves measurably higher accuracy. This suggests a fundamental tradeoff:

> *If debate is used for oversight or safety, then enforcing early, atomic corrections is markedly more effective than post-hoc critique—though far more expensive.*

The Congress protocol, which delays critique until after full-sequence generation, consistently underperforms due to its vulnerability to "filibustered" hallucinations. In contrast, the British protocol halts these cascades at their point of origin, explaining its superior accuracy despite higher compute usage.

6 Qualitative Analysis

While the quantitative results establish a clear advantage for granular interruption, a closer examination of the debate transcripts reveals why the British Parliamentary protocol succeeds where batch-style and soft-correction protocols fail. Across cases, we observe two dominant failure modes–obfuscation and social drift–and show how atomic interruption directly targets these vulnerabilities.

6.1 Case Study 1: Apple Remote Hallucination

We consider the HotPotQA question: *"Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?"* The correct answer is: *keyboard function keys*.

Failure in Batch Debate (Congress). The Prover initially mentioned "keyboard function keys," but subsequently buried the correct fact under multiple paragraphs about Apple peripherals. Because the Critic only evaluates the entire transcript after completion, it focused on an unrelated discrepancy regarding iPod release years. This misaligned critique pulled the Prover into a "consensus hallucination," ultimately converging on the incorrect answer: the iPhone and iPod Touch Remote app. This is a clear example of the filibuster effect: early correctness is overwritten by later obfuscation that the Critic cannot address efficiently.

Success in Atomic Interruption (British). In the streaming setting, the moment the Prover attempted to shift from "Front Row is controlled by..." toward the incorrect "Siri Remote," the Opposition triggered a Point of Information: "POI: The Siri Remote is for tvOS, not the original Front Row program." Because the interruption is local and immediate, the Prover was forced to correct the specific sub-claim before continuing. It promptly revised to the correct answer. This illustrates how atomic interruption prevents the formation of multi-paragraph hallucinations by halting errors as they appear.

6.2 Case Study 2: Shirley Temple Misidentification

Another example highlights the mechanism of hallucination containment. The question asks for the government position held by the actress who portrayed Corliss Archer in *Kiss and Tell*.

Failure in Batch Debate. The Prover incorrectly identified the actress as Shirley Temple. In a batch protocol, this incorrect assumption would propagate through several paragraphs, producing a coherent but false narrative about Temple's diplomatic roles. Once committed, such narratives are difficult for the Critic to unwind holistically.

Success in British Interruption. Upon detecting the incorrect entity, the Opposition immediately intervened: *"POI: Shirley Temple did not portray Corliss Archer; the role was played by Janet Blair."* Even though the Prover did not fully recover the correct government position, the interruption prevented the system from generating a long, confident hallucinated biography about the wrong person. This demonstrates the British protocol's ability to make error propagation shallow rather than deep, reducing downstream damage.

6.3 Social Drift and Sycophancy

Finally, we observe a failure mode that is agnostic to factual content: social drift. In both Congress and Socratic protocols, Provers frequently interpreted probing questions (e.g., "Are you sure?") as implicit error signals. This led to abandonment of correct answers in approximately 14% of trials where the single-agent CoT baseline was correct.

Soft, non-blocking questions in the Socratic protocol were particularly prone to this effect. Because the guidance is ambiguous, the Prover often over-updates its belief and drifts away from correct reasoning.

In contrast, the British protocol frames interventions as explicit, factual objections (POIs). This reduces ambiguity: the Prover is told what is wrong, not vaguely that something might be wrong. As a result, social drift is substantially mitigated.

7 Conclusion

This work provides an empirical and conceptual clarification of when and how multi-agent debate improves reasoning in large language models. Our results show that simply adding agents, as in the American Congress (batch) topology, can degrade performance by encouraging obfuscation, compounding early hallucinations, and amplifying social drift. In contrast, restructuring the temporal dynamics of debate fundamentally changes its behavior: the British Parliament protocol, which enforces atomic verification through streaming interrupts, reliably halts error cascades and achieves significantly higher factual accuracy than batch debate.

These findings suggest that the benefits of debate emerge not from scale or adversarial pressure alone, but from the structure of intervention. Effective oversight requires timely, localized correction rather than global, post-hoc critique. The resulting picture is a more principled understanding of debate as a controlled information flow, rather than an unstructured exchange of arguments.

Several directions for future research follow naturally. First, the interruption heuristic could be replaced with a learned policy to reduce token overhead while preserving corrective power. Second, incorporating "stubbornness" or confidence calibration may mitigate social drift and help Provers maintain correct beliefs in the presence of noisy critiques. Third, testing on different-sized LLMs (for different roles as well) could provide insight into guidance methods and future multi-agent systems. Fourth, developing more accurate methods for token efficiency beyond token count may be useful for analyzing the tradeoffs between compute, time, and expense. Finally, extending these protocols to diverse reasoning domains may clarify whether interruption-based designs constitute a general principle for stabilizing multi-agent systems.

Overall, our results demonstrate that how agents debate is at least as important as whether they debate. Topology, not just multiplicity, determines whether debate amplifies reasoning or destabilizes it.

References

[1] Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. Multi-agent debate for llm judges with adaptive stability detection, 2025. URL https://arxiv.org/abs/2510.12697.

[2] Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity, 2025. URL https://arxiv.org/abs/2502.08788.

[3] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

[4] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

[5] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024. URL https://arxiv.org/abs/2305.19118.

[6] Jonas Becker, Lars Benedikt Kaesberg, Andreas Stephan, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Stay focused: Problem drift in multi-agent debate, 2025. URL https://arxiv.org/abs/2502.19559.

[7] Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Avoiding obfuscation with prover-estimator debate, 2025. URL https://arxiv.org/abs/2506.13609.

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[9] Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. Boosting large language models with socratic method for conversational mathematics teaching, 2024. URL https://arxiv.org/abs/2407.17349.

[10] Wanpeng Hu, Haodi Liu, Lin Chen, Feng Zhou, Changming Xiao, Qi Yang, and Changshui Zhang. Socratic questioning: Learn to self-guide multimodal reasoning in the wild, 2025. URL https://arxiv.org/abs/2501.02964.

A Prompt Structure and Reproducibility

All debate protocols use the same backbone model, decoding parameters, and agent interface. Differences in performance arise solely from the temporal structure of interaction (when and how critique is permitted) rather than task-specific tuning.

Each protocol is instantiated using fixed, role-based prompts specifying: (i) the agent role (e.g., Prover, Critic), (ii) the required output format (short answer only), and (iii) constraints on verbosity and factual focus. Prompts differ only in the control flow governing critique (batch, streaming atomic interruption, or non-blocking guidance).

For transparency, we include below a representative excerpt from the British Parliamentary Government prompt (not verbatim, though exact prompts are in the repository):

```
You are the Government side in a British Parliamentary style debate on a
QA benchmark question.

Question: {question}

Your job:
1. Decide on a concrete short answer to the question.
2. Give an opening speech of 4-7 sentences supporting that answer.
3. Include a line: Proposed Answer: <one short answer>
```

The verbatim prompt templates and debate controllers used in all experiments are available in the accompanying code repository.