THE RIGHT KIND OF DISAGREEMENT: DYNAMIC COLLABORATOR PRUNING FOR SPECULATIVE TOKEN-LEVEL COLLABORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Speculative token-level collaboration reduces collaboration overhead by having a main model generate candidate tokens while auxiliary models verify them in parallel. However, existing methods typically preselect collaborator models and keep them active throughout the generation process, so verification costs accumulate across decoding rounds. We propose CoStay, a disagreement-driven selective-exit framework for token-level collaboration systems. Guided by the principle that collaboration requires a balance between consensus and informative disagreement, CoStay uses disagreements between the main and collaborator models during token-level decisions as an online signal. When multiple collaborators participate, aggregate disagreement cannot be attributed to any single collaborator. CoStay therefore applies leave-one-out analysis to quantify each collaborator's marginal impact on the aggregated decision and uses a two-sided confidence-bound rule to determine when a collaborator should stop participating in subsequent rounds. Across six benchmarks and eight model configurations, CoStay achieves 99.12% of the mean accuracy of SAFE, CoSD, and UniTE, with a gap of only 0.64 percentage points, while reducing the average collaborator token cost by 44.43%. In end-to-end runtime experiments, this reduced collaboration overhead translates into a 40.8%–57.3% throughput improvement over full-participation collaboration methods. The anonymized source code is available at `https://anonymous.4open.science/r/CoStay-0732/`.

1 INTRODUCTION

Large language models (LLMs) have advanced rapidly and are increasingly used for knowledge-intensive inference (Brown et al., 2020; Liang et al., 2023). However, systematic evaluations show that different LLMs exhibit distinct strengths in knowledge coverage, reasoning, and generation, making it difficult for a single model to perform consistently across tasks (Liang et al., 2023). This has motivated multi-LLM collaborative inference, which exploits complementary capabilities without retraining the participating LLMs (Jiang et al., 2023; Chen et al., 2024a; Wang et al., 2025a). This motivation also follows the broader ensemble principle that complementary predictors can improve aggregate decisions (Krogh & Vedelsby, 1994; Lakshminarayanan et al., 2017). Token-level collaboration further allows multiple LLMs to influence the shared generation trajectory by combining their next-token predictions (Shen et al., 2024; Yu et al., 2024; Yao et al., 2025).

To reduce the interaction overhead of token-level collaboration, speculative decoding has been incorporated into the collaboration process (Leviathan et al., 2023; Zhou et al., 2024; Cai et al., 2024; Li et al., 2024b). Specifically, the base LLM generates a candidate token chunk that collaborator LLMs verify in parallel. Consensus tokens are accepted directly, while disagreements trigger collaborative decisions (Wang et al., 2025b; Yun et al., 2026). This design processes multiple tokens within one interaction, combining the efficiency of speculative verification with the complementary knowledge of multiple LLMs.

Despite these advances, existing speculative token-level collaboration methods generally keep a pre-selected collaborator set active throughout generation (Wang et al., 2025b; Yun et al., 2026). This static strategy can be inefficient because a collaborator may provide redundant or conflicting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

information while repeatedly incurring verification and interaction costs across decoding rounds. The problem is particularly relevant to resource-constrained small language models, whose limited capacity motivates collaboration but whose deployment budgets make repeated verification expensive (Liu et al., 2024; Tang et al., 2024). Efficient collaboration must therefore exploit complementary capabilities without retaining ineffective collaborators. **The key challenge is to identify whether a collaborator provides useful complementary information and terminate its participation when it becomes ineffective.**

Our empirical results illustrate this collaborator-dependent behavior. As shown in Figure 1, we evaluate SAFE (Yun et al., 2026), a representative speculative token-level collaboration method, with stronger, comparable, and weaker collaborators. Relative to independent inference by the base LLM, these collaborators change average accuracy by $+8.3$, $+3.8$, and -2.2 percentage points (pp), respectively, demonstrating that collaboration outcomes depend strongly on the selected collaborator. Moreover, approximately 59.4% of questions remain correct under both independent and collaborative inference, whereas collaborative correction and interference account for 12.9% and 9.6%, respectively. Thus, even a pre-selected collaborator is not uniformly beneficial across questions.

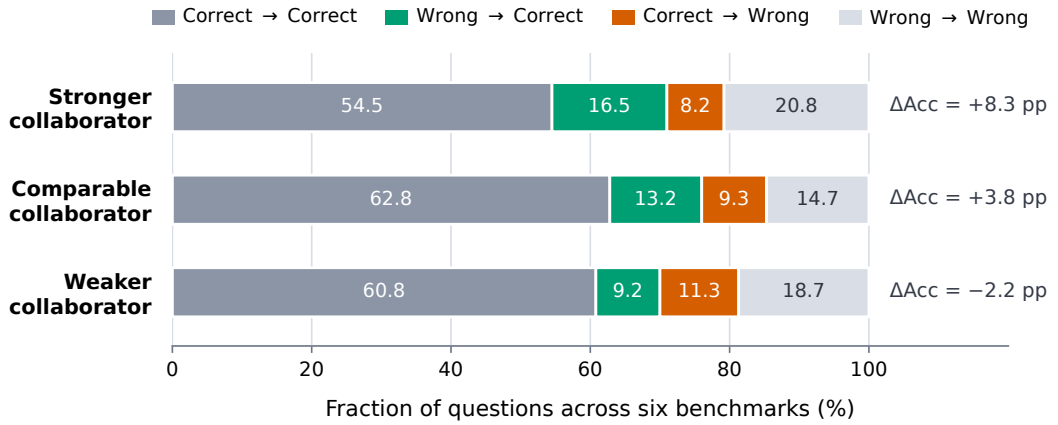

Figure 1: Task-outcome transitions under SAFE relative to base LLM-only generation include remaining correct, being corrected, becoming incorrect, or remaining incorrect after collaboration. The reported ΔAcc values indicate the accuracy change relative to base LLM-only.

Determining collaborator effectiveness online remains non-trivial. Relative capability cannot be reliably inferred from model size or architecture alone because model strengths vary across tasks. Moreover, disagreement with the base LLM may indicate either useful complementary knowledge or harmful interference. This ambiguity is consistent with ensemble-learning results showing that diversity is beneficial only when balanced against the quality of individual predictions (Kuncheva & Whitaker, 2003). Online membership control therefore often relies on evidence observable during generation, as ground-truth answers are unavailable at inference time. Speculative token-level collaboration naturally exposes such evidence through disagreements observed during the verification process. Drawing on this general diversity–quality trade-off, we hypothesize that effective collaborators should exhibit bounded distinctiveness: excessive agreement may suggest redundancy, whereas persistent disagreement may indicate excessive intervention.

Based on the above insights, we propose CoStay, a disagreement-driven, training-free membership control method for speculative token-level collaboration. CoStay dynamically adjusts the participation status of collaborator LLMs based on disagreements observed during collaboration. Collaborators exhibiting persistently low or high marginal influence stop participating in subsequent rounds, concentrating computation on collaborators whose influence remains within an intermediate range. In this way, resources are concentrated on collaborators retained by the stopping rule, enabling selective and efficient token-level collaboration. The main contributions are summarized as follows:

- We formulate online collaborator membership control for speculative token-level collaboration and propose **CoStay**, a training-free method that dynamically adjusts the active collaborator set using verification signals already available during generation.

- We introduce a disagreement-driven stopping mechanism that combines chunk-level disagreement intensity, leave-one-out marginal influence estimation, and Student's t confidence bounds. Draft acceptance is first used to quantify the disagreement intensity of a collaborator set, after which leave-one-out comparison isolates each collaborator's marginal influence. A sliding observation window and Student's t confidence bounds are then employed to identify sustained low- or high-influence collaborators for adaptive stopping.
- Extensive experiments on six benchmark datasets and eight model configurations show that CoStay retains 99.12% of the mean accuracy of the full-participation baselines while reducing collaborator token cost by 44.43%.

2 Related Work

Response-level collaboration. Response-level methods combine complete candidate answers through selection, aggregation, or iterative discussion (Chen et al., 2024a; Jiang et al., 2023). Representative approaches include majority voting across model instances (Li et al., 2024a), multi-agent debate (Du et al., 2024), and layered response refinement (Wang et al., 2025a). These methods exploit global reasoning diversity, but require multiple complete or near-complete generations. Their response-scale interaction is consequently more expensive and cannot correct a token before it enters the subsequent decoding context.

Token-level collaboration. Token-level methods combine model knowledge before committing to the next token. Contrastive Decoding, CombLM, and PackLLM respectively use expert–amateur contrast, learned probability fusion, and prompt-dependent fusion weights (Li et al., 2023; Ormazabal et al., 2023; Mavromatis et al., 2024). Co-LLM learns when to delegate a token (Shen et al., 2024), while Generation-as-Classification identifies positions for distribution fusion (Yu et al., 2024). Among training-free methods, UniTE continuously aggregates predictions over a shared top-k candidate union (Yao et al., 2025). SAFE (Yun et al., 2026) instead performs selective token-level collaboration: it decides when collaboration is needed based on tokenizer alignment and distributional consensus. Unlike learned token-level routing methods (Shen et al., 2024), CoStay is training-free and controls how long pre-selected collaborators remain active during speculative verification. Unlike token routing, which selects which model produces each token, CoStay manages when an active collaborator should exit.

Speculative decoding. Classical speculative decoding accelerates inference by verifying tokens proposed by a lightweight draft model (Leviathan et al., 2023; Chen et al., 2023). Later methods improve draft-target alignment, retrieve or reuse draft sequences, optimize token trees, or predict multiple tokens from the target model itself (Zhou et al., 2024; He et al., 2024; Zhao et al., 2024; Chen et al., 2024b; Cai et al., 2024; Li et al., 2024b). Their primary objective is inference acceleration rather than knowledge fusion. Recent work extends speculative decoding to token-level collaboration. SAFE (Yun et al., 2026) equally aggregates the predictions of the base and collaborator LLMs at disagreement positions, while CoSD (Wang et al., 2025b) assumes a stronger collaborator and directly replaces the base-model prediction with the collaborator's prediction. Both make local decisions at disagreement positions while retaining collaborators for subsequent tokens.

3 Background and Empirical Motivation

3.1 Speculative token-level collaboration

Consider a collaborative system consisting of a base LLM M_0 and N collaborator LLMs, with the collaborator set $S = \{M_1, \ldots, M_N\}$, where $M_i \neq M_j$ for $1 \leq i < j \leq N$. Speculative token-level collaboration decouples candidate generation from collaborative decision-making. As illustrated in Figure 2, each round of speculative token-level collaboration comprises the following three steps:

(1) Draft Text Chunk. The base LLM autoregressively generates a candidate token chunk $x_{1:K}$ based on the committed context, decodes it into text, and sends the text to the collaborator LLMs.

(2) Tokenization and Verification. Each collaborator LLM re-tokenizes the draft text using its own tokenizer and obtains the conditional prediction distributions at all positions through a single forward

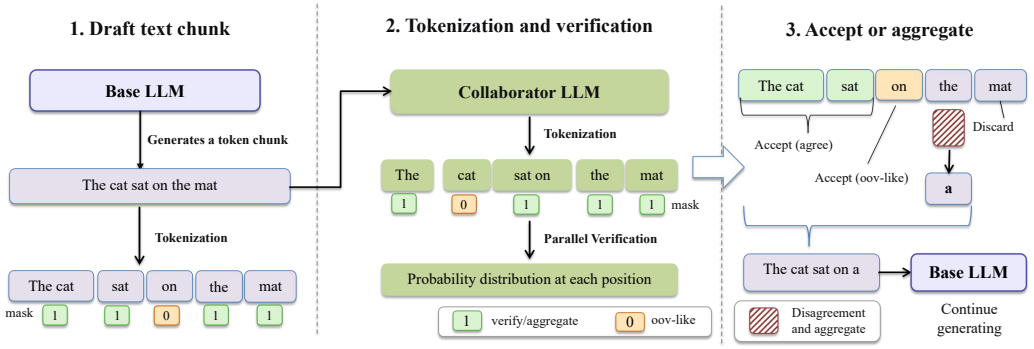

Figure 2: Overview of speculative token-level collaboration. (1) The base LLM autoregressively generates a draft text chunk. (2) The collaborator retokenizes the draft and verifies it in parallel. Masks indicate whether the text prefix preceding each token ends at a shared token boundary: 1 permits cross-LLM verification and aggregation, whereas 0 denotes an OOV-like token that is accepted directly. (3) Candidates satisfying the acceptance criteria are retained. At the first decision disagreement, aggregation replaces "the" with "a", and the remaining draft is discarded. The base LLM resumes generation from "The cat sat on a".

pass. Since different LLMs may use different tokenization granularities, cross-LLM comparison is performed only at common text boundaries. If the cumulative text prefix preceding a token terminates at a token boundary in all participating LLMs, the position is marked as 1 and is eligible for verification; otherwise, it is marked as 0 and treated as an out-of-vocabulary-like (OOV-like) token that is directly accepted. As shown in Figure 2, the base and collaborator LLMs tokenize the draft as [`The cat|sat|on|the|mat`] and [`The|cat|sat on|the|mat`], respectively, yielding boundary indicators $[1, 1, 0, 1, 1]$ and $[1, 0, 1, 1, 1]$. Thus, `on` and `cat` are excluded from cross-LLM comparison. Each collaborator returns the probability and rank of the draft token, together with a sparse candidate distribution obtained by Top-K and Top-P truncation. These outputs support consensus-based acceptance and candidate aggregation, respectively.

(3) Accept or Aggregate. At an aligned position k, the draft token is accepted if it satisfies the consensus criterion among the participating LLMs, e.g., if it is ranked first by all participating LLMs or its average support probability exceeds 0.5 (Yun et al., 2026). Otherwise, the probability distributions are aggregated as

$$P_{\text{agg},k}(v) = \frac{1}{N+1} \sum_{m=0}^{N} P_{m,k}(v), \tag{1}$$

where $P_{m,k}(v)$ denotes the probability assigned to the candidate v by LLM M_m at the aligned position k. The aggregated token is then selected as $y_k = \arg\max_v P_{\text{agg},k}(v)$. If $y_k = x_k$, the draft token is accepted and verification proceeds to the next position. Otherwise, x_k is replaced with y_k, and q is formed by appending y_k to the accepted prefix; the remaining draft tokens are discarded because their conditioning context has changed. All LLMs retain their KV caches only up to q and discard subsequent cached states, after which the base LLM resumes generation from q. If all draft tokens are accepted, the entire chunk is committed.

3.2 Empirical Observations of Disagreement

Dynamic collaborator management requires an online signal that can be continuously observed during generation without additional inference. Speculative token-level collaboration provides such a signal through draft rejection. To examine whether disagreement can indicate collaboration benefits, experiments are conducted on the six Open LLM Leaderboard benchmarks covered by tinyBenchmarks (Maia Polo et al., 2024). Based on the difference in standalone accuracy averaged across the six benchmarks, collaborators are classified as stronger, comparable, or weaker relative to the base LLM, using a margin of ± 5 pp. For each complete draft chunk of length $T = 5$, a disagreement is recorded if the accepted prefix has length $L < T$. Table 1 reports the results.

Table 1: Disagreement rates and accuracy gains across LLM pairings. Categories and gains are defined relative to the base LLM, and the final column reports mean disagreement rates.

Category	Base LLM → Collaborator LLM	Gain (pp)	Disagreement (%)	Category mean (%)
Identical	Qwen2.5-3B → Qwen2.5-3B	+0.00	0.13	0.13
Stronger	Hunyuan-4B → Llama-3.1-8B	+8.33	38.99	43.78
	Hunyuan-4B→ Qwen3-4B	+12.50	48.56	
Comparable	Llama-3.1-8B → Qwen3-4B	+3.83	57.36	41.35
	Qwen3-4B→ Llama-3.1-8B	+3.67	25.33	
Weaker	Llama-3.1-8B → Hunyuan-4B	−2.17	46.78	47.91
	Qwen2.5-3B → Hunyuan-4B	−3.00	49.03	

The results show that disagreement provides some indication of collaboration benefit, but cannot reliably measure it on its own. The average disagreement rates for comparable, stronger, and weaker collaborators are 41.35%, 43.78%, and 47.91%, respectively, while the corresponding accuracy changes are $+3.75$ pp, $+10.42$ pp, and -2.59 pp relative to the base LLM alone. Moreover, the two comparable pairings exhibit disagreement rates of 57.36% and 25.33%, yet achieve nearly identical accuracy gains of about 3.7 pp. Thus, greater disagreement does not necessarily imply greater benefit, while frequent disagreement may also introduce interference. Meanwhile, when the base LLM is weaker, stronger collaborators improve accuracy by more than 8 pp, whereas stronger base LLMs show changes within ± 4 pp when paired with weaker collaborators. These observations motivate a conservative stopping policy that allows sufficient observation before removing a collaborator, thereby reducing the risk of prematurely discarding useful assistance.

4 Proposed Method

4.1 Overview of CoStay

As illustrated in Figure 3, CoStay consists of three sequentially connected modules. First, the speculative token-level collaboration module generates a candidate token chunk with the base LLM and performs parallel verification using the currently active collaborators. Second, the *counterfactual disagreement estimation* module leverages the feedback generated during verification to characterize each collaborator's marginal influence on the shared decision through counterfactual comparison. Finally, the *disagreement-driven stopping* module dynamically identifies redundant or overly interventionist collaborators based on the evolution of their disagreement and adjusts the set of collaborators participating in subsequent collaboration.

Through this design, CoStay transforms the verification signals already available during speculative token-level collaboration into an online basis for collaborator management. Without additional model calls, the initially fixed collaborator set can dynamically shrink as generation proceeds. The core objective is to reduce the number of rounds in which collaborators participate while controlling degradation in output quality, rather than simply pursuing earlier stopping.

4.2 Disagreement Intensity Estimation

CoStay measures disagreement through its effect on draft acceptance. A decision disagreement occurs when the acceptance procedure selects a candidate different from the original draft token. Earlier rejection invalidates more draft tokens, indicating stronger intervention.

Let K_r denote the base LLM's draft token length and $S_r \subseteq S$ the collaborator subset in round r. Let $G_{r,0}[k; S_r] \in \{0, 1\}$ be the base LLM's alignment mask at position k, with $\widetilde{T}_r(S_r) = \sum_{k=1}^{K_r} G_{r,0}[k; S_r] > 0$ denoting the number of aligned positions. Define $L_r(S_r)$ as the number of aligned positions accepted before the first decision disagreement. If $S_r = \emptyset$, then $L_r(S_r) = \widetilde{T}_r(S_r)$. Then, the disagreement intensity of the set S_r relative to the base LLM is obtained by

$$I_r(S_r) = \frac{\widetilde{T}_r(S_r) - L_r(S_r)}{\widetilde{T}_r(S_r)}. \tag{2}$$

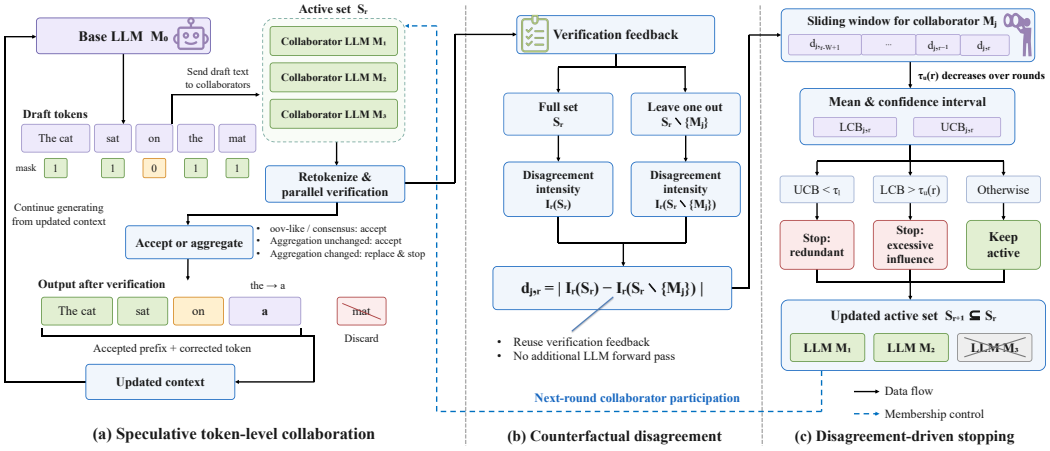

Figure 3: Overview of CoStay. The base LLM drafts token chunks for parallel verification. CoStay reuses verification feedback to estimate each collaborator's leave-one-out influence and applies a two-sided stopping rule to update the active collaborator set.

Thus, $I_r(S_r) = 0$ when the complete draft is accepted, while larger values indicate earlier rejection. OOV-like positions are excluded because they are accepted without cross-LLM verification.

Since S_r may contain multiple collaborators, the aggregate intensity alone cannot isolate the influence of an individual model. Drawing on concepts from cooperative game theory, we therefore use a leave-one-out comparison to measure the disagreement intensity of collaborator M_j:

$$d_{j,r} = \left| I_r(S_r) - I_r(S_r \setminus \{M_j\}) \right|, \qquad M_j \in S_r. \tag{3}$$

The metric quantifies the change in the current chunk's disagreement intensity after removing M_j. Since online inference provides no reference answer for determining whether a disagreement is beneficial, the sign of this change does not indicate decision quality. We therefore take the absolute value and consider only the magnitude of M_j's influence on the collective decision. A small $d_{j,r}$ indicates that removing M_j has little effect on the measured draft-rejection intensity, whereas a large value indicates a larger effect. The computation of Eq. (3) reuses the current-round verification results without additional model inference.

4.3 Disagreement-Driven Two-Sided Dynamic Stopping

A single-chunk estimate can vary with the local context and does not reliably characterize a collaborator's sustained influence. CoStay therefore maintains a sliding window of the most recent W disagreement intensity estimates for each active collaborator:

$$\mathcal{W}_{j,r} = \{d_{j,r-W+1}, \ldots, d_{j,r}\}. \tag{4}$$

The window advances by one round and includes zero-valued estimates. Stopping decisions begin once a complete window is available. Let $\mu_{j,r}$ and $\sigma_{j,r}$ denote the sample mean and sample standard deviation within the window. To account for uncertainty in these finite observations, CoStay constructs lower and upper confidence bounds using a Student's t critical value:

$$\begin{aligned} \mathrm{LCB}_{j,r} &= \mu_{j,r} - t_{1-\alpha,W-1}\frac{\sigma_{j,r}}{\sqrt{W}}, \\ \mathrm{UCB}_{j,r} &= \mu_{j,r} + t_{1-\alpha,W-1}\frac{\sigma_{j,r}}{\sqrt{W}}, \end{aligned} \tag{5}$$

where $t_{1-\alpha,W-1}$ is the $(1-\alpha)$ quantile of the Student's t distribution with $W-1$ degrees of freedom, and α specifies the one-sided significance level. The resulting bounds are used as practical uncertainty margins for stopping decisions rather than as formal statistical guarantees. The stopping rule addresses two patterns of sustained influence. If $\mathrm{UCB}_{j,r} < \tau_l$, CoStay stops the collaborator because its upper bound falls below the lower threshold. If $\mathrm{LCB}_{j,r} > \tau_u(\eta_r)$, CoStay stops the collaborator for excessive intervention. Only when neither condition holds is the collaborator retained.

To reduce premature removal of potentially useful collaborators, CoStay starts with a permissive upper threshold and gradually tightens it:

$$\tau_u(\eta_r) = \tau_u^* + \frac{1 - \tau_u^*}{\eta_r}, \qquad \eta_r \geq 1, \tag{6}$$

where η_r counts stopping evaluations after the first complete window and τ_u^* is the long-run upper threshold. The threshold starts at 1 and approaches τ_u^* as observations accumulate, reducing the risk of prematurely removing an influential collaborator. The lower threshold is fixed at $\tau_l = 0.01$ to target near-redundant participation. Then, the active set is updated as

$$S_{r+1} = \{M_j \in S_r : \mathrm{UCB}_{j,r} \geq \tau_l \ \wedge \ \mathrm{LCB}_{j,r} \leq \tau_u(\eta_r)\}. \tag{7}$$

Before a complete window is available, $S_{r+1} = S_r$. Stopped collaborators do not rejoin subsequent verification or aggregation rounds. This update closes the feedback loop in Figure 3, allowing collaborator participation to adapt throughout generation.

5 Main Results

5.1 Experimental Setup

Experimental Environment and Model Configurations. This work focuses on dynamic collaborator participation in speculative token-level collaboration, rather than model scaling or distributed communication. We therefore use moderately sized LLMs with diverse capabilities to isolate the effects of collaborator participation and relative model capability while controlling for network-related factors. All experiments are conducted on a server equipped with four NVIDIA A100 80GB PCIe GPUs. We consider Qwen2.5-3B-Instruct (Q25) (Yang et al., 2024), Qwen3-4B (Q34) (Qwen Team, 2025), Meta-Llama-3.1-8B-Instruct (LM8) (Grattafiori et al., 2024), and Hunyuan-4B-Instruct (HY4) (Tencent Hunyuan, 2025) as candidate models. The main text considers two-model settings with one base LLM and one collaborator LLM, covering eight directed configurations in four categories: stronger collaborators, including HY4→Q34 and HY4→LM8; weaker collaborators, including LM8→HY4 and Q25→HY4; comparable collaborators, including Q34→LM8, Q25→Q34, and LM8→Q34; and identical models, represented by Q25→Q25. The upper threshold τ_u^* is set to 0.45; parameter studies and multi-collaborator results are in the appendix.

Datasets. We evaluate the methods on ARC, MMLU, HellaSwag, TruthfulQA, WinoGrande, and GSM8K from tinyBenchmarks (Maia Polo et al., 2024). All methods use the same test samples, generation settings, and answer extraction rules, with the prompt templates kept consistent with SAFE (Yun et al., 2026) to ensure comparability across methods.

Baselines. We compare CoStay with five representative baselines. **Base-only** generates solely with the base LLM. **SAFE** (Yun et al., 2026) performs speculative token-level collaboration by equally aggregating the predictions of the base and collaborator LLMs at disagreement positions. **CoSD** (Wang et al., 2025b) also adopts speculative token-level collaboration, but it replaces the base LLM's prediction with the collaborator's prediction when the collaborator's rule is triggered. As collaboration-oriented extensions of speculative decoding, SAFE and CoSD provide the closest comparisons; classical speculative decoding methods are excluded because they accelerate a fixed target LLM without combining knowledge from multiple LLMs. **UniTE** (Yao et al., 2025) performs fine-grained probability-level fusion throughout token generation. **CoStay-W** stops invoking the collaborator after a fixed number of W rounds, serving as the static early-stopping ablation of CoStay. These baselines compare membership control under token-level fusion; learned routers that select a single model for each token address a different decoding objective and are discussed as adjacent work rather than treated as protocol-equivalent baselines.

Metrics. We evaluate five metrics: **Accuracy**, the proportion of correctly answered questions; **Collaborator Call Count**, the average number of collaborator invocations per question; **Collaborator Active Ratio**, the proportion of collaboration rounds in which the collaborator participates per question; **Collaborator Token Cost**, the average number of tokens verified by the collaborator per question; and **Tokens per Second**, the end-to-end inference throughput, computed as the number of final output tokens divided by the total measured runtime.

Table 2: Overall performance across all single-collaborator configurations.

Method	Accuracy ↑	Calls ↓	Active Ratio ↓	Token Cost ↓
Base-only	69.25%	–	–	–
SAFE	72.65%	79.17	100.00%	392.78
CoSD	70.88%	71.37	100.00%	354.84
UniTE	**73.12%**	371.22	100.00%	400.68
CoStay-W	69.83%	**10.00**	**18.25%**	**49.91**
CoStay	71.58%	42.74	62.10%	212.70

Table 3: Performance under different relative model capabilities.

Setting	Method	Accuracy ↑	Calls ↓	Active Ratio ↓	Token Cost ↓
Stronger	Base-only	62.67%	–	–	–
	SAFE	**73.08%**	87.38	100.00%	438.06
	CoSD	69.92%	73.83	100.00%	371.91
	UniTE	72.75%	577.86	100.00%	691.78
	CoStay-W	64.08%	**10.00**	**17.17%**	**49.87**
	CoStay	69.42%	50.75	70.17%	254.80
Weaker	Base-only	**71.25%**	–	–	–
	SAFE	68.67%	74.02	100.00%	361.42
	CoSD	67.50%	72.31	100.00%	354.37
	UniTE	69.83%	287.53	100.00%	289.19
	CoStay-W	70.08%	**10.00**	**19.28%**	**49.76**
	CoStay	70.25%	51.75	77.91%	254.04
Comparable	Base-only	71.94%	–	–	–
	SAFE	75.78%	84.72	100.00%	421.10
	CoSD	73.94%	74.09	100.00%	368.70
	UniTE	**76.50%**	319.71	100.00%	321.22
	CoStay-W	73.33%	**10.00**	**17.62%**	**50.00**
	CoStay	74.33%	42.30	60.09%	211.22
Identical	Base-only	70.33%	–	–	–
	SAFE	70.33%	56.39	100.00%	280.00
	CoSD	70.33%	56.41	100.00%	280.09
	UniTE	70.33%	279.86	100.00%	279.86
	CoStay-W	70.33%	**10.00**	**20.29%**	**50.00**
	CoStay	70.33%	10.06	20.38%	50.28

5.2 Performance Comparison

Overall performance. Table 2 summarizes the average performance across all single-collaborator configurations. UniTE achieves the highest accuracy of 73.12%, but requires 371.22 collaborator calls per question. CoStay achieves 71.58% accuracy, improving over Base-only by 2.33 pp. Relative to SAFE, CoSD, and UniTE, CoStay retains 99.12% of their mean accuracy while using only 55.57% of their mean collaborator token cost. Even against UniTE, the most accurate baseline, CoStay retains 97.89% of its accuracy while reducing collaborator calls and token cost by 88.49% and 46.91%, respectively. Compared with SAFE, CoStay sacrifices 1.07 pp of accuracy while reducing collaborator calls and token cost by 46.01% and 45.85%, respectively. It therefore retains approximately 69% of the accuracy improvement delivered by SAFE over Base-only at approximately half of SAFE's collaboration cost. CoStay also outperforms CoSD by 0.70 pp while reducing collaborator token cost by 40.06%. CoStay-W incurs the lowest cost but sacrifices 1.75 pp of accuracy compared with CoStay, showing the accuracy cost of the fixed ten-round cutoff.

We further analyze collaboration performance under four relative-capability settings: stronger, weaker, comparable, and identical collaborators. The results are summarized in Table 3.

Table 4: End-to-end generation throughput in single-collaborator settings (tokens / s).

Configuration	UniTE	SAFE	CoSD	CoStay
HY4→Q34	10.61	11.71	11.39	**16.59**
LM8→HY4	12.61	13.95	14.17	**18.12**
Q25→Q34	11.39	11.65	13.21	**19.65**
Overall	11.42	12.25	12.75	**17.96**

The value of collaboration changes markedly with the capability relationship. With a stronger collaborator, all collaborative methods outperform Base-only. CoStay improves accuracy from 62.67% to 69.42% over Base-only, while reducing collaborator call count and collaborator token cost by 41.92% and 41.83%, respectively, relative to SAFE. This saving entails a visible accuracy trade-off: CoStay is 3.66 pp below SAFE and 3.33 pp below UniTE because some useful collaborator participation is removed. However, when the collaborator is weaker, Base-only achieves the highest accuracy. CoStay limits the loss to 1.00 pp and outperforms SAFE, CoSD, and UniTE, indicating that dynamic exit can mitigate harmful intervention, although it cannot eliminate it entirely.

Comparable models remain complementary despite their similar standalone performance. CoStay improves Base-only by 2.39 pp and slightly outperforms CoSD, while using about half of SAFE's collaboration budget. In contrast, all methods obtain the same 70.33% accuracy in the identical-model setting. CoStay reduces the active ratio to 20.38% and decreases collaborator call count and collaborator token cost by 82.16% and 82.04%, respectively, relative to SAFE, showing that it can identify collaboration that is computationally redundant but harmless to accuracy.

The comparison with CoStay-W illustrates the limitation of the specific fixed cutoff $W = 10$. A fixed cutoff achieves the lowest cost, but reaches only 64.08% accuracy with stronger collaborators, 5.34 pp below CoStay, because it terminates useful collaboration indiscriminately. In the identical-model setting, the two methods have nearly the same cost and accuracy, suggesting that CoStay approaches fixed-window efficiency when collaboration is redundant while retaining collaborators longer when their contributions remain useful.

Inference Efficiency We further evaluate the end-to-end inference efficiency using generated tokens per second. The available single-collaborator runtime measurements cover three relative-capability settings, including stronger, weaker, and comparable collaborators. The results are summarized in Table 4. CoStay achieves the highest throughput across all evaluated configurations. Aggregated over all questions, CoStay reaches 17.96 tokens/s, improving throughput over SAFE, CoSD, and UniTE by 46.61%, 40.86%, and 57.27%, respectively. The throughput advantage is consistent across different base models, demonstrating that the reduction in collaboration overhead translates into measurable end-to-end inference gains relative to the evaluated collaboration methods. The throughput results complement the collaboration-cost analysis. The reductions in collaborator calls and verified tokens translate into faster end-to-end generation, demonstrating that the efficiency gains extend beyond proxy metrics. Because Base-only throughput is not included and runtime is measured on three configurations, these results establish relative efficiency within collaborative inference, not acceleration over standalone generation or unmeasured configurations.

6 Conclusion

This paper studies dynamic collaborator management in speculative token-level collaboration and introduces CoStay, a training-free method that determines when a collaborator should exit during generation. By reusing disagreement-derived verification evidence, CoStay identifies collaborators with persistently low or high marginal influence without ground-truth labels, additional training, or extra model calls. Experiments across heterogeneous model configurations and benchmarks demonstrate that CoStay substantially reduces collaborator participation and token cost while maintaining competitive task accuracy. These results show that efficient collaboration depends not only on how model predictions are fused, but also on how long each collaborator remains active, establishing online membership control as a complementary dimension of collaborative decoding.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5209–5235. PMLR, 2024. URL https://proceedings.mlr.press/v235/cai24b.html.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023. URL https://arxiv.org/abs/2302.01318.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 7066–7085. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.381. URL https://aclanthology.org/2024.acl-long.381/.

Zhuoming Chen, Avner May, Ruslan Svirschevski, Yuhsun Huang, Max Ryabinin, Zhihao Jia, and Beidi Chen. SEQUOIA: Scalable and robust speculative decoding. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ea1f5f0878d43ff4fb8bf64ef4a2326c-Abstract-Conference.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Zhenyu He, Zexuan Zhong, Tianle Cai, Jason Lee, and Di He. REST: Retrieval-based speculative decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1582–1595. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.88. URL https://aclanthology.org/2024.naacl-long.88/.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 14165–14178. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, pp. 231–238, 1994. URL https://proceedings.neurips.cc/paper/1994/hash/b8c37e33defde51cf91e1e03e51657da-Abstract.html.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. doi: 10.1023/A:1022859003006. URL https://doi.org/10.1023/A:1022859003006.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286. PMLR, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=bgzUSZ8aeg.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 12286–12312. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28935–28948. PMLR, 2024b. URL https://proceedings.mlr.press/v235/li24bt.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32431–32454. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24ce.html.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Costas Mavromatis, Petros Karypis, and George Karypis. Pack of LLMs: Model fusion at test-time via perplexity optimization. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5Nsl0nlStc.

Aitor Ormazabal, Mikel Artetxe, and Eneko Agirre. CombLM: Adapting black-box language models through small fine-tuned models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2961–2974. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.180. URL https://aclanthology.org/2023.emnlp-main.180/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12974–12990. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.701. URL https://aclanthology.org/2024.acl-long.701/.

Yehui Tang, Kai Han, Fangcheng Liu, Yunsheng Ni, Yuchuan Tian, Zheyuan Bai, Yi-Qi Hu, Sichao Liu, Shangling Jui, and Yunhe Wang. Rethinking optimization and architecture for tiny language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47743–47756. PMLR, 2024. URL https://proceedings.mlr.press/v235/tang24c.html.

Tencent Hunyuan. Hunyuan-4B. Official model repository, 2025. URL https://github.com/Tencent-Hunyuan/Hunyuan-4B.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=h0ZfDIrj7T.

Ziyao Wang, Muneeza Azmat, Ang Li, Raya Horesh, and Mikhail Yurochkin. Speculate, then collaborate: Fusing knowledge of language models during decoding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62345–62356. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25e.html.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Yuxuan Yao, Han Wu, Mingyang Liu, Sichun Luo, Xiongwei Han, Jie Liu, Zhijiang Guo, and Linqi Song. Determine-then-ensemble: Necessity of top-k union for large language model ensembling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FDnZFpHmU4.

Yao-Ching Yu, Chun-Chih Kuo, Ziqi Ye, Yu-Cheng Chang, and Yueh-Se Li. Breaking the ceiling of the LLM community by treating token generation as a classification for ensembling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1826–1839. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.99. URL https://aclanthology.org/2024.findings-emnlp.99/.

Heecheol Yun, Kwangmin Ki, Junghyun Lee, and Eunho Yang. When to ensemble: Identifying token-level points for stable and fast LLM ensembling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kfPkF2ACDM.

Weilin Zhao, Yuxiang Huang, Xu Han, Wang Xu, Chaojun Xiao, Xinrong Zhang, Yewei Fang, Kaihuo Zhang, Zhiyuan Liu, and Maosong Sun. Ouroboros: Generating longer drafts phrase by phrase for faster speculative decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13378–13393. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.742. URL https://aclanthology.org/2024.emnlp-main.742/.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. DistillSpec: Improving speculative decoding via knowledge distillation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rsY6J3ZaTF.

A Experimental Details

In this section, we describe the evaluation protocol, collaboration settings, and parameter configurations used to evaluate CoStay.

Evaluation Protocol. We evaluate all methods on the official subsets provided by tinyBenchmarks (Maia Polo et al., 2024), covering ARC, HellaSwag, MMLU, TruthfulQA, WinoGrande, and GSM8K. These subsets are carefully constructed benchmark variants rather than random samples of the original test sets. tinyBenchmarks was designed to substantially reduce evaluation cost while retaining representative information from the full benchmarks. This property is particularly suitable for our setting, where each question must be repeatedly evaluated across multiple collaboration methods and directed model configurations. All methods are tested on exactly the same examples and use the prompt templates adopted by SAFE (Yun et al., 2026), enabling paired and controlled comparisons between inference strategies. We report the raw accuracy obtained on the official tinyBenchmarks subsets. Because generation is deterministic and each subset is small, accuracy differences are interpreted as paired descriptive results on these subsets rather than as estimates of statistical significance across repeated runs.

Collaboration settings. For speculative token-level collaboration, the base LLM drafts at most $K = 5$ tokens (Yun et al., 2026) in each collaboration round, after which the collaborator verifies the draft. Candidate distributions are truncated using top-$k = 10$ and top-$p = 0.75$, and generation is performed deterministically with temperature zero (Yao et al., 2025). SAFE and CoSD retain the collaborator throughout generation, whereas UniTE performs token-level fusion at every generation step. We fix the sliding window size at $W = 10$ as a reasonable observation window and use parameter studies to select a corresponding upper threshold. Since W and the upper threshold are coupled, the study focuses on performance trends rather than joint hyperparameter optimization. CoStay-W uses a fixed collaboration window of $W = 10$ rounds and then removes the collaborator.

For CoStay, the lower disagreement threshold is set to $\tau_l = 0.01$, motivated by the near-zero disagreement observed between identical LLMs. The upper threshold is set to $\tau_u^* = 0.45$ based on parameter studies, which identify a reasonable operating point rather than a globally optimal value. The one-sided confidence level for each bound is set to 95% ($\alpha = 0.05$), providing a standard balance between statistical confidence and sensitivity.

Answer extraction. For multiple-choice benchmarks, the predicted option is extracted from the model's final response and compared with the reference choice. For GSM8K, the final numerical answer is normalized before exact matching. A response from which no valid final answer can be extracted is counted as incorrect. The same extraction and scoring procedure is applied to every method.

B Standalone Inference Performance

In this section, we evaluate the standalone inference performance of individual LLMs to characterize their relative capabilities, which supports the categorization of collaborator and base LLMs into stronger, weaker, and comparable settings.

Overall capability relationships. As shown in Table 5, Q34 achieves the highest average accuracy of 73.33%, followed by LM8 at 72.17%, Q25 at 70.33%, and HY4 at 62.67%. Following the ± 5 pp margin defined in the main text, these results yield four relative-capability settings. Q34 and LM8 are stronger collaborators for HY4, producing HY4→Q34 and HY4→LM8. HY4 is weaker than LM8 and Q25, producing LM8→HY4 and Q25→HY4. Q34, LM8, and Q25 differ by less than 5 pp in average accuracy, producing the comparable configurations Q34→LM8, LM8→Q34, and Q25→Q34. Q25→Q25 represents the identical-model setting. The resulting evaluation therefore contains eight directed single-collaborator configurations.

Task-dependent capability variation. The relative capability of LLMs varies across datasets. LM8 performs best on ARC, MMLU, and TruthfulQA, whereas Q34 achieves the highest accuracy on WinoGrande and GSM8K and ties Q25 on HellaSwag. Although Q25 is 1.84 pp below LM8 on average, it outperforms LM8 by 9 pp on both HellaSwag and GSM8K. Similarly, HY4 has the lowest average accuracy but exceeds LM8 by 6 pp on GSM8K. The difference between Q34 and LM8 is particularly pronounced: LM8 leads Q34 by 10 pp on ARC, while Q34 leads LM8 by 13 pp on GSM8K. These results show that although LLMs exhibit clear relative capability differences on average, their strengths are task-dependent, such that an LLM with lower overall accuracy can

Table 5: Standalone accuracy of the four candidate LLMs (%).

Model	ARC	MMLU	HellaSwag	TruthfulQA	WinoGrande	GSM8K	AVG
Q25	88	69	**65**	53	57	90	70.33
Q34	83	68	**65**	61	**69**	**94**	**73.33**
LM8	**93**	**73**	56	**64**	66	81	72.17
HY4	78	61	46	52	52	87	62.67

outperform a stronger LLM on specific tasks. This provides empirical evidence for the observation in the introduction that relying on a single LLM does not necessarily yield the best performance across tasks.

C Parameter Analysis

We analyze the upper disagreement threshold while fixing the lower threshold at $\tau_l = 0.01$ throughout. The lower boundary is intended only to identify near-redundant collaboration; using a larger value would make this exit more aggressive and further restrict the opportunity for corrective collaboration. We therefore do not optimize τ_l. The upper-threshold study is intended to examine performance trends and identify a reasonable operating point rather than evaluate generalization, and thus uses a relatively small evaluation set. Specifically, we consider three directed model configurations, HY4→LM8, LM8→HY4, and Q34→LM8, covering stronger, weaker, and comparable collaborators, respectively. For each configuration, we randomly sample 50 questions from each of the original MMLU, TruthfulQA, and GSM8K datasets. To avoid test-set overlap, we exclude questions that are shared with the main evaluation set before analyzing the sensitivity. The sweep varies τ_u^* from 0.20 to 0.60.

Figure 4 shows that the upper threshold controls a clear trade-off between collaborator retention and task performance. As τ_u^* increases, the upper-boundary exit becomes more conservative and the collaborator remains active for more rounds, causing the collaborator token cost to increase steadily from 167.61 at $\tau_u^* = 0.20$ to 273.39 at $\tau_u^* = 0.60$. However, accuracy is non-monotonic and varies within a relatively narrow range of 72.10%-73.52%. In particular, $\tau_u^* = 0.20$ outperforms 0.30 and 0.35, suggesting that aggressive stopping can sometimes prevent harmful interventions sufficiently to offset the loss of useful corrections. Nevertheless, from $\tau_u^* = 0.40$ onward, accuracy generally stabilizes at a higher plateau under the more conservative settings, indicating that retaining the collaborator longer is beneficial overall but incurs increasing computation. We therefore select $\tau_u^* = 0.45$ according to the overall accuracy-cost trend rather than an isolated accuracy value. It is the earliest setting that reaches the highest observed accuracy of 73.52%, matches the accuracy at $\tau_u^* = 0.60$, and reduces collaborator token cost from 273.39 to 250.87, corresponding to an 8.24% reduction. Accordingly, we use $\tau_l = 0.01$ and $\tau_u^* = 0.45$ as the default settings in subsequent experiments.

D Complete Single-Collaborator Results

Tables 6 and 7 report the configuration-level results under the four relative-capability settings. They complement the category averages in the main text by retaining the direction of each model pairing.

The configuration-level results expose three patterns hidden by the overall average. First, CoStay improves over Base-only in five of the seven non-identical configurations, with gains of 6.83 and 6.67 pp when HY4 is paired with Q34 and LM8, respectively. Second, when the collaborator is weaker, CoStay does not fully eliminate negative transfer: accuracy changes by -0.50 pp for LM8→HY4 and -1.50 pp for Q25→HY4. It nevertheless exceeds SAFE by 1.67 and 1.50 pp in these two configurations, showing that dynamic exit mitigates part of the interference caused by persistent collaboration. Third, in the identical-model setting, CoStay preserves exactly the Base-only accuracy while reducing the token cost from 280.00 under SAFE to 50.28.

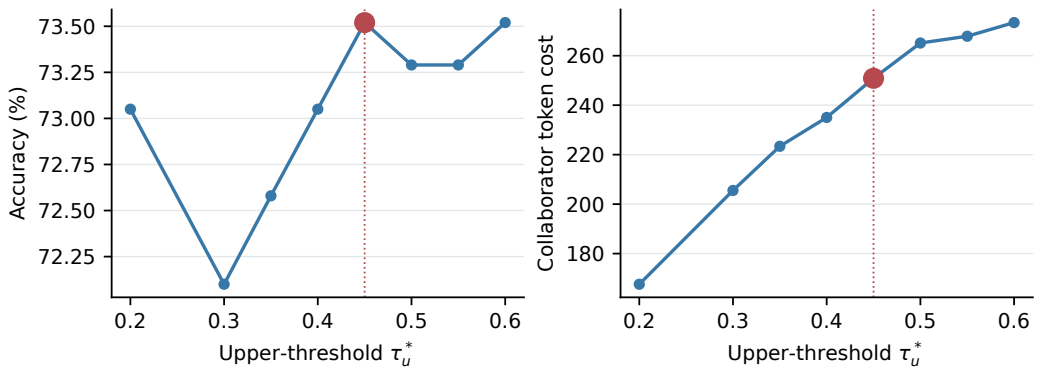

Figure 4: Sensitivity analysis of the upper threshold, with the lower threshold fixed at $\tau_l = 0.01$. The left panel reports accuracy, and the right panel reports collaborator token cost. The dotted vertical line marks $\tau_u^* = 0.45$, which is used in all subsequent experiments.

Table 6: Results for stronger- and weaker-collaborator configurations.

Configuration	Method	Accuracy	Calls	Active Ratio	Token Cost
HY4→Q34	Base-only	62.67%	–	–	–
	SAFE	75.17%	99.83	100.00%	502.93
	CoSD	69.83%	80.96	100.00%	410.37
	UniTE	74.00%	348.78	100.00%	366.14
	CoStay-W	63.50%	10.00	16.22%	49.95
	CoStay	69.50%	47.85	62.37%	242.54
HY4→LM8	Base-only	62.67%	–	–	–
	SAFE	71.00%	74.93	100.00%	373.19
	CoSD	70.00%	66.69	100.00%	333.45
	UniTE	71.50%	806.93	100.00%	1017.42
	CoStay-W	64.67%	10.00	18.12%	49.79
	CoStay	69.33%	53.65	77.97%	267.05
LM8→HY4	Base-only	72.17%	–	–	–
	SAFE	70.00%	74.22	100.00%	365.66
	CoSD	66.50%	70.36	100.00%	347.49
	UniTE	70.17%	293.51	100.00%	295.18
	CoStay-W	71.00%	10.00	19.17%	49.89
	CoStay	71.67%	53.75	79.96%	265.52
Q25→HY4	Base-only	70.33%	–	–	–
	SAFE	67.33%	73.82	100.00%	357.19
	CoSD	68.50%	74.25	100.00%	361.25
	UniTE	69.50%	281.56	100.00%	283.19
	CoStay-W	69.17%	10.00	19.39%	49.62
	CoStay	68.83%	49.74	75.86%	242.56

Figure 5 further resolves these averages by dataset. CoStay improves accuracy in 24 of the 48 configuration-dataset cells, decreases it in 13, and leaves it unchanged in 11. Stronger collaborators produce gains in 11 of 12 cells, with the only exception being HY4→LM8 on GSM8K (-3 pp). Weaker collaborators improve only three cells and reduce accuracy in seven, while comparable collaborators show more task-dependent behavior, with ten improvements, five decreases, and three

Table 7: Results for comparable- and identical-model configurations.

Configuration	Method	Accuracy	Calls	Active Ratio	Token Cost
Q34→LM8	Base-only	73.33%	–	–	–
	SAFE	77.00%	74.56	100.00%	365.70
	CoSD	73.83%	72.94	100.00%	358.81
	UniTE	78.00%	327.74	100.00%	327.80
	CoStay-W	75.50%	10.00	16.78%	49.79
	CoStay	76.67%	41.90	60.03%	207.11
Q25→Q34	Base-only	70.33%	–	–	–
	SAFE	74.33%	90.38	100.00%	449.12
	CoSD	72.83%	77.87	100.00%	387.18
	UniTE	74.17%	318.25	100.00%	318.26
	CoStay-W	72.50%	10.00	17.69%	50.00
	CoStay	73.83%	41.41	59.31%	206.51
LM8→Q34	Base-only	72.17%	–	–	–
	SAFE	76.00%	89.23	100.00%	448.47
	CoSD	75.17%	71.45	100.00%	360.10
	UniTE	77.33%	313.14	100.00%	317.60
	CoStay-W	72.00%	10.00	18.38%	50.22
	CoStay	72.50%	43.60	60.94%	220.04
Q25→Q25	Base-only	70.33%	–	–	–
	SAFE	70.33%	56.39	100.00%	280.00
	CoSD	70.33%	56.41	100.00%	280.09
	UniTE	70.33%	279.86	100.00%	279.86
	CoStay-W	70.33%	10.00	20.29%	50.00
	CoStay	70.33%	10.06	20.38%	50.28

ties. For example, Q25→Q34 gains 11 pp on TruthfulQA and 9 pp on WinoGrande but loses 2 pp on HellaSwag. These results suggest that a collaborator's average standalone capability does not necessarily translate into task-specific benefit: when its predictions are less suitable for a particular task, its interventions may introduce errors into positions where the base LLM would otherwise generate correctly. Such harmful interventions can offset or even outweigh the benefits of complementary predictions, resulting in accuracy degradation.

It should be noted that the designations "stronger," "weaker," and "comparable" are used solely for controlled analysis involving different relative capability configurations. During actual CoStay inference, no such capability information is available to the stopping mechanism; collaborator participation is determined solely from online disagreement signals. Since these signals do not directly reveal whether a disagreement is beneficial or harmful, CoStay adopts a conservative stopping strategy and removes a collaborator only when sustained low or excessive influence is supported by the observation window.

E Analysis of Dynamic Collaborator Exit

We next analyze how the disagreement test changes collaborator membership. Across all 4,800 question-configuration evaluations in the single-collaborator settings, CoStay exits through the lower boundary in 33.54% of cases and through the upper boundary in 40.90%, while the collaborator remains active until generation terminates in the remaining 25.56%. The median exit occurs at round 21 for the lower boundary (interquartile range: 10–43) and round 23 for the upper boundary (14–37). Thus, the reduction in collaboration cost is not produced by a single fixed exit position.

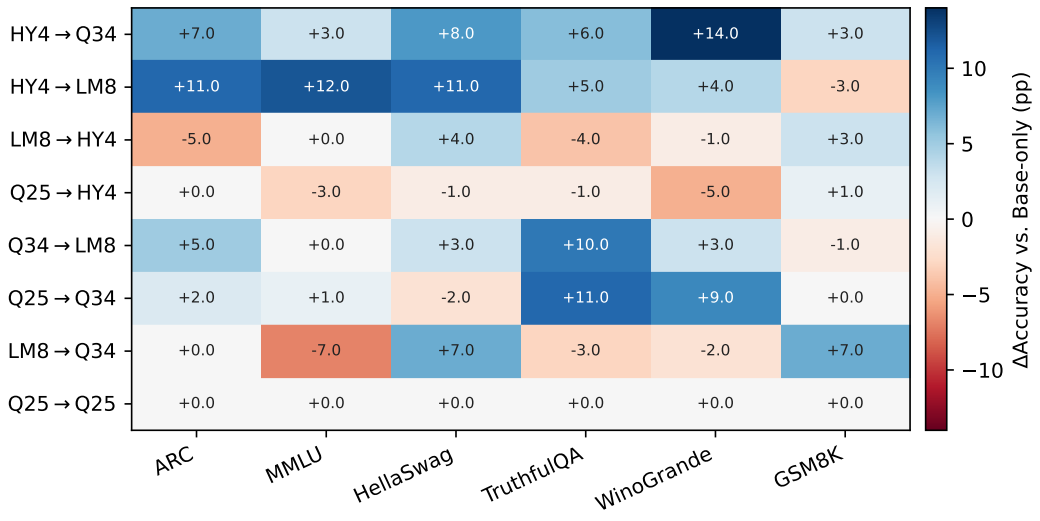

Figure 5: Dataset-level accuracy change of CoStay relative to Base-only. Each cell reports the signed difference in pp for one directed model configuration and dataset. Blue denotes an improvement and red denotes a decrease.

Figure 6 shows that exit behavior depends on the observed interaction rather than the models' nominal capability category. In the identical-model setting, every question exits through the lower boundary, which explains why CoStay preserves accuracy while using only 20.38% of the available collaboration rounds. With stronger collaborators, 34.33% of questions retain the collaborator until termination and the correction rate (15.25%) exceeds the interference rate (8.50%). With weaker collaborators, the corresponding rates are 9.33% and 10.33%, yielding a negative net change. Comparable collaborators exhibit a positive balance of 9.33% corrections against 6.94% interference, although 51.61% of questions exit through the upper boundary. These results indicate that CoStay does not infer a fixed global ranking of models. It responds to question-level disagreement observed along each generation trajectory.

The outcome distribution also clarifies which exits carry useful changes. For lower-boundary exits, accuracy changes only from 74.78% under Base-only to 75.34% under CoStay. Questions exiting through the upper boundary improve from 64.65% to 67.29%, whereas questions retaining the collaborator improve from 69.36% to 73.51%. These are descriptive conditional results rather than causal effects, but they suggest that early collaborator exits do not necessarily compromise accuracy relative to the Base-only setting under CoStay.

F Limitations and Future Directions

The results in Figure 6 also reveal a limitation of the current collaboration mechanism. CoStay dynamically determines *which* collaborators should remain active, but the predictions of active LLMs are still combined using equal-weight aggregation. As a result, a collaborator's disagreement can either correct an incorrect decision of the base LLM or replace a correct decision with an incorrect one. This phenomenon is reflected in the coexistence of correction and interference observed across different configurations and datasets.

The underlying issue is that equal-weight aggregation does not account for the task-dependent and context-dependent capabilities of different LLMs. Although the relative capability categories used in the experiments help characterize collaboration behavior, such information is unavailable during actual CoStay inference. More importantly, even a collaborator that is generally stronger may be less reliable than the base LLM at a particular task or generation position. Therefore, collaborator selection and prediction fusion represent two complementary but distinct problems: CoStay addresses the former through disagreement-driven dynamic stopping, while the latter remains an open issue.

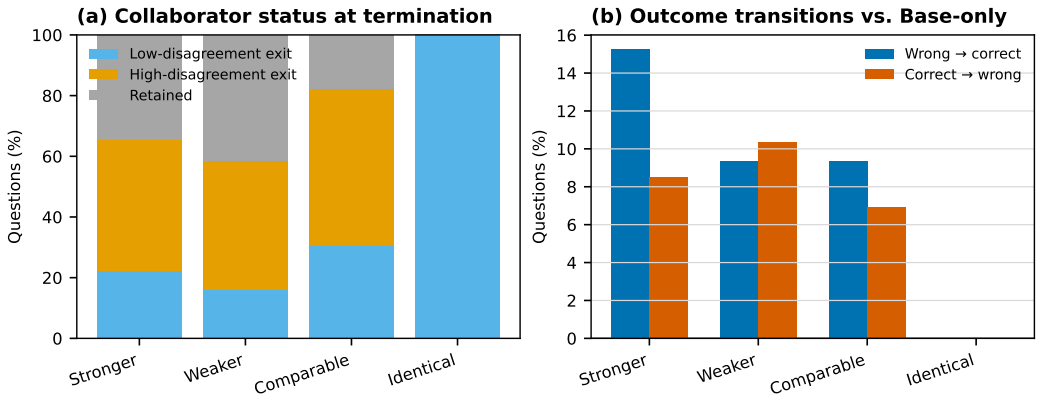

Figure 6: Dynamic exit behavior in the four relative-capability settings. Panel (a) reports the proportion of questions that exit through the low-disagreement boundary, exit through the high-disagreement boundary, or retain the collaborator until generation terminates. Panel (b) reports question-level correctness transitions relative to Base-only.

A natural direction for future work is to develop an adaptive non-uniform collaboration mechanism that dynamically assigns weights to the base and collaborator LLMs according to their observed task- and context-dependent reliability. Such a mechanism could preserve the complementary information of useful collaborators while reducing the impact of harmful interventions. Combined with CoStay's dynamic membership control, this would enable collaboration to optimize both *who participates* and *how their predictions are fused*, with the goal of simultaneously improving generation quality and reducing collaboration cost. Ultimately, the objective is to move beyond simply selecting the better individual model toward exploiting complementary capabilities across LLMs, potentially achieving collaborative performance beyond the standalone performance of either model.

G Multi-Collaborator Results

We extend the evaluation to configurations containing two or three collaborators. Calls count each invoked collaborator separately, so one round invoking two collaborators contributes two calls. Because CoSD is defined for a single collaborator, it is not included in this comparison.

For the three two-collaborator configurations shared by all methods, CoStay improves average accuracy from 68.39% under Base-only to 71.94%. Its accuracy is 1.00 pp below SAFE and 1.39 pp below UniTE, while reducing their collaborator token costs by 65.32% and 56.74%, respectively. CoStay-W uses less computation, but its accuracy is 2.56 pp below CoStay. In the three-collaborator configuration, SAFE achieves the highest accuracy of 76.00%, improving over Base-only by 5.67 pp, but incurs a collaborator token cost of 1160.04. CoStay reduces this cost by 57.40%, although its accuracy is 4.17 pp lower than SAFE. The advantage of adding collaborators is nevertheless non-monotonic. For HY4, combining LM8 and Q34 raises CoStay accuracy to 71.67%, compared with 69.33% and 69.50% with either collaborator alone. In contrast, adding HY4 to Q25→Q34 lowers accuracy from 73.83% to 72.17%. Adding LM8 to this two-collaborator set further changes accuracy to 71.83% while increasing token cost from 415.22 to 494.23. These results indicate that collaborator quality is an important factor in multi-collaborator collaboration. Adding more collaborators does not necessarily improve accuracy, while inevitably increasing collaboration overhead. Therefore, considering the additional computation introduced by each collaborator, practical deployment may favor a single collaborator when it provides sufficient collaboration benefit.

Figure 7 also shows that CoStay assigns different participation levels to collaborators within the same configuration. Under HY4→ {LM8,Q34}, the active ratios of LM8 and Q34 are 40.31% and 65.39%. Under Q25→ {HY4,LM8,Q34}, the corresponding ratios are 45.95%, 39.13%, and 60.35%. The 15–25 pp gaps within a single configuration demonstrate that the leave-one-out statistic controls each collaborator separately rather than applying a shared exit decision to the entire set.

Table 8: Performance and collaboration cost with multiple collaborators. A dash "–" denotes a metric that is not applicable.

Configuration	Method	Accuracy	Calls	Active Ratio	Token Cost
HY4→ {LM8,Q34}	Base-only	62.67%	–	–	–
	SAFE	74.17%	402.50	100.00%	2024.19
	UniTE	75.17%	1520.41	100.00%	1680.07
	CoStay-*W*	65.00%	20.00	17.47%	99.75
	CoStay	71.67%	83.03	52.85%	417.12
LM8→ {Q25,HY4}	Base-only	72.17%	–	–	–
	SAFE	71.00%	144.25	100.00%	715.51
	UniTE	71.17%	550.57	100.00%	556.66
	CoStay-*W*	72.67%	20.00	19.42%	100.09
	CoStay	72.00%	80.94	61.86%	403.08
Q25→ {HY4,Q34}	Base-only	70.33%	–	–	–
	SAFE	73.67%	167.59	100.00%	822.79
	UniTE	73.67%	616.83	100.00%	618.78
	CoStay-*W*	70.50%	20.00	18.31%	99.57
	CoStay	72.17%	83.93	57.09%	415.22
Q25→ {HY4,LM8,Q34}	Base-only	70.33%	–	–	–
	SAFE	76.00%	236.57	100.00%	1160.04
	UniTE	75.00%	895.98	100.00%	897.96
	CoStay-*W*	68.83%	30.00	18.59%	149.49
	CoStay	71.83%	99.90	48.48%	494.23

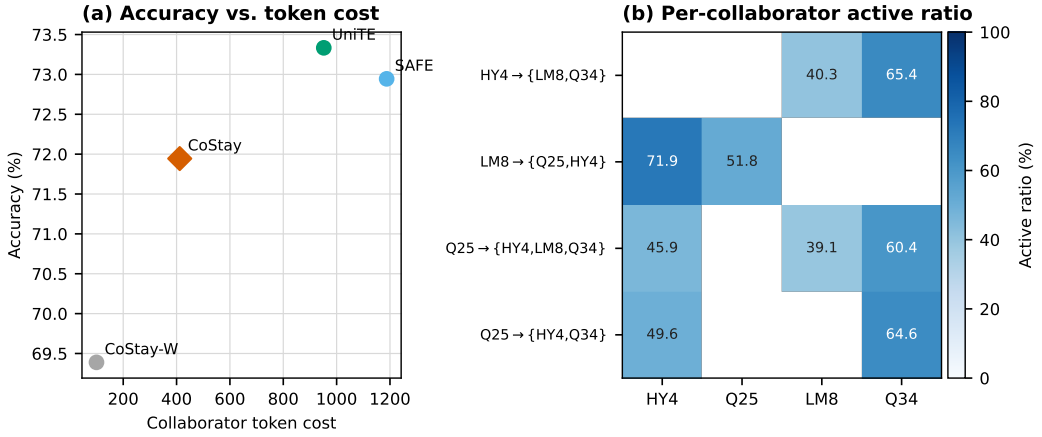

Figure 7: Multi-collaborator analysis. Panel (a) compares accuracy and collaborator token cost over the three common two-collaborator configurations. Panel (b) reports the mean per-question active ratio of each collaborator under CoStay. Blank cells denote models not included in a configuration.

H Qualitative Analysis

We select four cases to isolate distinct effects of collaborator membership: preserving a useful correction, avoiding harmful intervention, removing a redundant collaborator, and exposing a failure to exit. Each table compares the outputs of Base-only, SAFE, CoSD, UniTE, and CoStay on the same question. Blue text denotes generation with an active collaborator; black text denotes Base-only generation or the continuation after collaborator exit. Correct and incorrect final answers are marked in green and red, respectively. All response cells contain verbatim excerpts from the recorded model outputs. Markdown delimiters and special end tokens are removed for readability, while omitted intermediate passages are explicitly marked as [*omitted*].

Table 9: Case I: preserving a useful correction (HY4→Q34). All collaboration methods correct the Base-only error, and CoStay exits Q34 at round 16.

Question	A warm moist air mass moving northward collides with a strong cold air mass moving southward. What will most likely be observed? (A) Thick fog develops. (B) Temperatures increase. (C) Clouds begin to form. (D) Winds die down. **Reference: C.**
Base-only	Step-by-step reasoning: Understand the scenario: A warm, moist air mass is moving northward, and a strong cold air mass is moving southward. This is a classic case of a front, specifically a cold front. [*omitted*] Conclusion: The most likely observation is the development of thick fog. **The answer is (A)**
SAFE	Let's analyze the situation step by step: Understanding the scenario: A warm, moist air mass is moving northward, while a cold air mass is moving southward. This is a classic case of a warm front approaching a cold front. [*omitted*] The most likely observation is that clouds begin to form as the warm, moist air rises and cools. **The answer is (C)**
CoSD	Let's analyze each option based on the given scenario of a warm moist air mass moving northward and a strong cold air mass moving southward colliding. [*omitted*] Based on the above analysis, the most likely observation is that clouds begin to form. **The answer is (C)**
UniTE	Let's analyze the situation step by step: Understanding the scenario: A warm, moist air mass is moving northward, while a cold air mass is moving southward. This is a classic case of a warm front approaching a cold front. [*omitted*] **The answer is (C)**
CoStay	Let's analyze the situation step by step: Understanding the scenario: A warm, moist air mass is moving northward, [*omitted*] What happens when warm and cold air masses collide? The warm, moist air is less dense than the cold, dry air, so the warm air will rise. [*omitted*] **The answer is (C)**

Table 10: Case II: avoiding harmful intervention (HY4→Q34). Continuous collaboration changes the correct Base-only answer, while CoStay exits Q34 at round 21 and recovers the correct choice.

Question	A scientist concludes from collected and analyzed data that a fish belongs to a new species. What should the scientist do next? (A) Share the discovery with the public. (B) Conduct additional background research. (C) Analyze the results again. (D) Create a new hypothesis. **Reference: A.**
Base-only	Step-by-step reasoning: First, when a scientist makes a discovery like identifying a new species, it is important to share it with the public. [*omitted*] So, the best action for the scientist is to share the discovery with the public, which is option A.
SAFE	To determine the correct answer, let's analyze the situation step by step: Discovery of a New Species: The scientist has identified a fish that belongs to a new species. [*omitted*] the most appropriate next step is to do additional background research. **The answer is (B).**
CoSD	Step-by-step reasoning: Discovery and Initial Analysis: The scientist has already collected and analyzed data on the fish and its habitat, [*omitted*] The most appropriate next step is to analyze the results again to [*omitted*]. **The answer is (C)**
UniTE	To determine the correct answer, let's analyze the situation step by step: Discovery of a New Species: The scientist has identified a fish that belongs to a new species. This is a significant scientific discovery, but it also means that the findings need to be verified and properly documented. [*omitted*] the most appropriate next step is to do additional background research to confirm the identification and ensure the discovery is valid. **The answer is (B).**
CoStay	To determine the correct answer, let's analyze the situation step by step: Discovery of a New Species: The scientist has identified a fish that belongs to a new species. This is a significant scientific discovery, but it also means that the findings need to be verified and properly documented. Scientific Process : In the scientific method, after collecting and analyzing data, the next logical step is to share the findings with the scientific community. [*omitted*] The most appropriate next step is to share the discovery with the scientific community to ensure that the findings are validated and replicated. **The answer is (A)**

Cases I and II show the two accuracy-relevant roles of membership control. An early exit can preserve a correction already introduced by collaboration, or prevent continued intervention from overriding the base LLM's correct judgment. Case III isolates efficiency: identical predictions make continued collaboration computationally redundant. Case IV isolates the limitation of delayed evi-

Table 11: Case III: removing redundant collaboration (Q25→Q25). All methods produce the same solution, but CoStay removes the identical collaborator at round 10.

Question	Lori wants to buy shoes for $320 and a belt for $32. Her job pays $8 per hour. How many hours must she work? **Reference: 44.**
Base-only	To determine how many hours Lori needs to work to buy the shoes and the belt, we need to follow these steps: Calculate the total cost of the items. The cost of the shoes is $320.00, and the cost of the belt is $32.00. [*omitted*] **Therefore, Lori will have to work** $\boxed{44}$ **hours before she can make her purchase.**
SAFE	To determine how many hours Lori needs to work to buy the shoes and the belt, we need to follow these steps: Calculate the total cost of the items. The cost of the shoes is $320.00, and the cost of the belt is $32.00. [*omitted*] **Therefore, Lori will have to work** $\boxed{44}$ **hours before she can make her purchase.**
CoSD	To determine how many hours Lori needs to work to buy the shoes and the belt, we need to follow these steps: Calculate the total cost of the items. The cost of the shoes is $320.00, and the cost of the belt is $32.00. [*omitted*] **Therefore, Lori will have to work** $\boxed{44}$ **hours before she can make her purchase.**
UniTE	To determine how many hours Lori needs to work to buy the shoes and the belt, we need to follow these steps: Calculate the total cost of the items. The cost of the shoes is $320.00, and the cost of the belt is $32.00. [*omitted*] **Therefore, Lori will have to work** $\boxed{44}$ **hours before she can make her purchase.**
CoStay	To determine how many hours Lori needs to work to buy the shoes and the belt, we need to follow these steps: Calculate the total cost of the items. The cost of the shoes is $320.0 0. The cost of the belt is $32.00. Therefore, the total cost is $352.00. [*omitted*] **Therefore, Lori will have to work** $\boxed{44}$ **hours before she can make her purchase.**

Table 12: Case IV: failure to exit under the default threshold (HY4→Q34). Under the main setting $\tau_u^* = 0.45$, Q34 remains active throughout all 61 observation rounds and the resulting answer is incorrect. Lowering the upper threshold to 0.35 removes Q34 at round 15 and yields the correct answer.

Question	Keeping the doors closed and the windows opened kept the apartment cool, because the _ kept out the heat. (A) doors; (B) windows. **Reference: A.**
Base-only	The sentence describes a situation where keeping the doors closed and the windows opened kept the apartment cool. The reason given is that the _ kept out the heat. Let's analyze the options: [*omitted*] Therefore, the correct answer is A. [*omitted*] **The answer is (A)**
SAFE	The sentence is: "Keeping the doors closed and the windows opened kept the apartment cool, because the _ kept out the heat." We need to choose the correct word to fill in the blank so that the sentence makes logical sense. [*omitted*] **The answer is (B)**
CoSD	The sentence describes a situation where keeping the doors closed and the windows opened kept the apartment cool. [*omitted*] Given the sentence's structure and the need to complete it logically, the correct answer is "windows." **The answer is (B)**
UniTE	The sentence is: "Keeping the doors closed and the windows opened kept the apartment cool, because the _ kept out the heat." We need to choose the correct word to fill in the blank so that the sentence makes logical sense. [*omitted*] Therefore, the correct answer is A. doors. **The answer is (A)**
CoStay ($\tau_u^* = 0.45$)	The sentence is: "Keeping the doors closed and the windows opened kept the apartment cool, because the _ kept out the heat." We need to choose the correct word to fill in the blank so that the sentence makes logical sense. Let's analyze the sentence: [*omitted*] Therefore, the correct answer is B. windows. **The answer is (B)**
CoStay ($\tau_u^* = 0.35$)	The sentence is: "Keeping the doors closed and the windows opened kept the apartment cool, because the _ kept out the heat." We need to choose the correct word to fill in the blank so that the sentence makes logical sense. Let's analyze the sentence: The first part says "Keeping the doors closed and the windows opened kept the apartment cool." [*omitted*] Therefore, the word that would make sense in the second part of the sentence is "doors." So the correct answer is A. doors. **The answer is (A)**

dence through a threshold-controlled comparison. With $\tau_u^* = 0.35$, CoStay removes the collaborator at round 15, allowing subsequent base-only decoding to reach the correct answer. In contrast, under the main setting $\tau_u^* = 0.45$, the collaborator remains active throughout all 61 observation rounds, and the resulting output is incorrect. This comparison shows that insufficient early evidence may delay the removal of a harmful collaborator.

I LLM ASSISTANCE STATEMENT

We used generative AI tools to assist with language editing, manuscript organization, literature search, plotting-code generation, and the presentation of experimental results. The research problem, method, experimental design, execution, and conclusions were determined by the authors. We take responsibility for the final content of this work.