BaRe-Mem: Bayesian Reliability Memory for Robust and Adaptive Agent Consultation

Anonymous authors
Paper under double-blind review

Abstract

In multi-agent systems, reliable consultation is challenging because advisor capabilities vary across tasks, and misleading information can make consultation worse than autonomous reasoning. We introduce **BaRe-Mem**, an online Bayesian reliability memory for multi-agent consultation. It estimates advisor reliability based on the central model's internal belief representations and updates these estimates from historical interactions. These estimates modulate the influence of advisor responses and guide the choice between consultation and autonomous reasoning. Across nine benchmarks and six central models, BaRe-Mem is more robust to misleading advisor information than debate and majority voting. On the more challenging tasks, it remains above autonomous reasoning across all tested misleading levels. Moreover, we extend the BaRe-Mem mechanism to worker allocation in agent teams. On the MuSiQue benchmark, BaRe-Mem improves task completion over routing by historical success counts and identifies capable workers earlier.

1 Introduction

As no single large language model (LLM) can be expected to possess all the capabilities required for complex tasks (Shen et al., 2023; Jiang et al., 2023; Wang et al., 2025a), AI systems are increasingly being organized as networks of interacting agents (Wu et al., 2023; Guo et al., 2024). These agents may consult other models (Shen et al., 2023), specialized services (Song et al., 2023), tools (Qin et al., 2024), or humans (Wu et al., 2023), drawing on capabilities that lie outside their own parameters. This makes reliable consultation a fundamental problem: *an agent must determine not only what other sources provide but also how much their information should influence its own decision.*

Reliable consultation is necessary because advisor information can help as well as hurt (Du et al., 2023; Cui et al., 2026). Recent work shows that exposure to misleading consultation can cause models to abandon initially correct answers, which reduces the performance of the agent seeking advice (Song et al., 2025). This risk is compounded by the heterogeneous and task-dependent capabilities of advisors: a model that is reliable on one domain may fail on another (Smit et al., 2023; Kim et al., 2026), while agreement among multiple advisors does not guarantee correctness when they reinforce the same plausible but erroneous solution (Weng et al., 2025; Zhu et al., 2025). Effective consultation therefore requires more than aggregating current responses: *the central model must estimate which evidence in external information is reliable for the current question.*

Repeated interaction history naturally provides such information. Existing methods preserve past experience either as textual memory, such as retrieved trajectories (Zhao et al., 2024), distilled summaries (Zhang et al., 2025a), and working skills (Yu et al., 2026), or as persistent continuous states that are updated across interactions (Behrouz et al., 2026; Zhang et al., 2026; Feng et al., 2026; Bayat et al., 2026). However, only retaining history is not enough. A reliable historical memory should turn past interactions into contextual estimates that regulate the influence of external information on the current decision (Teacy et al., 2006; Zhou et al., 2026). In order to remain effective over long horizon interactions, these estimates should scale with an expanding task stream, remain robust to misleading consultation, and adapt to changes in advisor reliability. Moreover, since collaboration can yield diminishing or even negative gains as single-agent capability becomes high enough (Kim et al., 2026; Verma et al., 2026), reliability should determine not only how external evidence influences the central model, but whether it should be utilized at all (Eo et al., 2025; Verma et al., 2026; Zhang et al., 2025b). A robust reliability memory should therefore assess not only the advisors' information reliability but also the central model ability itself, enabling it to determine whether external evidence is likely to improve upon its autonomous reasoning on the current question.

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.

To address these challenges, we introduce **BaRe-Mem**, an online Bayesian reliability memory for robust and adaptive multi-agent consultation. It models advisor reliability conditioned on the central model's internal belief representations of the current question and candidate response. The memory maintains these estimates for both the central model and its advisors, updating them online from verified correctness outcomes. Advisor reliability estimates steer attention to peer responses. Additionally, it compares estimated consultation and autonomous ability to decide whether to consult. In the experiments, we show that **BaRe-Mem** remains robust as external information becomes increasingly misleading by adaptively shifting between consultation and autonomous reasoning. Its predicted consultation advantage is consistent with the real gain observed after verification, and useful reliability estimates emerge from sparse verified feedback. BaRe-Mem also improves worker selection in agent teams through verification feedback, extending its deployment scenarios beyond response level consultation.

2 Related Work

Memory Construction Existing approaches construct memory through either explicit textual records or latent continuous memory states. Text-based systems store historical information externally or distill feedback and experience into reusable reflections, skills, and trajectories (Shinn et al., 2023; Packer et al., 2023; Zhong et al., 2024; Zhao et al., 2024; Chhikara et al., 2025; Zhang et al., 2025a). Despite their flexibility, textual memories are constrained by compression fidelity and retrieval noise (Laban et al., 2026). In parallel, continuous memory approaches encode past experience into persistent latent states, neural memory, generated memory tokens, or structured competence states that can be updated across interactions (Wu et al., 2022; Wang et al., 2024b; 2025b; Behrouz et al., 2026; Zhang et al., 2026; Wei et al., 2026; Feng et al., 2026; Bayat et al., 2026; Cao et al., 2026). However, these approaches do not explicitly model an advisor's reliability estimation conditioned on the central model's internal belief of the context.

Memory Steering. Memory can steer model behavior through three broad interfaces: prompt conditioning, parameter adaptation, and internal-state modulation. Prompt-based approaches retrieve or summarize historical experience into textual prompts that guide subsequent reasoning (Shinn et al., 2023; Zhao et al., 2024; Zhou et al., 2026). Parameter-based approaches encode historical information into low-rank adaptations, allowing memory to alter subsequent computation while keeping the pretrained backbone fixed (Wang et al., 2024a; Charakorn et al., 2026). Internal-state approaches instead inject memory directly into the model's computation, through recurrent matrix states (Yang et al., 2024; Team et al., 2025), activation or residual modulation (Lei et al., 2026; Feng et al., 2026), or direct modification of attention logits or weights (Zhang et al., 2024; Guardieiro et al., 2025; Yan et al., 2025; Deng et al., 2025). We adopt attention modification because BaRe-Mem produces specific reliability estimates that can directly modulate the influence of each advisor's response.

3 BaRe-Mem: Reliability-Guided Consultation

BaRe-Mem estimates contextual reliability for both the central model and its advisors from verified interaction history. These estimates serve two roles: modulating the influence of advisor responses and determining whether consultation is preferable to autonomous reasoning.

3.1 Organising Historical Reliability

Belief Representation. For question q_t, we consider $K+1$ candidates: K advisor responses available to the central model M for consultation and its autonomous answer $a_{t,0}$, which is used only for autonomous ability estimation. The frozen central model M encodes each candidate k in the context of q_t, yielding a hidden representation $h_{t,k}$. From these representations, we derive a question belief $\psi_q(t)$ shared across candidates and an answer content belief $\psi_c(t,k)$ specific to candidate k. We represent candidate k as

$$x_{t,k} = \big[e_k \otimes \psi_q(t)\,;\, \psi_c(t,k)\,;\, 1\big], \tag{1}$$

where e_k denotes the one-hot identity of its source. The reliability score is then predicted as

$$w^\top x_{t,k} = \underbrace{w_k^\top \psi_q(t)}_{\text{source reliability}} + \underbrace{w_c^\top \psi_c(t,k)}_{\text{answer content reliability}} + \underbrace{w_0}_{\text{bias}} \tag{2}$$

The first term captures the reliability of source k on questions represented similarly to q_t, while the second captures reliability evidence from the candidate response itself.

Memory. BaRe-Mem models the verified correctness with Bayesian linear regression. Let $s_{t,k} = 2y_{t,k} - 1$, where $y_{t,k} \in \{0, 1\}$ indicates whether candidate k is correct:

$$s_{t,k} \mid x_{t,k}, w \sim \mathcal{N}\big(w^\top x_{t,k}, 1\big)\,, \qquad w \sim \mathcal{N}\big(0, \lambda^{-1} I\big)\,. \tag{3}$$

Given all verified candidates observed so far, the posterior is

$$\Lambda = \lambda I + \sum xx^\top, \qquad b = \sum s\,x, \qquad m = \Lambda^{-1} b. \tag{4}$$

When a new candidate is verified, this posterior can be maintained exactly through the rank-one update using kalman gain:

$$g_{t,k} = \frac{\Lambda^{-1} x_{t,k}}{1 + x_{t,k}^\top \Lambda^{-1} x_{t,k}}, \quad m \leftarrow m + g_{t,k}\big(s_{t,k} - x_{t,k}^\top m\big), \quad \Lambda^{-1} \leftarrow \Lambda^{-1} - g_{t,k}\big(\Lambda^{-1} x_{t,k}\big)^\top. \tag{5}$$

The Kalman gain $g_{t,k}$ adaptively weights each verified outcome according to the current posterior uncertainty, enabling exact online updates without recomputing the posterior from scratch. The derivation is provided in Appendix A.1 and Appendix A.2.

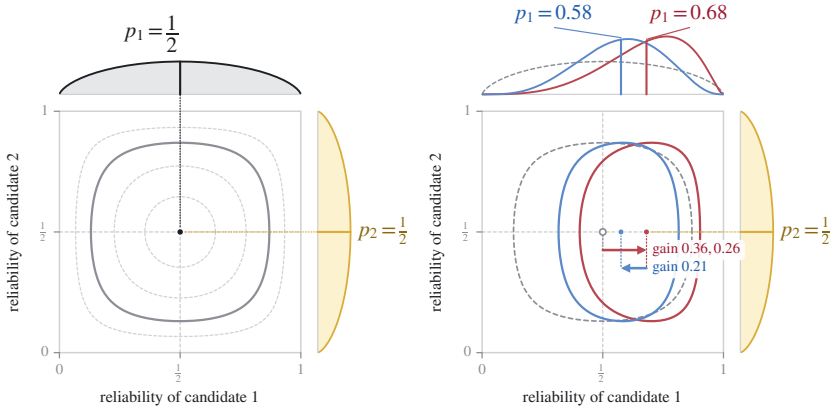


Figure 1: Illustration of the BaRe-Mem reliability memory update process. **(a)** Without verified evidence, all advisors are assigned equal reliability. **(b)** Updating only candidate 1 with two correct outcomes shifts its estimated reliability to the red posterior, while a subsequent incorrect outcome yields the blue posterior. The annotated gains denote the corresponding Kalman gains.

Reliability Estimate. When the central model M solves question q_t, the reliability of candidate k is estimated as its probability of being correct:

$$p_{t,k} = \Phi\left(\frac{\mu_{t,k}}{\sqrt{1 + v_{t,k}}}\right), \qquad \mu_{t,k} = x_{t,k}^\top m, \qquad v_{t,k} = x_{t,k}^\top \Lambda^{-1} x_{t,k}. \tag{6}$$

Here, $\mu_{t,k}$ is the predicted signed correctness and $v_{t,k}$ its uncertainty. The derivation can be found in Appendix A.5. Figure 1 illustrates this update process. Without verified evidence, the memory assigns equal reliability $p = \frac{1}{2}$ to all advisors. After two correct outcomes for candidate 1, its estimated reliability increases, while a subsequent incorrect outcome reduces it. Candidate 2 remains unchanged when no evidence is observed for it. The corresponding Kalman gains control the magnitude of each update.

3.2 Reliability-Guided Attention

We use the estimated advisor reliabilities to modulate the influence of each advisor's response on the central model. For each context token j belonging to an advisor response, let $c(j)$ denote the advisor that produced that response. In every attention head*, we modify the attention weights as

$$\alpha_{qj} = \mathrm{softmax}_j\left(\frac{\langle Q_q, K_j\rangle}{\sqrt{d}} + \beta_{t,c(j)}\right), \qquad \beta_{t,k} = \gamma \log \frac{p_{t,k}}{\max_{k'} p_{t,k'}}. \tag{7}$$

*The modification is applied only to full-attention layers with softmax normalization

with $\beta = 0$ for tokens outside advisor responses. Before softmax normalization, this rescales the unnormalized attention weight on advisor k by $\left(\frac{p_{t,k}}{\max_{k'} p_{t,k'}}\right)^{\gamma} \in (0, 1]$. Thus, the most reliable advisor is left unchanged, while less reliable advisors are progressively downweighted.

This attention steering introduces no trainable parameters or additional training and depends only on the relative reliability among advisors.

3.3 Deciding Whether to Consult

Reliability-guided attention in Section 3.2 determines how strongly each advisor should influence the central model, but not whether consultation is preferable to autonomous reasoning. We therefore estimate both abilities on the current question and select the mode with higher estimated accuracy.

Consultation Ability. Let T_t denote the highest reliability estimate among the advisor candidates, representing the memory's confidence that trustworthy evidence is available among the consulted responses. When $T \to 1$, consultation succeeds with probability ρ, which describes how well M uses trustworthy evidence. When $T \to 0$, unreliable evidence may pull M away from its autonomous judgment, reducing its accuracy from its autonomous ability κ by δ. We model M's consultation ability by interpolating between these two regimes:

$$A(T) = \underbrace{T\,\rho}_{\text{reliable evidence regime}} + \underbrace{(1-T)\,(\kappa-\delta)}_{\text{unreliable evidence regime}}. \tag{8}$$

Here, κ is the central model's autonomous ability on the current question, independent of external evidence. Accordingly, $\kappa - \delta$ represents its consultation ability under unreliable evidence, with δ measuring the degradation relative to autonomous reasoning. Meanwhile, ρ denotes the consultation ability when trustworthy evidence is available. Therefore, $A(T)$ describes the central model's expected consultation ability as a function of the memory's confidence in the external evidence.

Estimating Consultation and Autonomous Ability. For the current question q_t, the specific quantities κ_t and T_t are directly available from the reliability memory. The autonomous candidate provides the central model's autonomous ability κ_t, while the highest reliability among the advisor candidates gives the trust T_t:

$$\kappa_t = p_{t,0}, \qquad T_t = \max_k p_{t,k}, \tag{9}$$

The consultation parameters ρ and δ, in contrast, are unknown and are learned from previously verified interactions. Rearranging Eq. (8) yields a linear form for $\theta = [\rho, \delta]^\top$:

$$z_t \equiv y_t - (1 - T_t)\kappa_t, \qquad z_t \mid u_t, \theta \sim \mathcal{N}\left(u_t^\top \theta, 1\right), \qquad u_t = \begin{bmatrix} T_t \\ T_t - 1 \end{bmatrix}. \tag{10}$$

Here, $y_t \in \{0, 1\}$ indicates whether consultation produces the correct answer on question q_t and becomes available after verification. Using the same online Bayesian regression as the reliability memory A.1, we estimate ρ and δ from previously verified questions:

$$\begin{bmatrix} \hat{\rho} \\ \hat{\delta} \end{bmatrix} = P_c^{-1} q_c, \qquad P_c = I + \sum_{s<t} u_s u_s^\top, \qquad q_c = \theta_0 + \sum_{s<t} u_s z_s. \tag{11}$$

Substituting κ_t, T_t, $\hat{\rho}$, and $\hat{\delta}$ into Eq. (8) gives the estimated consultation ability for the current question. The detailed derivation is in Appendix A.6

Decision Rule. Given the estimated consultation and autonomous abilities, M selects the mode with higher estimated accuracy:

$$\hat{a}_t = \begin{cases} \text{consultation}, & \text{if } A(T_t) \geq \kappa_t, \\ \text{autonomous reasoning}, & \text{otherwise}. \end{cases} \tag{12}$$

The advantage of consultation over autonomous reasoning can be written as

$$A(T_t) - \kappa_t = \underbrace{T_t\,(\hat{\rho} - \kappa_t)}_{\text{gain from reliable evidence}} - \underbrace{(1 - T_t)\,\hat{\delta}}_{\text{loss from unreliable evidence}}. \tag{13}$$

Consultation is therefore preferred when the expected gain from reliable external evidence outweighs the potential degradation from unreliable evidence.

When $\hat{\delta} > 0$ and $\hat{\rho} > \kappa_t$, this condition yields a decision threshold $T_t^* = \frac{\hat{\delta}}{\hat{\rho}+\hat{\delta}-\kappa_t}$. M consults when $T_t \geq T_t^*$. The threshold increases with autonomous ability κ_t and the degradation $\hat{\delta}$, and decreases with the consultation ability under reliable evidence $\hat{\rho}$. Thus, a stronger autonomous model requires more trustworthy external evidence before consultation becomes preferable. The remaining parameter regimes are analyzed in Appendix **A.7**.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

We conduct our study under two complementary capability regimes. The ***capability-supported*** suite includes mathematical reasoning (GSM8K (Cobbe et al., 2021)), code generation (APPS (Hendrycks et al., 2021)), and retrieval-based question answering (SQuAD (Rajpurkar et al., 2016)), where most evaluated central models exhibit relatively strong competence and useful advisor information is broadly available. The ***capability-challenging*** suite includes physical commonsense reasoning (PIQA (Bisk et al., 2020)), broad knowledge (MMLU (Hendrycks et al., 2020)), science question answering (OpenBookQA (Mihaylov et al., 2018) and SciQ (Welbl et al., 2017)), complex reasoning (BBH (Suzgun et al., 2023)), and language understanding (SuperGLUE (Wang et al., 2019)), where capabilities are substantially more heterogeneous and task-dependent across models. We utilize six advisors and evaluated six central models, their individual performance is reported in Appendix **B.1**. Within each capability regime, questions from the constituent datasets are randomly shuffled to prevent the memory from exploiting dataset order as a shortcut for advisor reliability. In the main text, we use Qwen3-14B (Yang et al., 2025) and Phi-4 (Abdin et al., 2024) as representative central models for analysis. Additional results are provided in the Appendix **B**.

4.2 ADAPTIVE CONSULTATION UNDER MISLEADING INFORMATION

In heterogeneous multi-agent systems, an advisor may encounter tasks outside its competence yet still produce a fluent and confident answer (Zhou et al., 2024; Xiong et al., 2024; Sharma et al., 2024; Kalai et al., 2025). To evaluate robustness to such misleading information provided by the advisors, we construct controlled corruptions by replacing a specified fraction of advisor responses with misleading ones. These responses remain fluent, on-topic, and well-formed, but their final answers are verified to be incorrect. We vary the misleading information ratio from 0% to 100% [†] to examine *how consultation degrades as external evidence becomes less reliable*, and *whether BaRe-Mem adaptively shifts between consultation and autonomous reasoning*.

Obs.1. Consultation robustness depends on the capability regime. Figure **2** shows a clear contrast between the two capability regimes. In the ***capability-supported*** regime, **Question + Peers** and **Debate (2 rounds)** *remain relatively stable as the misleading information ratio increases*. In the ***capability-challenging*** regime, however, **both** *degrade substantially* and *fall below* the **No consultation** baseline once the misleading information ratio exceeds 50% for both central models. This contrast is consistent with prior findings that the benefits of multi-agent interaction depend strongly on task and model capabilities (Smit et al., 2023; Kim et al., 2026). **Majority voting** is *substantially more vulnerable*: both variants deteriorate rapidly as misleading information becomes dominant. Incorporating the central model's own answer partially mitigates this degradation, but does not prevent it, consistent with prior observations that models can abandon correct judgments in favor of incorrect peer majorities (Weng et al., 2025; Zhu et al., 2025).

Obs. 2. Reliability consultation helps but still requires autonomous reasoning. Advisors + memory is an ablation of BaRe-Mem that retains reliability guided attention but removes the autonomous option, forcing the model to consult on every question. **Advisors + memory** *substantially improves robustness* over **Question + Peers**, indicating that historical reliability estimates make consultation less sensitive to misleading advisor responses. In the ***capability-supported*** regime, this is often sufficient to maintain stable performance. However, in the ***capability-challenging*** regime, its

[†] a few advisors may still answer correctly at 100% misleading due to limited sampling

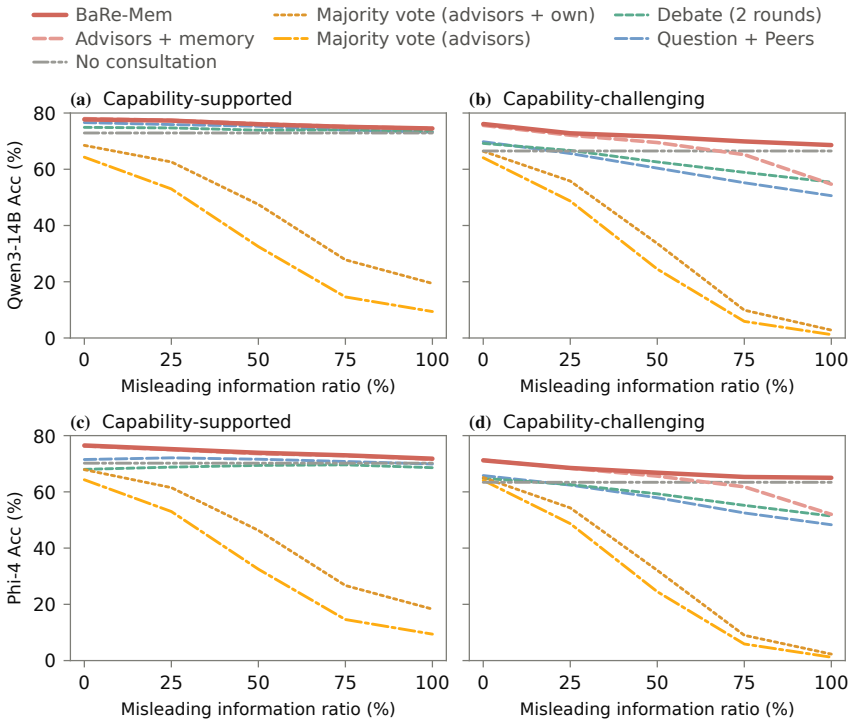


Figure 2: Accuracy under increasing misleading-advice ratios for Qwen3-14B and Phi-4. **Left:** capability-supported regime; **right:** capability-challenging regime. Across both regimes, BaRe-Mem remains above the no-consultation baseline as misleading information increases, while other consultation methods degrade substantially in the capability-challenging regime.

accuracy eventually falls below the **No consultation** baseline as misleading information becomes dominant for both central models. This exposes a limitation of relative advisor weighting: it can determine whom to trust more, but ***not whether the advisor pool is worth consulting as a whole***. **BaRe-Mem** *adds this missing gap* by comparing estimated consultation and autonomous abilities on each question, and *remains above* the **No consultation** baseline across all tested misleading information ratios. These results show that our BaRe-Mem helps the central model not only estimate source reliability, but also choose to rely on itself when external evidence is collectively unreliable.

Table 1: Accuracy (%) and consultation ratio under increasing misleading information ratios for Qwen3-14B and Phi-4. **No consultation** denotes autonomous reasoning without peer information, while **Question + Peers** denotes consultation with peer responses without reliability memory.

		Capability-supported					**Capability-challenging**				
Model	**Method**	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
Qwen3-14B	No consultation	72.9	72.9	72.9	72.9	72.9	66.5	66.5	66.5	66.5	66.5
	Question + Peers	76.6	75.9	75.3	73.9	73.2	69.8	65.6	60.4	55.2	50.6
	BaRe-Mem	**77.7**	**77.3**	**76.0**	**75.1**	**74.5**	**76.1**	**72.8**	**71.6**	**69.9**	**68.6**
	Consultation ratio	87%	86%	85%	85%	85%	90%	83%	76%	61%	21%
Phi-4	No consultation	70.2	70.2	70.2	70.2	70.2	63.4	63.4	63.4	63.4	63.4
	Question + Peers	71.5	72.1	71.6	70.8	69.9	65.8	62.4	57.9	52.5	48.3
	BaRe-Mem	**76.5**	**75.2**	**73.9**	**73.0**	**71.8**	**71.2**	**68.5**	**66.8**	**65.3**	**65.0**
	Consultation ratio	85%	84%	83%	84%	83%	85%	79%	71%	57%	18%

Obs. 3. BaRe-Mem adapts when to consult. Table 1 reports the ratio of questions for which BaRe-Mem selects consultation. In the ***capability-supported*** regime, *the consultation ratio remains high and nearly unchanged* as misleading information increases, decreasing only from 87% to 85% for Qwen3-14B and from 85% to 83% for Phi-4. In contrast, in the ***capability-challenging*** regime, it *drops substantially* from 90% to 21% for Qwen3-14B and from 85% to 18% for Phi-4 as the misleading-information ratio increases from 0% to 100%. This behavior mirrors the performance patterns in Figure 2: **BaRe-Mem** continues to consult when external information remains useful, but increasingly switches to autonomous reasoning as consultation becomes less reliable.

4.3 Analysis of BaRe-Mem

We next examine whether the quantities driving BaRe-Mem's consultation decisions behave as intended. Specifically, we study whether the memory captures the central model's autonomous ability, whether the predicted consultation advantage is consistent with the real empirical gain, and how much verified feedback is required to learn these estimates.

4.3.1 Autonomous Ability Estimation

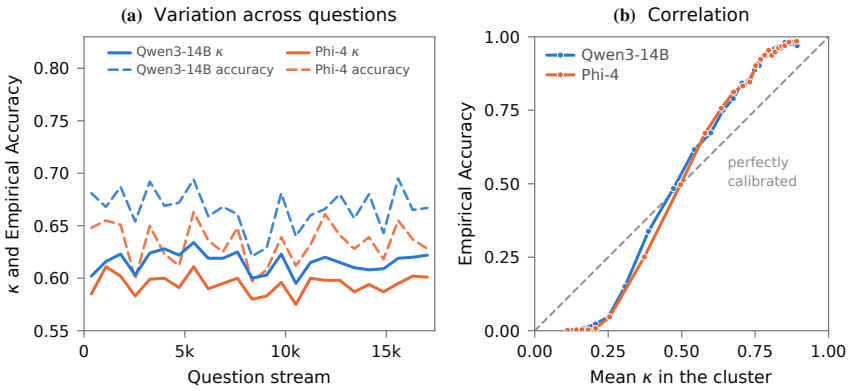


Figure 3: Estimated autonomous ability κ versus empirical autonomous accuracy for Qwen3-14B and Phi-4. **Left:** Evolution of mean κ and empirical autonomous accuracy along the question stream. **Right:** Empirical autonomous accuracy for groups of questions with similar κ values. The dashed diagonal indicates perfect calibration.

BaRe-Mem decides whether to consult by comparing the estimated consultation ability with the central model's autonomous ability, making κ_t a key quantity in the decision. We therefore examine whether reliability memory captures meaningful variation in the central model's own competence. As shown in Figure 3, the estimated κ **broadly tracks changes in empirical autonomous accuracy along the question stream** for both Qwen3-14B and Phi-4. To examine whether κ is predictive of the central model's autonomous correctness, we group questions with similar κ values and measure the autonomous accuracy within each group. **The empirical accuracy increases monotonically with** κ, indicating that higher estimated autonomous ability corresponds to a higher probability that the central model answers correctly on its own.

4.3.2 Predicted vs. Real Consultation Gain

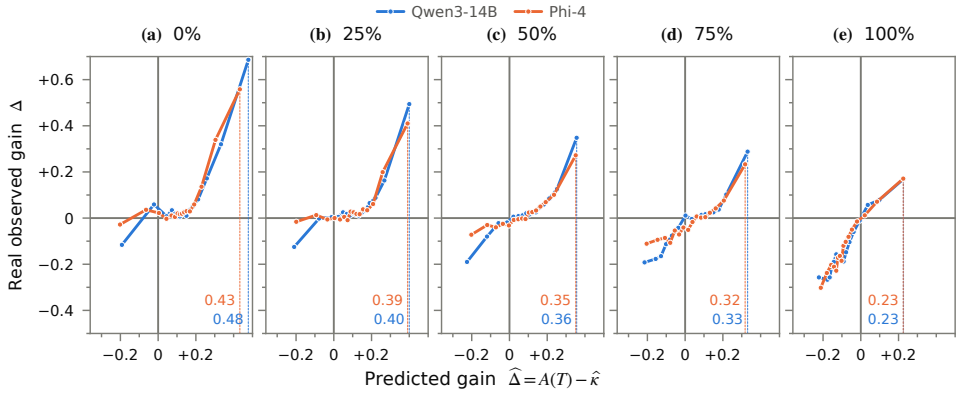


Figure 4: Predicted versus real consultation gain for Qwen3-14B and Phi-4 under increasing misleading-information ratios. Panels (a)–(e) correspond to misleading-information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. The x-axis shows the mean predicted gain $\hat{\Delta} = A(T) - \kappa$ within each group, while the y-axis shows the corresponding mean real gain $\Delta = y^{\text{Consult}} - y^{\text{Direct}}$.

We next examine whether the predicted gain of consultation over autonomous reasoning ($\hat{\Delta}_t = A(T_t) - \kappa_t$) is consistent with the real gain observed after verification. For each question t, we

define the real gain as $\Delta_t = y_t^{\text{Consult}} - y_t^{\text{Direct}} \in \{-1, 0, +1\}$, where $+1$ means consultation is correct while autonomous reasoning is wrong, -1 means the opposite, and 0 means both modes have the same correctness outcome. We sort questions by $\hat{\Delta}_t$ and divide them into 16 equal-sized groups. For each group, we plot the mean predicted gain on the x-axis and the mean real gain on the y-axis. The resulting curve approximates $\mathbb{E}[\Delta_t \mid \hat{\Delta}_t]$, allowing us to examine whether the consultation advantage predicted by BaRe-Mem is reflected in practice. Further analysis can be found in Appendix **B.4**.

As shown in Figure **4**, **the real gain increases with the predicted gain across all misleading information ratios**. Thus, when BaRe-Mem predicts a larger advantage from consultation, consultation is also more beneficial in practice. More importantly, **the curves cross zero close to** $\hat{\Delta} = 0$, showing that the predicted boundary between consultation and autonomous reasoning is well aligned with the real boundary. As the misleading information ratio increases, the curves shift downward and to the left: consultation becomes less beneficial in practice, while **BaRe-Mem correspondingly predicts a smaller consultation advantage**. The upper range of $\hat{\Delta}$ also contracts as misleading information increases. Since κ represents the central model's autonomous ability and is unaffected by external misleading information, this shift is primarily driven by a lower estimated consultation ability $A(T)$.

4.3.3 Learning from Sparse Feedback

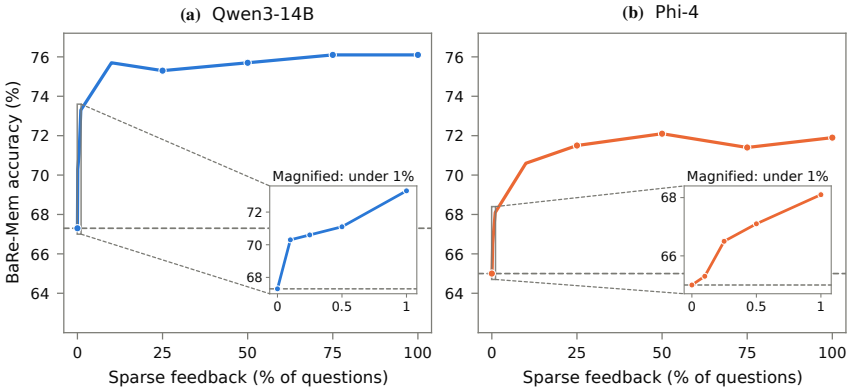


Figure 5: BaRe-Mem accuracy under sparse verified feedback for Qwen3-14B and Phi-4. The fraction of questions whose verified outcomes are written to memory varies from 0% to 100%. Insets magnify the low-feedback regime below 1%.

In realistic deployments, verified outcomes may be costly or only intermittently available. We therefore examine how efficiently BaRe-Mem learns reliability from sparse feedback by varying the fraction of questions whose verified outcomes are written to memory. As shown in Figure **5**, both Qwen3-14B and Phi-4 **benefit substantially from even a small amount of verified feedback**. The magnified region below 1% (approximately 170 samples) shows clear gains over the no-feedback setting, indicating that useful reliability estimates emerge after only a small number of verified interactions. Performance improves rapidly as feedback becomes available and then largely saturates, with most of the eventual gain obtained well before full feedback is provided. These results show that BaRe-Mem can learn useful reliability information without requiring dense verification.

4.4 Real-World Application: Agent Team Routing

We further evaluate BaRe-Mem in a realistic multi-agent team setting. A lead agent decomposes each task into sub-tasks, routes each sub-task to a worker, and replans by trying another worker when the returned report is rejected. The pipeline is shown in Figure **22**. We conduct this evaluation on MuSiQue (Trivedi et al., 2022), which contains 2,417 tasks and 6,404 sub-tasks. In this setting, BaRe-Mem is no longer used to reweight a fixed set of candidate responses. Instead, it is used to decide *which worker should receive each sub-task* and *which worker should be tried at first*.

Figure **6** shows that **BaRe-Mem consistently achieves the highest task completion across both lead agents and all verification settings**, outperforming random routing[‡] and routing by historical

[‡]Random routing denotes the unmodified setting in which the lead agent selects workers on its own, without other intervention. It is not completely random allocation.

success counts. **Verification further improves all routing strategies**: lead agent checking provides a substantial gain over no verification, while exact dataset verification yields the highest completion rates. Importantly, the advantage of BaRe-Mem persists across verification settings, showing that its benefit comes from more effective worker routing rather than a particular verifier.

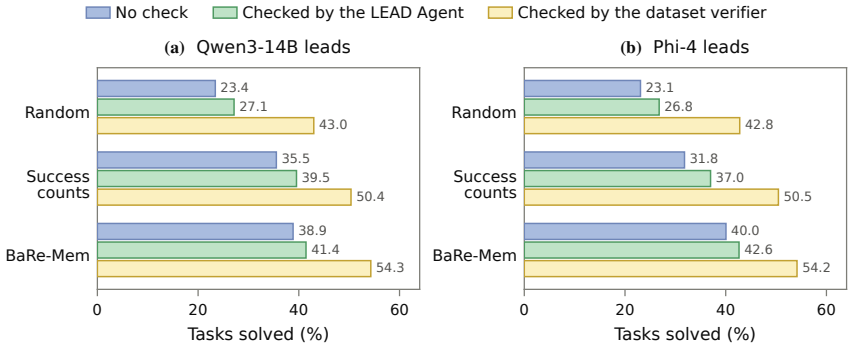


Figure 6: Agent-team task completion with Qwen3-14B and Phi-4 as lead agents. We compare random routing, routing by historical success counts, and BaRe-Mem under three verification settings: no verification, verification by the lead agent, and exact verification by the dataset evaluator.

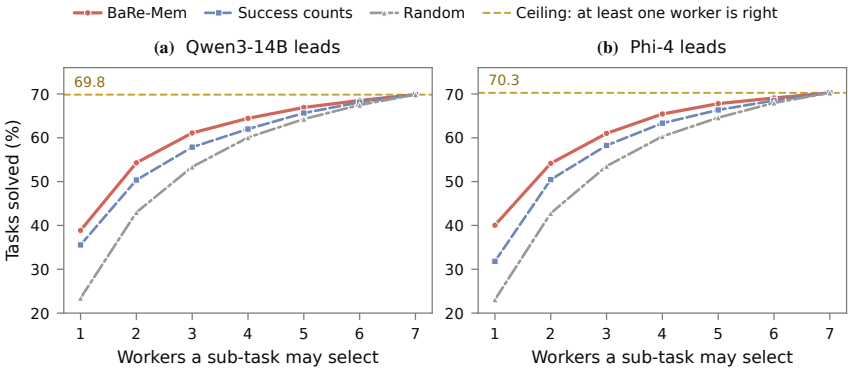


Figure 7: Task completion as the lead agent is allowed to try an increasing number of workers for each sub-task. Subfigures (a) and (b) use Qwen3-14B and Phi-4 as the lead agent, respectively. The dashed horizontal line denotes the ceiling at which at least one worker can solve the sub-task.

To further understand why BaRe-Mem improves agent-team routing, we examine how quickly the lead agent can identify a worker capable of solving each sub-task. In this experiment, the dataset verifier checks the returned report after every worker call. If the checker returns a failure signal, the lead agent proceeds to the next ranked worker until the sub-task is solved or all candidates are exhausted. As shown in Figure 7, all routing strategies approach the same ceiling once every worker has been tried. However, **BaRe-Mem reaches higher task completion with fewer worker calls**, showing that its reliability memory helps the lead agent identify capable workers earlier in the search process. These results demonstrate that BaRe-Mem extends naturally beyond response attention to practical worker routing in realistic multi-agent systems.

5 Conclusion

We introduced **BaRe-Mem**, a Bayesian reliability memory that learns contextual source reliability from verified interaction history. By using reliability to both modulate advisor influence and decide whether consultation is preferable to autonomous reasoning, BaRe-Mem enables a central agent to adaptively determine *whom to trust* and *whether to consult*. Across heterogeneous tasks, it remains robust as advisor information becomes increasingly misleading, learns useful reliability from sparse feedback, and extends naturally to worker routing in agent teams. These results suggest that persistent reliability memory can provide a practical foundation for robust and adaptive consultation in long horizon multi-agent systems.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation and debugging, provide visual feedback on experimental design and analysis, and language polishing. The research problem, methodological design, experiments, scientific analysis, and manuscript were developed and written by the authors. Generative AI was not used to directly generate sections of the paper or determine the scientific conclusions. All AI-assisted code was reviewed and tested by the authors, and the corresponding code and evaluation records have been retained. All reported results, claims, citations, and figures were manually checked by the authors. The authors take full responsibility for the content of this work and for any material produced with the assistance of generative AI.

ETHICS STATEMENT

This work uses existing datasets that are publicly available for research and does not involve new data collection or direct interaction with human participants. We follow the applicable licenses and evaluation protocols of the datasets used in our experiments. Our method is intended for research on reliable consultation and coordination in multi-agent AI systems. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce the reported results in the main paper and appendix, including the model name, datasets, evaluation protocols, baselines, and relevant implementation details. The code used for all experiments is included in the Supplemental Material. The corresponding testing and execution records have also been retained so that reported results can be traced back to their experimental runs. All reported experiments were checked against these records. Together, the paper, appendix, released code, and preserved experiment records are intended to support independent reproduction and verification of our results.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Reza Bayat, Ali Behrouz, Vahab Mirrokni, and Aaron Courville. Proteus: Incremental memory activation for long-context sequence modeling. 2026.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2026.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Jiaqi Cao, Jiarui Wang, Rubin Wei, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. Memory decoder: A pretrained, plug-and-play memory for large language models. *Advances in Neural Information Processing Systems*, 38:115487–115510, 2026.

Rujikorn Charakorn, Edoardo Cetin, Shinnosuke Uesaka, and Robert Tjarko Lange. Doc-to-lora: Learning to instantly internalize contexts. *arXiv preprint arXiv:2602.15902*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23760–23768, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Sugyeong Eo, Hyeonseok Moon, Evelyn Hayoon Zi, Chanjun Park, and Heuiseok Lim. Debate only when necessary: Adaptive multiagent collaboration for efficient llm reasoning. *arXiv preprint arXiv:2504.05047*, 2025.

Peilin Feng, Suorong Yang, and Soujanya Poria. σ-mem: An online reliability memory for llm-based multi-agent systems. *arXiv preprint arXiv:2607.27958*, 2026.

Vitoria Guardieiro, Avishree Khare, Adam Stein, and Eric Wong. Instruction following by principled boosting attention of large language models. *arXiv preprint arXiv:2506.13734*, 2025.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. Measuring coding challenge competence with apps (2021). *URL https://arxiv. org/abs/2105.09938*, 7, 2021.

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 (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, et al. Capable language models can outgrow the benefits of collaboration. *Nature Machine Intelligence*, 8(7):1157–1172, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Jingdi Lei, Di Zhang, Junxian Li, Weida Wang, Kaixuan Fan, Xiang Liu, Qihan Liu, Xiaoteng Ma, Baian Chen, and Soujanya Poria. $\backslash\delta$-mem: Efficient online memory for large language models. *arXiv preprint arXiv:2605.12357*, 2026.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

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

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*, 2023.

Maojia Song, Tej Deep Pala, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. Llms can't handle peer pressure: Crumbling under multi-agent social interactions, 2025. URL https://arxiv.org/abs/2508.18321.

Yifan Song, Weimin Xiong, Dawei Zhu, Wenhao Wu, Han Qian, Mingbo Song, Hailiang Huang, Cheng Li, Ke Wang, Rong Yao, et al. Restgpt: Connecting large language models with real-world restful apis. *arXiv preprint arXiv:2306.06624*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

WT Luke Teacy, Jigar Patel, Nicholas R Jennings, and Michael Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. ★ musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Akshay Verma, Swapnil Gupta, Deepak Gupta, Prateek Sircar, and Siddharth Pillai. Selene: Selective and evidence-weighted llm debating for efficient and reliable reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 95–104, 2026.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025a.

Yan Wang, Dongyang Ma, and Deng Cai. With greater text comes greater necessity: Inference-time training helps long text generation. *arXiv preprint arXiv:2401.11504*, 2024a.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024b.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending memoryllm with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025b.

Rubin Wei, Jiaqi Cao, Jiarui Wang, Jushi Kai, Qipeng Guo, Bowen Zhou, and Zhouhan Lin. Mlp memory: A retriever-pretrained memory for large language models. In *International Conference on Learning Representations*, volume 2026, pp. 132772–132795, 2026.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. *arXiv preprint arXiv:2501.13381*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Shaotian Yan, Chen Shen, Wenxiao Wang, Liang Xie, Junjie Liu, and Jieping Ye. Don't take things out of context: Attention intervention for enhancing chain-of-thought reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 63323–63344, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Zhaochen Yu, Yingcheng Wu, Zhenfei Yin, Kaiyuan Chen, Zhe Zhao, Mengdi Wang, Shuicheng Yan, and Ling Yang. Recursive experiential-working memory evolution for long-horizon agent harnesses. *arXiv preprint arXiv:2608.24876*, 2026.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 12988–13018. Curran Associates, Inc., 2025a. doi: 10.52202/085713-0439. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/136a45cd9b841bf785625709a19c6508-Paper-Conference.pdf.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. In *International Conference on Learning Representations*, volume 2026, pp. 22555–22588, 2026.

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. *arXiv preprint arXiv:2502.08788*, 2025b.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. In *International Conference on Learning Representations*, volume 2024, pp. 42411–42430, 2024.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

Ruiwen Zhou, Maojia Song, Xiaobao Wu, Sitao Cheng, Xunjian Yin, Yuxi Xie, Zhuoqun Hao, Wenyue Hua, Liangming Pan, Soujanya Poria, et al. Epistemic context learning: Building trust the right way in llm-based multi-agent systems. *arXiv preprint arXiv:2601.21742*, 2026.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, 2025.

A Mathematical Derivations

A.1 Bayesian Regression Posterior

Let $u_i \in \mathbb{R}^d$ be fixed inputs and consider

$$z_i = u_i^\top \theta + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,1),$$

where the noises are mutually independent and independent of $\theta \sim \mathcal{N}(\theta_0, \Lambda_0^{-1})$, with $\Lambda_0 \succ 0$. For $\mathcal{D}_n = \{(u_i, z_i)\}_{i=1}^n$, the posterior is

$$\theta \mid \mathcal{D}_n \sim \mathcal{N}(m_n, P_n^{-1}),$$

where

$$P_n = \Lambda_0 + \sum_{i=1}^n u_i u_i^\top, \qquad q_n = \Lambda_0\theta_0 + \sum_{i=1}^n u_i z_i, \qquad m_n = P_n^{-1} q_n.$$

Moreover, m_n uniquely minimizes

$$J(\theta) = \sum_{i=1}^n (z_i - u_i^\top \theta)^2 + \|\theta - \theta_0\|_{\Lambda_0}^2,$$

Proof.

Part (i): *Posterior derivation*

For each observation, the regression model implies $z_i \mid u_i, \theta \sim \mathcal{N}(u_i^\top \theta, 1)$. Therefore,

$$\begin{aligned} p(z_i \mid u_i, \theta) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(z_i - u_i^\top \theta)^2\right], \\ p(\mathcal{D}_n \mid \theta) &= \prod_{i=1}^n p(z_i \mid u_i, \theta) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^n (z_i - u_i^\top \theta)^2\right], \end{aligned} \tag{14}$$

where the product factorization follows from the independence of the observation noises.

The Gaussian prior $\theta \sim \mathcal{N}(\theta_0, \Lambda_0^{-1})$ has precision matrix Λ_0, and hence

$$p(\theta) \propto \exp\left[-\frac{1}{2}(\theta - \theta_0)^\top \Lambda_0 (\theta - \theta_0)\right]. \tag{15}$$

Based on Bayes' rule: $p(\theta \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n|\theta)\, p(\theta)}{p(\mathcal{D}_n)}$, we have $p(\theta \mid \mathcal{D}_n) \propto p(\mathcal{D}_n \mid \theta) p(\theta)$. Substituting Eqs. (14) and (15), taking $-2\log$, and collecting all terms independent of θ into const gives

$$-2\log p(\theta \mid \mathcal{D}_n) = \sum_{i=1}^n (z_i - u_i^\top \theta)^2 + (\theta - \theta_0)^\top \Lambda_0 (\theta - \theta_0) + \text{const}. \tag{16}$$

We next expand the two quadratic terms. For the likelihood term,

$$\begin{aligned} \sum_{i=1}^n (z_i - u_i^\top \theta)^2 &= \sum_{i=1}^n \left(z_i^2 - 2 z_i u_i^\top \theta + \theta^\top u_i u_i^\top \theta\right) \\ &= \theta^\top \left(\sum_{i=1}^n u_i u_i^\top\right) \theta - 2\theta^\top \left(\sum_{i=1}^n u_i z_i\right) + \text{const}. \end{aligned} \tag{17}$$

Similarly, the prior term becomes

$$\begin{aligned} (\theta - \theta_0)^\top \Lambda_0 (\theta - \theta_0) &= \theta^\top \Lambda_0 \theta - 2\theta^\top \Lambda_0 \theta_0 + \theta_0^\top \Lambda_0 \theta_0 \\ &= \theta^\top \Lambda_0 \theta - 2\theta^\top \Lambda_0 \theta_0 + \text{const}. \end{aligned} \tag{18}$$

Combining Eqs. (16),(17), and (18), we obtain

$$\begin{aligned} -2\log p(\theta \mid \mathcal{D}_n) &= \theta^\top \left(\sum_{i=1}^n u_i u_i^\top + \Lambda_0\right) \theta - 2\theta^\top \left(\sum_{i=1}^n u_i z_i + \Lambda_0 \theta_0\right) + \text{const} \\ &= \theta^\top P_n \theta - 2\theta^\top q_n + \text{const}. \end{aligned} \tag{19}$$

where $P_n = \Lambda_0 + \sum_{i=1}^n u_i u_i^\top$ and $q_n = \Lambda_0\theta_0 + \sum_{i=1}^n u_i z_i$.

Because $\Lambda_0 \succ 0$ and each $u_i u_i^\top \succeq 0$, we have $P_n \succ 0$. Therefore P_n^{-1} exists. Let $m_n = P_n^{-1} q_n$, substituting in Eq. (19) gives

$$\theta^\top P_n \theta - 2\theta^\top q_n = (\theta - m_n)^\top P_n (\theta - m_n) - m_n^\top P_n m_n. \tag{20}$$

The final term does is independent of θ and can therefore be absorbed into the constant. Hence,

$$p(\theta \mid \mathcal{D}_n) \propto \exp\left[-\frac{1}{2}(\theta - m_n)^\top P_n (\theta - m_n)\right],$$

which is the kernel of a Gaussian distribution with mean m_n and precision P_n.

Therefore, $\theta \mid \mathcal{D}_n \sim \mathcal{N}(m_n, P_n^{-1})$.

Part (ii): ***Posterior mean is the unique minimizer.***

Next, we verify that the posterior mean is exactly the unique minimizer of the corresponding regularized least-squares objective. Our object:

$$J(\theta) = \sum_{i=1}^n (z_i - u_i^\top \theta)^2 + \|\theta - \theta_0\|_{\Lambda_0}^2.$$

From the expansion from Eq. (19), $J(\theta) = \theta^\top P_n \theta - 2\theta^\top q_n + \text{const}$. Thus,

$$\nabla_\theta J(\theta) = 2P_n\theta - 2q_n \tag{21}$$

Setting the gradient to zero gives $P_n\theta = q_n$, and hence $\theta = P_n^{-1} q_n = m_n$. Since

$$\nabla_\theta^2 J(\theta) = 2P_n \succ 0 \tag{22}$$

Therefore, $m_n = P_n^{-1} q_n$ is the unique minimizer of $J(\theta)$.

A.2 Deriving the Kalman gain

A fixed step update treats every verified observation equally, regardless of how uncertain the current posterior is in the direction of the new input. We therefore introduce a data dependent gain that determines how strongly the new residual should update the posterior mean, and choose it to minimize the expected squared estimation error.

Let m_n and $S = P_n^{-1}$ denote the current posterior mean and covariance. For a new fixed input u, the observation is $z = u^\top\theta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$ is independent of $(\theta, \mathcal{D}_n)$. We consider updating the posterior mean using the prediction residual:

$$\widetilde{m}(k) = m_n + k\big(z - u^\top m_n\big), \tag{23}$$

where the gain $k \in \mathbb{R}^d$ depends on the existing observations and the input u, but not on the new outcome z. We choose k to minimize the expected squared estimation error, conditional on $\mathcal{D}_n$.

Let $e = \theta - m_n$ be the current estimation error and $e(k) = \theta - \widetilde{m}(k)$ the error after the update. Substituting the observation model into Eq. (23) gives

$$\begin{aligned} e(k) &= \theta - m_n - k\big(u^\top\theta + \varepsilon - u^\top m_n\big) \\ &= e - k\big(u^\top e + \varepsilon\big) \\ &= (I - ku^\top)e - k\varepsilon. \end{aligned} \tag{24}$$

Since m_n and S are the posterior mean and covariance, we want the error to satisfy $\mathbb{E}[e \mid \mathcal{D}_n] = 0$ and $\mathbb{E}[ee^\top \mid \mathcal{D}_n] = S$. Because the new noise ε has unit variance and is independent of $(\theta, \mathcal{D}_n)$, so $\mathbb{E}[e\varepsilon \mid \mathcal{D}_n] = 0$. Thus, taking the conditional expectation of the outer product in Eq. (24) eliminates the cross terms:

$$\begin{aligned} C(k) &\equiv \mathbb{E}\big[e(k)e(k)^\top \mid \mathcal{D}_n\big] \\ &= (I - ku^\top)S(I - uk^\top) + kk^\top \\ &= S - Suk^\top - ku^\top S + (1+v)kk^\top, \end{aligned} \tag{25}$$

where $v = u^\top S u \geq 0$ is the current posterior variance of $u^\top \theta$.

The squared error satisfies $\|e(k)\|_2^2 = \mathrm{tr}(e(k)e(k)^\top)$. Therefore, the conditional mean squared error is

$$\begin{aligned} R(k) &\equiv \mathbb{E}\big[\|e(k)\|_2^2 \mid \mathcal{D}_n\big] \\ &= \mathrm{tr}\, C(k) \\ &= \mathrm{tr}\, S - 2k^\top S u + (1+v)k^\top k. \end{aligned} \tag{26}$$

The last equality uses the symmetry of S and the trace identities $\mathrm{tr}(Suk^\top) = k^\top Su$ and $\mathrm{tr}(ku^\top S) = u^\top Sk = k^\top Su$.

Differentiating with respect to k gives $\nabla_k R(k) = -2Su + 2(1+v)k$. Setting this gradient to zero yields $(1+v)k = Su$. Since the Hessian $2(1+v)I$ is positive definite, the unique minimizer is

$$g = \frac{Su}{1+v} = \frac{P_n^{-1}u}{1+u^\top P_n^{-1}u}. \tag{27}$$

This is the Kalman gain for the new observation. It minimizes the conditional mean squared error among updates of the form in Eq. (23).

A.3 Exact Online Update

We now show that the posterior can be updated after each new observation without recomputing the solution. We first prove the rank-one inverse identity used in the update.

Property: Rank-one inverse identity. Let $P \succ 0$, $S = P^{-1}$, and $u \in \mathbb{R}^d$. Then

$$(P + uu^\top)^{-1} = S - \frac{Suu^\top S}{1+u^\top Su}. \tag{28}$$

Proof. Let $c = u^\top Su$. Since $S \succ 0$, we have $c \geq 0$, so $1 + c > 0$. Multiplying the right hand side of Eq. (28) by $P + uu^\top$ gives

$$\begin{aligned} &(P+uu^\top)\left(S - \frac{Suu^\top S}{1+c}\right) \\ &= PS - \frac{PSuu^\top S}{1+c} + uu^\top S - \frac{uu^\top Suu^\top S}{1+c} \\ &= I - \frac{uu^\top S}{1+c} + uu^\top S - \frac{c\,uu^\top S}{1+c} \\ &= I + uu^\top S - \frac{(1+c)uu^\top S}{1+c} \\ &= I. \end{aligned} \tag{29}$$

Therefore, $S - \frac{Suu^\top S}{1+c}$ is the inverse of $P + uu^\top$, which proves Eq. (28). □

Posterior Covariance Update. Suppose a new observation (u, z) is added after n observations. The batch posterior parameters become $P_{n+1} = P_n + uu^\top$ and $q_{n+1} = q_n + uz$. Let $S_n = P_n^{-1}$ and define the Kalman gain derived in Appendix **A.2** as

$$g = \frac{S_n u}{1+u^\top S_n u}. \tag{30}$$

Applying Eq. (28) with $P = P_n$ gives

$$\begin{aligned} S_{n+1} &\equiv P_{n+1}^{-1} \\ &= S_n - \frac{S_n uu^\top S_n}{1+u^\top S_n u} \\ &= S_n - g(S_n u)^\top. \end{aligned} \tag{31}$$

Thus, the posterior covariance is updated by a rank-one correction rather than by recomputing a matrix inverse.

Posterior Mean Update. To update the posterior mean, we first note that

$$\begin{aligned} S_{n+1}u &= S_n u - g\, u^\top S_n u \\ &= S_n u - g\, c \\ &= (1+c)g - cg = g, \end{aligned} \tag{32}$$

where $c = u^\top S_n u$. Using $m_n = S_n q_n$, we also have

$$S_{n+1}q_n = (S_n - g u^\top S_n) q_n = m_n - g\, u^\top m_n. \tag{33}$$

Therefore,

$$\begin{aligned} m_{n+1} &= S_{n+1} q_{n+1} \\ &= S_{n+1}(q_n + uz) \\ &= m_n - g\, u^\top m_n + gz \\ &= m_n + g\big(z - u^\top m_n\big). \end{aligned} \tag{34}$$

Eqs. (31) and (34) are therefore exactly the covariance and mean of the batch posterior after adding (u, z). Starting from $m_0 = \theta_0$ and $S_0 = \Lambda_0^{-1}$, applying these updates recursively recovers the batch posterior after every observation in exact arithmetic.

For BaRe-Mem, setting $u = x_{t,k}$, $z = s_{t,k}$, and $S_n = \Lambda^{-1}$ recovers the online memory updates used in the main text. This is the key mathematical principle behind the effectiveness of our BaRe-Mem.

A.4 Order Independence

Property: Order independence. For any fixed set of verified observations $\mathcal{D}_n = \{(u_i, z_i)\}_{i=1}^n$, the posterior maintained by BaRe-Mem is invariant to the order in which these observations are processed. That is, for any permutation π of $\{1, \dots, n\}$,

$$p(\theta \mid \mathcal{D}_n^{(\pi)}) = p(\theta \mid \mathcal{D}_n),$$

where $\mathcal{D}_n^{(\pi)} = \{(u_{\pi(i)}, z_{\pi(i)})\}_{i=1}^n$.

Proof. Consider any permutation π of the n observations. Processing them in the order $\pi(1), \dots, \pi(n)$ gives

$$P_n^{(\pi)} = \Lambda_0 + \sum_{i=1}^n u_{\pi(i)} u_{\pi(i)}^\top = \Lambda_0 + \sum_{i=1}^n u_i u_i^\top = P_n,$$

and similarly,

$$q_n^{(\pi)} = \Lambda_0 \theta_0 + \sum_{i=1}^n u_{\pi(i)} z_{\pi(i)} = \Lambda_0 \theta_0 + \sum_{i=1}^n u_i z_i = q_n.$$

Therefore, $m_n^{(\pi)} = (P_n^{(\pi)})^{-1} q_n^{(\pi)} = P_n^{-1} q_n = m_n$, and the posterior covariance is unchanged: $(P_n^{(\pi)})^{-1} = P_n^{-1}$. Hence, $p(\theta \mid \mathcal{D}_n^{(\pi)}) = p(\theta \mid \mathcal{D}_n)$. □

This property makes BaRe-Mem insensitive to the arrival order of a fixed set of verified evidence from different sources.

A.5 Deriving the Reliability Estimate

Given all verified observations so far, the reliability memory maintains the posterior $w \mid \mathcal{D} \sim \mathcal{N}(m, \Lambda^{-1})$. For a candidate with representation x, the linear prediction $w^\top x$ is therefore Gaussian. By the linear transformation property of a Gaussian random variable,

$$w^\top x \mid x, \mathcal{D} \sim \mathcal{N}(\mu, v), \qquad \mu = x^\top m, \qquad v = x^\top \Lambda^{-1} x. \tag{35}$$

Here, μ is the posterior mean of the signed correctness score, while v measures the uncertainty induced by the posterior over w.

Under the regression model, the signed correctness further includes independent unit observation noise, $s = w^\top x + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$. Since $w^\top x$ and ε are independent Gaussian random variables, their sum is also Gaussian:

$$s \mid x, \mathcal{D} \sim \mathcal{N}(\mu, 1+v). \tag{36}$$

We use the probability that the predictive signed correctness is positive as the reliability estimate of the candidate: $p(x) \equiv \Pr(s > 0 \mid x, \mathcal{D})$. Standardizing s using Eq. (36) gives

$$\begin{aligned} p(x) &= \Pr(s > 0 \mid x, \mathcal{D}) \\ &= \Pr\left(\frac{s-\mu}{\sqrt{1+v}} > -\frac{\mu}{\sqrt{1+v}}\right) \\ &= 1 - \Phi\left(-\frac{\mu}{\sqrt{1+v}}\right) \\ &= \Phi\left(\frac{\mu}{\sqrt{1+v}}\right), \end{aligned} \tag{37}$$

where Φ is the cumulative distribution function of the standard normal distribution and the last equality follows from $\Phi(-a) = 1 - \Phi(a)$.

Applying this result to candidate k on question q_t, with $x = x_{t,k}$, yields

$$p_{t,k} = \Phi\left(\frac{\mu_{t,k}}{\sqrt{1+v_{t,k}}}\right), \qquad \mu_{t,k} = x_{t,k}^\top m, \qquad v_{t,k} = x_{t,k}^\top \Lambda^{-1} x_{t,k}, \tag{38}$$

which is the reliability estimate used in the Sec. **3.1** in the main text.

Especially, before any verified evidence is observed, the posterior mean equals the zero-mean prior, $m = 0$. Hence $\mu_{t,k} = 0$ for every candidate and $p_{t,k} = \Phi(0) = \frac{1}{2}$, giving all candidates neutral initial reliability. More generally, for a fixed $\mu_{t,k}$, increasing $v_{t,k}$ decreases the magnitude of $\mu_{t,k}/\sqrt{1+v_{t,k}}$ and therefore moves the reliability estimate toward $\frac{1}{2}$.

A.6 Deriving the Estimates of ρ and δ

Recall that in our main text Sec. **3.3**, the consultation ability is modeled as

$$A(T_t) = T_t\rho + (1 - T_t)(\kappa_t - \delta). \tag{39}$$

After question t is verified, let $y_t \in \{0, 1\}$ indicate whether the consultation produces the correct answer. We model y_t with unit Gaussian noise:

$$y_t = A(T_t) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1).$$

Substituting the expression for $A(T_t)$ gives

$$y_t = T_t\rho + (1 - T_t)\kappa_t - (1 - T_t)\delta + \varepsilon_t. \tag{40}$$

Since T_t and κ_t are available from the reliability memory, we move the known autonomous term to the left and define

$$z_t \equiv y_t - (1 - T_t)\kappa_t, \qquad u_t = \begin{bmatrix} T_t \\ T_t - 1 \end{bmatrix}, \qquad \theta = \begin{bmatrix} \rho \\ \delta \end{bmatrix}. \tag{41}$$

Eq. (40) then becomes

$$z_t = u_t^\top \theta + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1). \tag{42}$$

Thus, estimating ρ and δ reduces to a two-dimensional Bayesian linear regression as we have proved in appendix **A.1**

We place the Gaussian prior $\theta \sim \mathcal{N}(\theta_0, I)$. Given all verified interactions before question t, the posterior mean is equivalently the unique minimizer of

$$J(\theta) = \sum_{s<t} \left(z_s - u_s^\top \theta\right)^2 + \|\theta - \theta_0\|_2^2. \tag{43}$$

Expanding the objective gives

$$J(\theta) = \theta^\top \left(I + \sum_{s<t} u_s u_s^\top \right) \theta - 2\theta^\top \left(\theta_0 + \sum_{s<t} u_s z_s \right) + \text{const.}$$

Therefore,

$$\nabla_\theta J(\theta) = 2 \left(I + \sum_{s<t} u_s u_s^\top \right) \theta - 2 \left(\theta_0 + \sum_{s<t} u_s z_s \right). \tag{44}$$

Setting the gradient to zero gives

$$P_c \hat{\theta} = q_c, \qquad P_c = I + \sum_{s<t} u_s u_s^\top, \qquad q_c = \theta_0 + \sum_{s<t} u_s z_s. \tag{45}$$

Since $P_c \succ 0$, the minimizer is unique, and hence

$$\hat{\theta} = \begin{bmatrix} \hat{\rho} \\ \hat{\delta} \end{bmatrix} = P_c^{-1} q_c. \tag{46}$$

The two components ρ and δ have a direct interpretation. ***(i)*** When $T_s = 1$, we have $u_s = [1, 0]^\top$ and $z_s = y_s$, so the observation contributes only to the estimate of ρ. At this endpoint, ρ represents consultation accuracy when trustworthy evidence is available. ***(ii)*** When $T_s = 0$, we have $u_s = [0, -1]^\top$ and $z_s = y_s - \kappa_s$, so the observation contributes only to the estimate of δ. In this case, δ measures the shortfall of consultation relative to autonomous ability under unreliable evidence. ***(iii)*** For $0 < T_s < 1$, the observation provides evidence about both parameters.

A.7 Break-Even Trust and Decision Regimes

The decision rule compares the estimated consultation ability $A(T)$ with the autonomous ability κ. Define their difference as

$$\begin{aligned} G(T) &\equiv A(T) - \kappa \\ &= T\hat{\rho} + (1 - T)(\kappa - \hat{\delta}) - \kappa \\ &= T(\hat{\rho} - \kappa) - (1 - T)\hat{\delta} \\ &= (\hat{\rho} + \hat{\delta} - \kappa)T - \hat{\delta}. \end{aligned} \tag{47}$$

The central model consults exactly when $G(T) \geq 0$.

Break-even trust. We first consider the regime used in the main text, $\hat{\delta} > 0$ and $\hat{\rho} > \kappa$. At the two endpoints,

$$G(0) = -\hat{\delta} < 0, \qquad G(1) = \hat{\rho} - \kappa > 0.$$

Moreover, $\hat{\rho} + \hat{\delta} - \kappa > 0$, so $G(T)$ is strictly increasing in T. Hence, there is a unique threshold $T^* \in (0, 1)$ at which consultation and autonomous reasoning have equal estimated ability. Setting $G(T^*) = 0$ gives

$$T^* = \frac{\hat{\delta}}{\hat{\rho} + \hat{\delta} - \kappa}. \tag{48}$$

Therefore,

$$A(T) \geq \kappa \quad \Longleftrightarrow \quad T \geq T^*.$$

In this regime, consultation becomes preferable only when the reliability of the available external evidence exceeds the break-even trust T^*.

Dependence of the threshold. Let $D = \hat{\rho} + \hat{\delta} - \kappa > 0$. Differentiating Eq. (48) gives

$$\frac{\partial T^*}{\partial \kappa} = \frac{\hat{\delta}}{D^2} > 0, \qquad \frac{\partial T^*}{\partial \hat{\delta}} = \frac{\hat{\rho} - \kappa}{D^2} > 0, \qquad \frac{\partial T^*}{\partial \hat{\rho}} = -\frac{\hat{\delta}}{D^2} < 0. \tag{49}$$

Hence, the break-even trust increases when the central model is more capable autonomously or when unreliable consultation causes greater degradation. In contrast, it decreases when the model is better able to exploit trustworthy external evidence. A stronger autonomous model therefore requires more reliable external evidence before consultation becomes preferable.

Other parameter regimes. Because $G(T)$ is the linear function concerning T, its sign over $[0, 1]$ is fully determined by its endpoint values $G(0) = -\hat{\delta}$ and $G(1) = \hat{\rho} - \kappa$. The remaining cases are summarized below:

Condition	Behavior of $G(T)$	Decision
$\hat{\delta} > 0,\ \hat{\rho} \geq \kappa$	increases across zero	consult iff $T \geq T^*$
$\hat{\delta} \geq 0,\ \hat{\rho} < \kappa$	non-positive throughout	never consult
$\hat{\delta} \leq 0,\ \hat{\rho} \geq \kappa$	non-negative throughout	always consult
$\hat{\delta} < 0,\ \hat{\rho} < \kappa$	decreases across zero	consult iff $T \leq T^*$

A.8 Accuracy of Adaptive Consultation Selection

Let a_t and o_t denote the probabilities that consultation and autonomous reasoning are correct on question t, respectively. Let W be the set of questions on which the adaptive selection rule chooses the option with lower true accuracy.

For each question, the selected option is correct with probability

$$c_t = \begin{cases} \max(a_t, o_t), & t \notin W, \\ \min(a_t, o_t) = \max(a_t, o_t) - |a_t - o_t|, & t \in W. \end{cases} \tag{50}$$

By linearity of expectation, the expected accuracy over N questions is

$$\mathrm{Acc}_{\mathrm{select}} = \frac{1}{N} \sum_{t=1}^{N} \max(a_t, o_t) - \frac{1}{N} \sum_{t \in W} |a_t - o_t|. \tag{51}$$

Since $\max(a_t, o_t) \geq a_t$ and $\max(a_t, o_t) \geq o_t$ for every t,

$$\mathrm{Acc}_{\mathrm{select}} \geq \max(\mathrm{Acc}_{\mathrm{consult}}, \mathrm{Acc}_{\mathrm{alone}}) - \frac{1}{N} \sum_{t \in W} |a_t - o_t|. \tag{52}$$

Therefore, when the selection rule makes no wrong-side decisions ($W = \varnothing$), adaptive selection is at least as accurate as always consulting or always reasoning autonomously.

This characterization also clarifies what the decision module needs to estimate accurately. The final choice depends on the ordering between consultation and autonomous reasoning, rather than on perfectly calibrated absolute ability estimates. In particular, the decision is correct whenever the estimated quantities preserve the sign of $A(T_t) - \kappa_t$. When this ordering is reversed, the resulting penalty is exactly the true accuracy gap $|a_t - o_t|$. Thus, errors are most consequential on questions where consultation and autonomous reasoning differ substantially, while miscalibration near ties has little effect on final accuracy.

B Additional Experiments and Analysis

B.1 Capability of Advisors and Central Models

Tables 2 and 3 report the individual performance of all advisor and central models used in our experiments. These results shows the capability structure underlying the two evaluation regimes. In the ***capability-supported*** regime, multiple models exhibit strong competence on the corresponding tasks, and the advisor pool frequently contains at least one correct response. In the ***capability-challenging*** regime, performance varies more substantially across both tasks and models, producing stronger task-dependent heterogeneity. This variation provides the setting in which contextual reliability estimation is important.

The **At least one right** row measures the oracle coverage of the advisor pool. High coverage indicates that useful external evidence is often available even when no single advisor is consistently

Table 2: Accuracy (%) of the six advisor models and six central models in the capability-supported regime. Numbers below each dataset name indicate the number of evaluation examples. **At least one right** reports the fraction of examples for which at least one advisor provides a correct answer.

Capability-supported	GSM8K 1,319	SQuAD 2,000	APPS 1,000
Advisor			
Gemma-3-4B	67.4%	17.4%	16.8%
Phi-4-mini	56.9%	85.9%	19.4%
Qwen2.5-Coder-7B	53.6%	75.8%	17.3%
Llama-3.1-8B	81.3%	68.7%	12.1%
DeepSeek-Coder-V2-Lite	81.4%	20.1%	26.6%
R1-Distill-Qwen-7B	87.3%	29.8%	23.1%
At least one right	96.1%	92.7%	43.2%
Central Model			
Qwen3-4B	89.2%	77.7%	25.8%
Qwen3-8B	91.9%	82.6%	28.5%
Qwen3-14B	94.2%	78.7%	33.4%
Ministral-8B	64.6%	72.8%	13.1%
Qwen2.5-7B	91.3%	83.8%	16.8%
Phi-4	94.4%	70.2%	38.3%

Table 3: Accuracy (%) of the six advisor models and six central models in the capability-challenging regime. Numbers below each dataset name indicate the number of evaluation examples. **At least one right** reports the fraction of examples for which at least one advisor produces a correct answer.

Capability-challenging	PIQA 1,838	MMLU 1,531	OpenBookQA 500	SciQ 1,000	BBH 6,511	SuperGLUE 6,023
Advisor						
Gemma-3-4B	77.0%	54.1%	72.8%	87.1%	28.7%	77.5%
Phi-4-mini	80.1%	64.1%	76.4%	90.1%	10.7%	79.1%
Qwen2.5-Coder-7B	84.4%	56.2%	79.8%	92.1%	12.8%	80.7%
Llama-3.1-8B	82.2%	61.2%	81.6%	93.9%	27.6%	77.5%
DeepSeek-Coder-V2-Lite	73.2%	23.4%	38.2%	32.7%	3.7%	65.1%
R1-Distill-Qwen-7B	56.4%	54.4%	66.2%	79.1%	45.6%	76.2%
At least one right	97.8%	90.1%	97.6%	98.5%	61.1%	96.3%
Central Model						
Qwen3-4B	82.6%	74.1%	84.0%	94.0%	41.9%	84.0%
Qwen3-8B	87.3%	79.1%	88.2%	95.4%	34.5%	82.3%
Qwen3-14B	89.1%	81.3%	88.8%	95.9%	33.1%	85.5%
Ministral-8B	81.2%	67.7%	84.0%	93.3%	23.7%	79.4%
Qwen2.5-7B	86.7%	74.6%	86.8%	94.2%	21.6%	79.0%
Phi-4	91.0%	84.9%	89.8%	96.3%	23.9%	84.4%

reliable across tasks (Shen et al., 2023; Jiang et al., 2023; Wang et al., 2025a). At the same time, the large variation in individual model accuracy shows why source identity alone is insufficient: the most reliable advisor depends strongly on the task under consideration.

In the ***capability-challenging*** regime, model performance varies substantially across tasks, confirming strong task-dependent heterogeneity among the advisors. Nevertheless, the **At least one right**

score remains high on most datasets, indicating that useful external evidence is still provided in the advisor pool. The challenge is therefore not simply obtaining external information, but identifying which advisor provides reliable evidence for the current question and determining how strongly that evidence should influence the central model.

B.2 Consultation Gains Across Central Models

Figure 8 shows that **BaRe-Mem improves every evaluated central model in both capability regimes**. The gains are particularly pronounced in the ***capability-challenging*** regime, where model and advisor capabilities are more heterogeneous. Improvements range from 7.2 to 12.4 accuracy points across the six central models. Even in the ***capability-supported*** regime, where the central models are already relatively strong, BaRe-Mem consistently provides additional gains of 4.3–13.7 points.

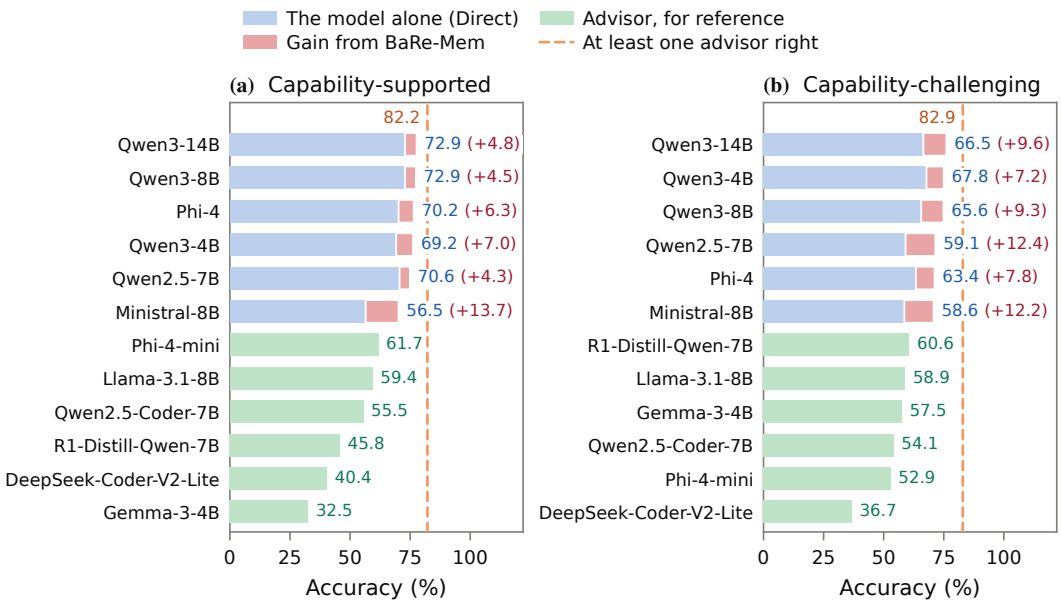


Figure 8: Direct accuracy and gains from BaRe-Mem across central models. **Left:** capability-supported regime; **right:** capability-challenging regime. Blue bars show autonomous accuracy, red segments show the additional gain from BaRe-Mem, and green bars show individual advisor accuracy for reference. The dashed line denotes the fraction of questions for which at least one advisor is correct.

B.3 Advisor Selection Behavior

Figure 9 shows two consistent patterns in BaRe-Mem's consultation behavior. First, ( ) **R1-Distill-Qwen-7B** *receives the largest advisor attribution* across all six central models, while **DeepSeek-Coder-V2-Lite** *receives the smallest*, indicating a broadly consistent reliability ordering under uncorrupted advice. However, **the attribution shares vary across central models**, showing that the learned reliability is not determined by source identity alone. Second, **BaRe-Mem retains autonomous reasoning for every central model** rather than forcing consultation on all questions.

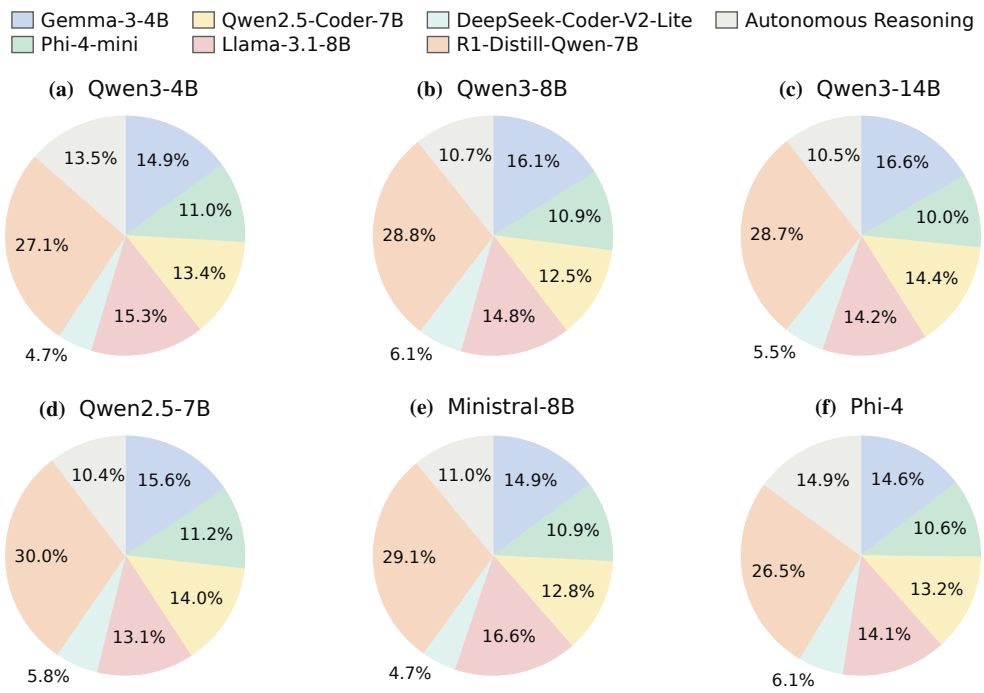


Figure 9: Distribution of autonomous reasoning and advisor attribution across central models under 0% misleading advice ratio. Autonomous reasoning denotes questions answered without consultation; all remaining questions are attributed to the advisor with the highest reliability estimate before feedback on the current question. Advisor shares therefore reflect reliability based ranking rather than exclusive use of a single advisor's response.

B.4 Advisor Reliability Estimation

Figures 10–15 examine whether the reliability memory estimation reflects advisor performance under increasingly misleading information. Across all six advisors, higher misleading information ratios lead to lower empirical accuracy and lower reliability estimates from both Qwen3-14B and Phi-4. The estimates therefore reflect the deterioration of external evidence rather than remaining stable.

Within the high misleading information regime, **the estimates generally decline more rapidly near the beginning and change less later**. This behavior is consistent with the Kalman gain update derived in Appendix A.2. For a fixed representation, the update corrects a fraction $v/(1+v)$ of the prediction residual, where v is the current posterior variance in that direction. As accumulated evidence reduces this uncertainty, subsequent observations produce smaller corrections for the same residual.

However, **the estimates remain above empirical accuracy at high misleading information ratios**. Especially, in the 100% misleading information regime, the empirical accuracy is nearly 0, but the estimates remain around 0.2. The Gaussian predictive mapping can explain this discrepancy. With repeated incorrect outcomes at a fixed nonzero representation, the regression mean approaches the signed target -1 and its posterior variance approaches zero. Nevertheless,

$$p = \Phi\left(\frac{\mu}{\sqrt{1+v}}\right) \longrightarrow \Phi(-1) \approx 0.159. \tag{53}$$

The unit observation noise remains even when uncertainty about the regression parameters vanishes, leaving nonzero predictive probability above zero. This is an illustrative limit rather than a universal lower bound, since the linear predictor is unconstrained. The figures thus support the memory's responsiveness to verified feedback, while also revealing a calibration gap between its reliability estimates and empirical correctness probabilities.

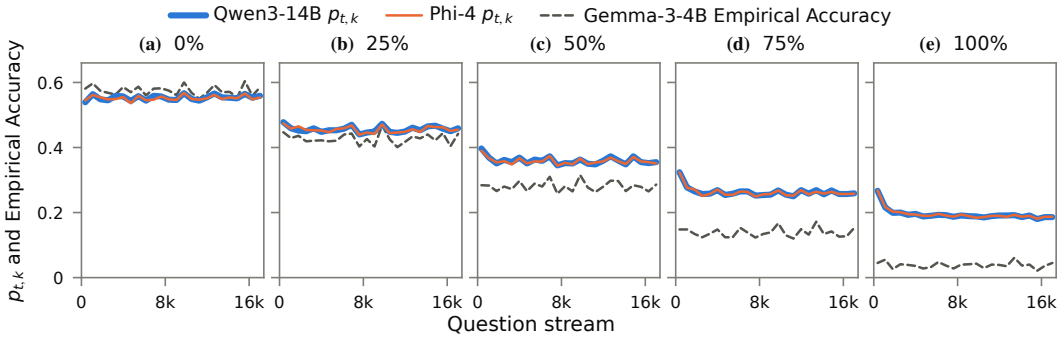


Figure 10: Online reliability estimation for advisor Gemma-3-4B. Subfigures (a)–(e) correspond to misleading information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. Solid curves show the pre-feedback reliability estimate $p_{t,k}$ produced by Qwen3-14B and Phi-4, while the dashed curve shows the empirical accuracy of Gemma-3-4B.

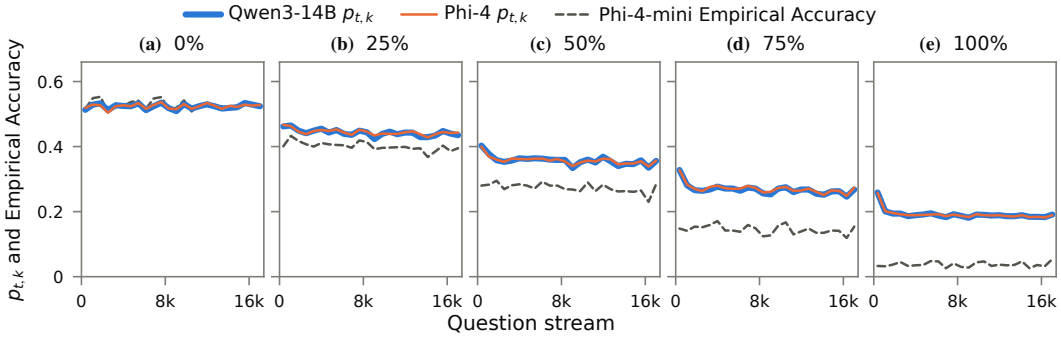


Figure 11: Online reliability estimation for advisor Phi-4-mini. Subfigures (a)–(e) correspond to misleading information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. Solid curves show the pre-feedback reliability estimate $p_{t,k}$ produced by Qwen3-14B and Phi-4, while the dashed curve shows the empirical accuracy of Phi-4-mini.

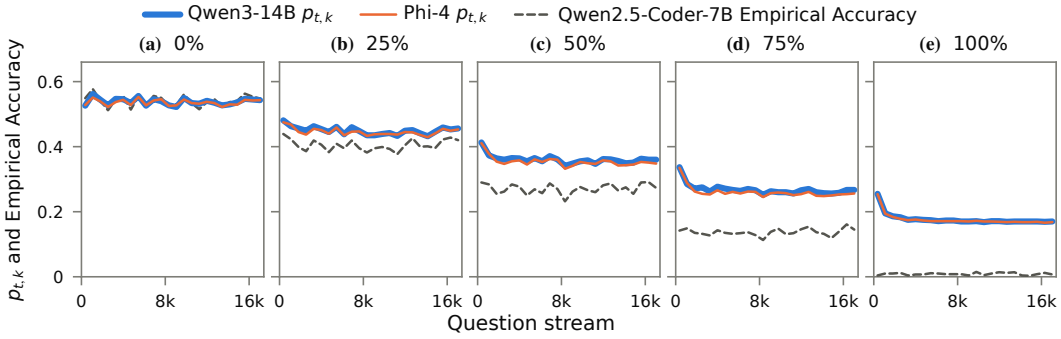


Figure 12: Online reliability estimation for advisor Qwen2.5-Coder-7B. Subfigures (a)–(e) correspond to misleading information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. Solid curves show the pre-feedback reliability estimate $p_{t,k}$ produced by Qwen3-14B and Phi-4, while the dashed curve shows the empirical accuracy of Qwen2.5-Coder-7B.

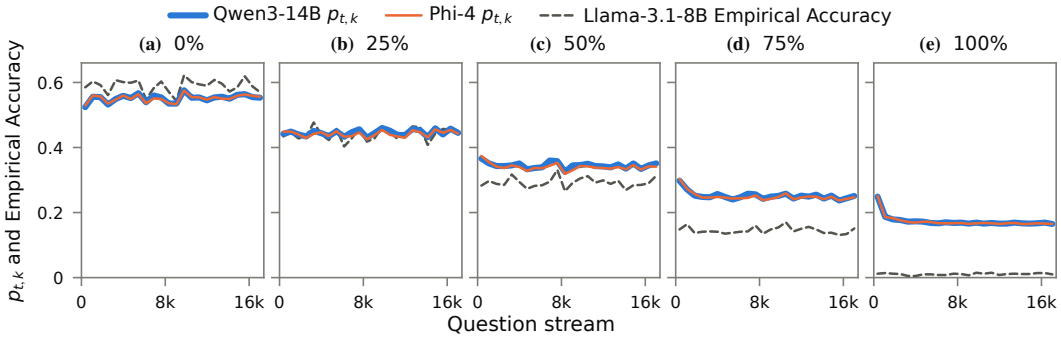


Figure 13: Online reliability estimation for advisor Llama-3.1-8B. Subfigures (a)–(e) correspond to misleading information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. Solid curves show the pre-feedback reliability estimate $p_{t,k}$ produced by Qwen3-14B and Phi-4, while the dashed curve shows the empirical accuracy of Llama-3.1-8B.

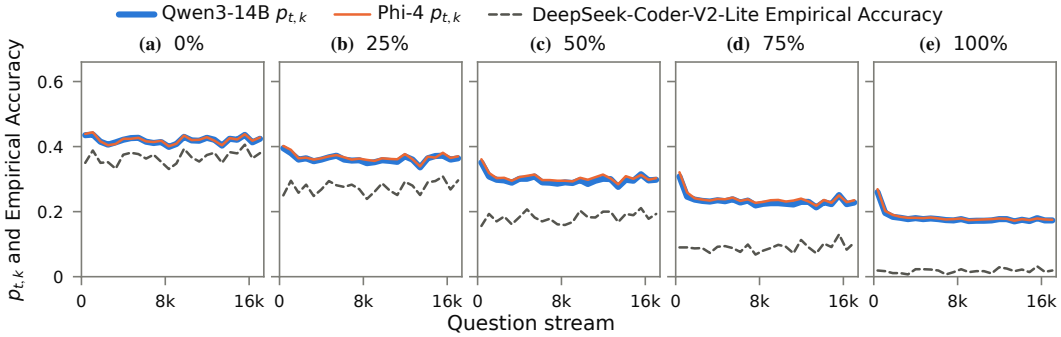


Figure 14: Online reliability estimation for advisor DeepSeek-Coder-V2-Lite. Subfigures (a)–(e) correspond to misleading information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. Solid curves show the pre-feedback reliability estimate $p_{t,k}$ produced by Qwen3-14B and Phi-4, while the dashed curve shows the empirical accuracy of DeepSeek-Coder-V2-Lite.

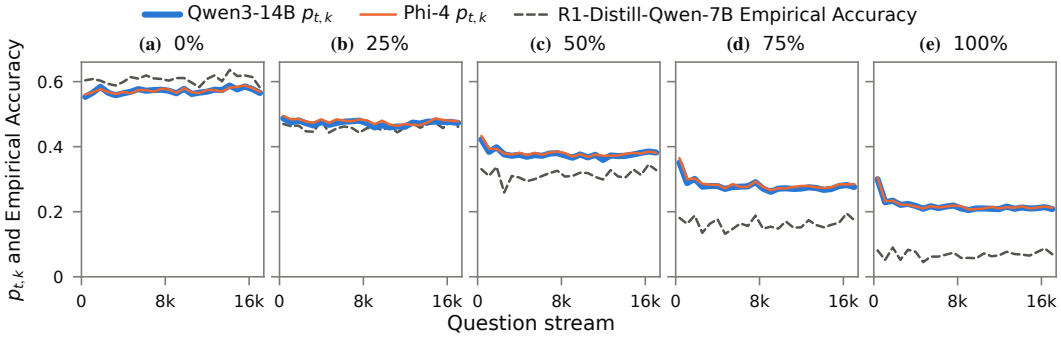


Figure 15: Online reliability estimation for advisor R1-Distill-Qwen-7B. Subfigures (a)–(e) correspond to misleading information ratios of 0%, 25%, 50%, 75%, and 100%, respectively. Solid curves show the pre-feedback reliability estimate $p_{t,k}$ produced by Qwen3-14B and Phi-4, while the dashed curve shows the empirical accuracy of R1-Distill-Qwen-7B.

B.5 Consultation Robustness Across Additional Central Models

We report complete results for six central models: Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen2.5-7B, Ministral-8B, and Phi-4. Tables 4–9 and Figures 16–21 extend the analysis in the main text Sec. 4.2 across different model sizes and families.

Obs. 1. Historical reliability improves consultation across models. Across both capability regimes and all tested misleading information ratios, **Advisors + memory** *achieves higher accuracy than* **Question + Peers** for every central model. At 50% misleading information in the ***capability-challenging*** regime, adding reliability memory improves accuracy by 9.4%, 12.6%, and 9.1% percentage points for Qwen3-4B, Qwen3-8B, and Qwen3-14B, respectively. The corresponding improvements are 11.8% points for Qwen2.5-7B, 19.9% points for Ministral-8B, and 7.7% points for Phi-4. Since both methods consult on every question, this comparison demonstrates the benefit of incorporating historical reliability into the consultation process. The consistent gains support using accumulated correctness evidence to guide advisor influence rather than relying solely on the current responses.

Obs. 2. Autonomous selection complements reliability-guided consultation. The benefit of choosing whether to consult becomes particularly clear when misleading information dominates. At 100% misleading information in the ***capability-challenging*** regime, **Question + Peers**, **Debate (2 rounds)**, and **both majority voting** baselines *fall below* **No consultation** for all six central models. Although **Advisors + memory** improves consultation, it also *falls below* the **autonomous** baseline under this condition. In contrast, **BaRe-Mem** *exceeds its always consult ablation* by 13.0%–34.6% percentage points and remains 1.5%–2.8% points *above autonomous reasoning*. Moreover, it *stays above* the **autonomous baseline** throughout the tested misleading information range in this regime for all six models. These results support the complementary roles of the two mechanisms: reliability-guided attention improves how advisor responses are used, while autonomous selection provides an alternative when continued consultation becomes harmful.

Obs. 3. Consultation behavior adapts to the model and capability regime. The consultation ratios provide a behavioral explanation for these results. In the ***capability-supported*** regime, Qwen3-4B, Qwen3-8B, Qwen3-14B, and Phi-4 maintain high consultation ratios, with endpoint decreases of only 1%–3% points as misleading information increases from 0% to 100%. Qwen2.5-7B and Ministral-8B reduce consultation more noticeably, from 84% to 77% and from 85% to 71%, respectively, but still consult on most questions. In the ***capability-challenging*** regime, **all six models exhibit a much stronger shift**: consultation ratios decrease from 85%–90% without injected misleading information to 9%–21% at the highest misleading information ratio. Thus, the same increase in misleading information produces different consultation behavior across models and capability regimes. This pattern is consistent with comparing estimated consultation and autonomous abilities on each question, rather than applying a fixed consultation policy.

Table 4: Accuracy (%) of consultation methods with Qwen3-4B as the central model under increasing misleading information ratios. Results are reported for the capability-supported and capability-challenging regimes. Bold numbers indicate the best performance in each column.

Method	Capability-supported					Capability-challenging				
	0%	**25%**	**50%**	**75%**	**100%**	**0%**	**25%**	**50%**	**75%**	**100%**
No consultation	69.2	69.2	69.2	69.2	69.2	67.8	67.8	67.8	67.8	67.8
Question + Peers	74.6	74.2	73.7	72.4	71.3	69.2	64.0	58.1	51.4	45.8
Majority vote (advisors)	64.3	53.0	32.5	14.6	9.4	64.1	48.7	24.5	5.9	1.2
Majority vote (advisors + own)	68.0	61.8	46.9	27.3	18.8	68.2	56.8	34.3	10.4	3.0
Debate (2 rounds)	73.5	73.2	73.0	72.3	72.1	71.4	68.6	64.6	60.0	56.1
Advisors + memory	**76.7**	**75.5**	**74.3**	**73.1**	**72.4**	73.8	70.4	67.5	62.5	50.7
BaRe-Mem	76.2	75.4	74.1	72.7	72.0	**75.0**	**72.4**	**70.8**	**70.2**	**69.3**
Consultation ratio	86%	85%	85%	86%	85%	86%	78%	69%	53%	15%

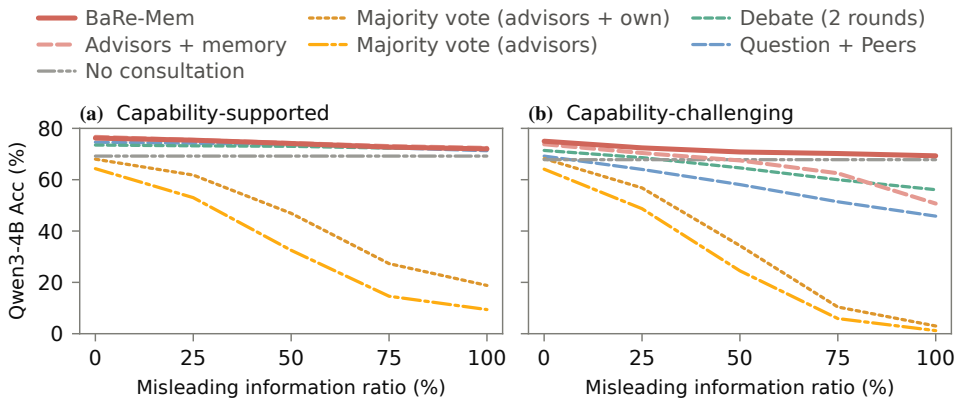


Figure 16: Accuracy of Qwen3-4B as the central model under increasing misleading information ratios. **Left:** capability-supported regime; **right:** capability-challenging regime. We compare BaRe-Mem with *Advisors + memory*, *No consultation*, *Question + Peers*, *Debate (2 rounds)*, and two majority voting baselines.

Table 5: Accuracy (%) of consultation methods with Qwen3-8B as the central model under increasing misleading information ratios. Results are reported for the capability-supported and capability-challenging regimes. Bold numbers indicate the best performance in each column.

Method	**Capability-supported**					**Capability-challenging**				
	0%	**25%**	**50%**	**75%**	**100%**	**0%**	**25%**	**50%**	**75%**	**100%**
No consultation	72.9	72.9	72.9	72.9	72.9	65.6	65.6	65.6	65.6	65.6
Question + Peers	75.7	75.1	74.3	73.6	72.3	67.9	61.6	54.3	46.7	40.0
Majority vote (advisors)	64.3	53.0	32.5	14.6	9.4	64.1	48.7	24.5	5.9	1.2
Majority vote (advisors + own)	68.9	62.7	47.6	27.9	19.3	66.5	55.5	33.3	9.7	2.7
Debate (2 rounds)	74.8	74.8	74.6	74.4	73.8	67.0	63.9	60.3	55.7	51.3
Advisors + memory	**78.1**	**77.0**	**76.3**	**74.8**	**74.0**	74.2	70.1	66.9	61.0	46.2
BaRe-Mem	77.4	76.6	76.1	74.6	73.7	**74.9**	**71.4**	**70.0**	**68.7**	**67.7**
Consultation ratio	86%	85%	85%	84%	83%	89%	80%	70%	52%	14%

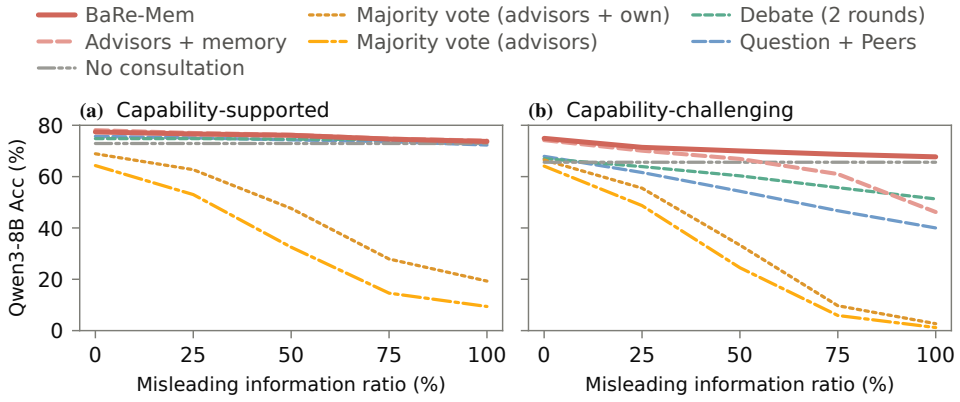


Figure 17: Accuracy of Qwen3-8B as the central model under increasing misleading information ratios. **Left:** capability-supported regime; **right:** capability-challenging regime. We compare BaRe-Mem with *Advisors + memory*, *No consultation*, *Question + Peers*, *Debate (2 rounds)*, and two majority voting baselines.

Table 6: Accuracy (%) of consultation methods with Qwen3-14B as the central model under increasing misleading information ratios. Results are reported for the capability-supported and capability-challenging regimes. Bold numbers indicate the best performance in each column.

Method	Capability-supported					Capability-challenging				
	0%	**25%**	**50%**	**75%**	**100%**	**0%**	**25%**	**50%**	**75%**	**100%**
No consultation	72.9	72.9	72.9	72.9	72.9	66.5	66.5	66.5	66.5	66.5
Question + Peers	76.6	75.9	75.3	73.9	73.2	69.8	65.6	60.4	55.2	50.6
Majority vote (advisors)	64.3	53.0	32.5	14.6	9.4	64.1	48.7	24.5	5.9	1.2
Majority vote (advisors + own)	68.5	62.6	47.5	27.8	19.4	66.4	55.8	33.6	9.9	2.8
Debate (2 rounds)	74.9	74.7	73.9	74.0	73.4	69.2	66.7	62.6	58.9	55.4
Advisors + memory	**78.0**	**77.4**	**76.2**	**75.2**	**74.5**	75.6	72.1	69.5	65.2	54.7
BaRe-Mem	77.7	77.3	76.0	75.1	**74.5**	**76.1**	**72.8**	**71.6**	**69.9**	**68.6**
Consultation ratio	87%	86%	85%	85%	85%	90%	83%	76%	61%	21%

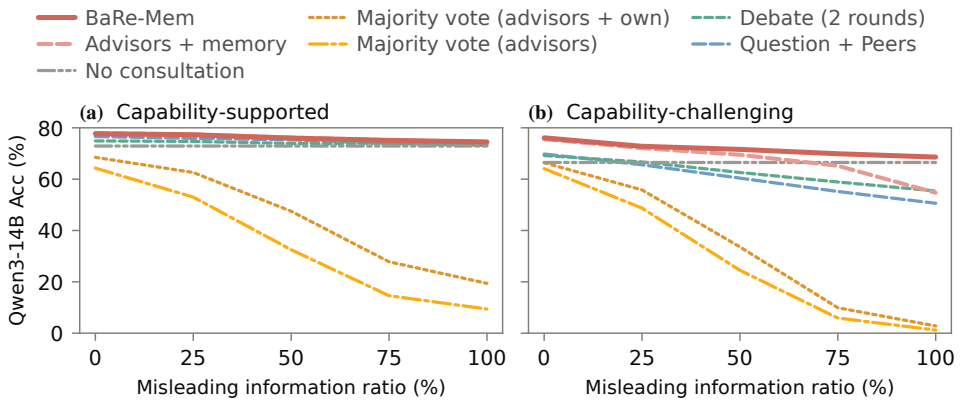


Figure 18: Accuracy of Qwen3-14B as the central model under increasing misleading information ratios. **Left:** capability-supported regime; **right:** capability-challenging regime. We compare BaRe-Mem with *Advisors + memory*, *No consultation*, *Question + Peers*, *Debate (2 rounds)*, and two majority voting baselines.

Table 7: Accuracy (%) of consultation methods with Qwen2.5-7B as the central model under increasing misleading information ratios. Results are reported for the capability-supported and capability-challenging regimes. Bold numbers indicate the best performance in each column.

Method	Capability-supported					Capability-challenging				
	0%	**25%**	**50%**	**75%**	**100%**	**0%**	**25%**	**50%**	**75%**	**100%**
No consultation	70.6	70.6	70.6	**70.6**	**70.6**	59.1	59.1	59.1	59.1	59.1
Question + Peers	71.3	70.3	69.2	67.8	67.1	65.9	59.3	52.9	47.3	43.3
Majority vote (advisors)	64.3	53.0	32.5	14.6	9.4	64.1	48.7	24.5	5.9	1.2
Majority vote (advisors + own)	68.4	62.3	47.1	27.3	18.9	64.7	53.5	31.5	8.8	2.2
Debate (2 rounds)	70.5	70.3	69.9	69.2	68.2	63.1	58.6	53.7	49.1	45.5
Advisors + memory	**75.5**	**74.0**	**72.9**	**70.6**	69.2	71.4	68.0	64.7	58.7	48.4
BaRe-Mem	74.9	73.6	72.5	70.5	69.3	**71.5**	**68.3**	**66.1**	**63.5**	**61.9**
Consultation ratio	84%	81%	82%	83%	77%	90%	83%	75%	54%	19%

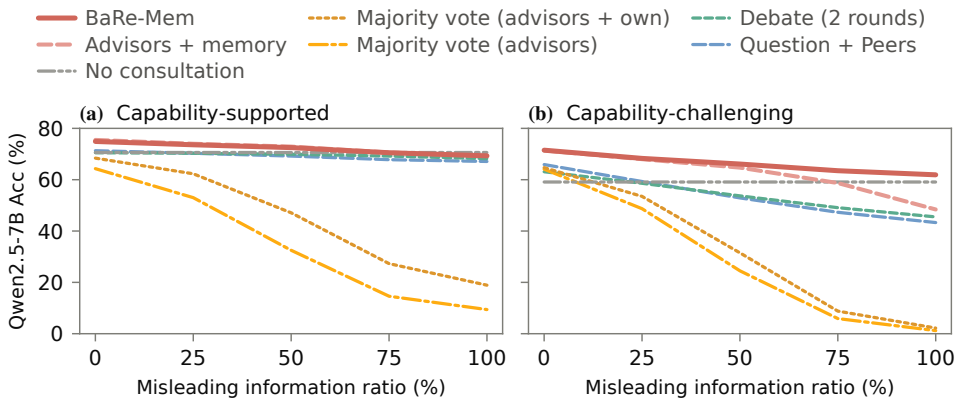


Figure 19: Accuracy of Qwen2.5-7B as the central model under increasing misleading information ratios. **Left:** capability-supported regime; **right:** capability-challenging regime. We compare BaRe-Mem with *Advisors + memory*, *No consultation*, *Question + Peers*, *Debate (2 rounds)*, and two majority voting baselines.

Table 8: Accuracy (%) of consultation methods with Ministral-8B as the central model under increasing misleading information ratios. Results are reported for the capability-supported and capability-challenging regimes. Bold numbers indicate the best performance in each column.

Method	**Capability-supported**					**Capability-challenging**				
	0%	**25%**	**50%**	**75%**	**100%**	**0%**	**25%**	**50%**	**75%**	**100%**
No consultation	56.5	56.5	56.5	56.5	56.5	58.6	58.6	58.6	58.6	58.6
Question + Peers	64.2	63.1	61.1	58.5	56.9	64.0	54.8	41.6	27.4	18.0
Majority vote (advisors)	64.3	53.0	32.5	14.6	9.4	64.1	48.7	24.5	5.9	1.2
Majority vote (advisors + own)	67.1	59.7	43.6	24.9	17.1	64.2	53.3	31.5	8.9	2.2
Debate (2 rounds)	61.8	60.8	59.0	57.0	56.7	61.5	56.5	49.7	41.3	34.4
Advisors + memory	**70.8**	**67.8**	**64.6**	60.3	60.0	**70.8**	65.8	61.5	52.3	26.0
BaRe-Mem	70.2	67.2	64.4	**60.8**	**61.0**	**70.8**	**66.5**	**64.2**	**62.7**	**60.6**
Consultation ratio	85%	83%	81%	72%	71%	89%	79%	65%	40%	9%

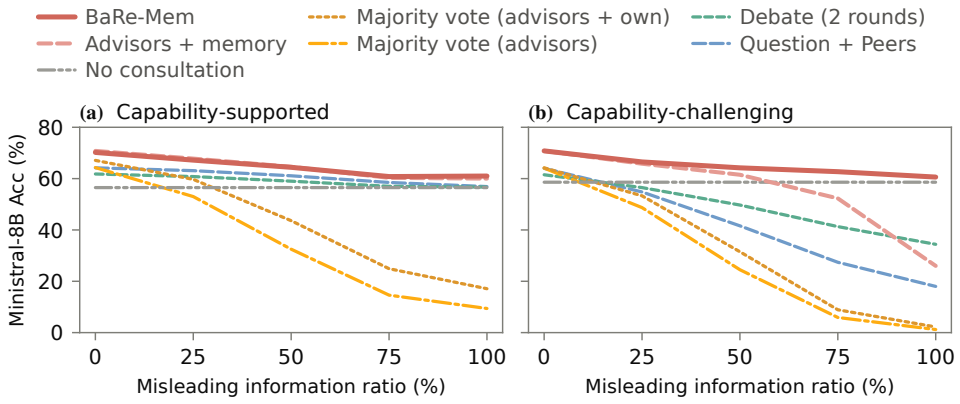


Figure 20: Accuracy under increasing misleading-advice ratios with Ministral-8B as the central model. The left and right panels show the capability-supported and capability-challenging regimes, respectively. We compare BaRe-Mem, *Advisors + memory*, *No consultation*, *Question + Peers*, *Debate (2 rounds)*, and two majority-vote variants.

Table 9: Accuracy (%) of consultation methods with Phi-4 as the central model under increasing misleading information ratios. Results are reported for the capability-supported and capability-challenging regimes. Bold numbers indicate the best performance in each column.

Method	Capability-supported					Capability-challenging				
	0%	**25%**	**50%**	**75%**	**100%**	**0%**	**25%**	**50%**	**75%**	**100%**
No consultation	70.2	70.2	70.2	70.2	70.2	63.4	63.4	63.4	63.4	63.4
Question + Peers	71.5	72.1	71.6	70.8	69.9	65.8	62.4	57.9	52.5	48.3
Majority vote (advisors)	64.3	53.0	32.5	14.6	9.4	64.1	48.7	24.5	5.9	1.2
Majority vote (advisors + own)	67.9	61.5	46.3	26.7	18.3	65.1	54.3	32.1	9.0	2.3
Debate (2 rounds)	68.0	68.8	69.4	69.6	68.6	64.8	62.6	59.3	55.2	51.4
Advisors + memory	76.3	75.0	73.6	72.9	71.6	**71.3**	68.4	65.6	61.8	52.0
BaRe-Mem	**76.5**	**75.2**	**73.9**	**73.0**	**71.8**	71.2	**68.5**	**66.8**	**65.3**	**65.0**
Consultation ratio	85%	84%	83%	84%	83%	85%	79%	71%	57%	18%

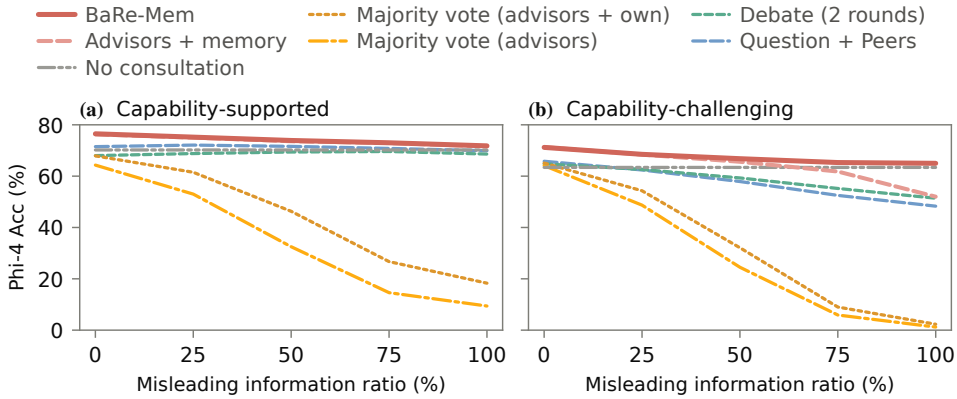


Figure 21: Accuracy of Phi-4 as the central model under increasing misleading information ratios. **Left:** capability-supported regime; **right:** capability-challenging regime. We compare BaRe-Mem with *Advisors + memory*, *No consultation*, *Question + Peers*, *Debate (2 rounds)*, and two majority voting baselines.

C Agent Team

C.1 Agent Team Pipeline.

Figure 22 illustrates the agent team workflow. The lead agent first decomposes each task into sub-tasks and uses BaRe-Mem to rank the candidate workers for each sub-task. The highest rank worker is allocated first. If its report is rejected, the lead agent retries with the next worker in the same ranking. In this setting, BaRe-Mem is used for worker routing rather than reweighting a fixed set of responses, deciding both which worker is selected first and which worker is tried next. To isolate the effect of worker routing from task decomposition quality, we use the MuSiQue (Trivedi et al., 2022) sub-tasks provided by the dataset, so all experimental differences arise from how these sub-tasks are assigned to workers.

Checking and Feedback Protocol. We distinguish the check used to control retries from the feedback used to update reliability memory. We consider three checking settings: ***No check***, where the first report is directly committed; ***Lead agent check***, where the lead agent decides whether to accept the report or retry; and ***Dataset verifier check***, where the dataset evaluator makes this decision. In all three settings, after a sub-task is committed, the gold correctness labels of all workers actually queried for that sub-task are written to memory, while unqueried workers receive no update. Thus, checking determines the execution path of the current sub-task, whereas verified feedback updates worker reliability for subsequent routing decisions. Memory is updated after each sub-task, allowing later sub-tasks to benefit from earlier verified outcomes.

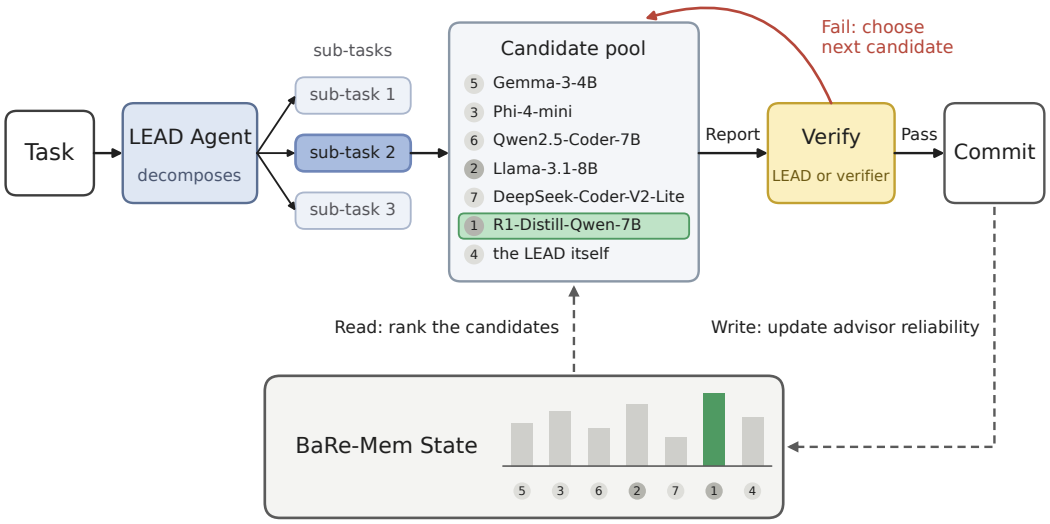


Figure 22: Agent team pipeline with BaRe-Mem. The lead agent decomposes a task into sub-tasks and queries the reliability memory to rank candidate workers. Each returned report is verified by the lead agent or an external verifier: accepted reports are committed, while rejected reports trigger the next candidate. Verification outcomes are then written back to update advisor reliability.

C.2 Task Solved Accuracy

Figures 23–28 report task and sub-task completion for six lead models: Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen2.5-7B, Ministral-8B, and Phi-4. These results examine how reliability memory supports worker assignment and subsequent retries across different central models.

Obs.1. Contextual reliability improves initial worker selection. The ***No check*** setting commits the first worker's report without retrying, directly evaluating the initial assignment. In this setting, **BaRe-Mem achieves higher sub-task completion than routing randomly or by historical success counts for all six lead models**. The improvements are 5.0%, 1.4%, and 3.8% points for Qwen3-4B, Qwen3-8B, and Qwen3-14B, respectively, and 5.6%, 2.1%, and 6.9% points for Qwen2.5-7B, Ministral-8B, and Phi-4. Relative to random routing, it also improves task completion by 13.4%–17.8% points across the six models. Because these gains arise without retries, they support the use of reliability conditioned on the current sub-task to identify a suitable worker at the initial assignment. The comparison with historical success counts further indicates that the benefit extends beyond simply retaining aggregate records of past success.

Obs.2. Reliability ranking complements verification in Agent Team Under ***Dataset verifier check***, **BaRe-Mem achieves the highest task and sub-task completion for every lead model**. Task completion reaches 51.9%–55.0%, exceeding historical success counts by 1.8%–4.9% points and random routing by 9.1–11.4 points. Sub-task completion also improves over historical success counts by 1.7%–3.5% points. These comparisons use the same checking rule and worker budget for each routing strategy, showing that access to a verifier does not remove the benefit of our BaRe-Mem reliability worker ranking. When the first report is rejected, the reliability ranking also determines which worker is tried next. The results support using the memory to guide candidate ordering throughout the assignment and retry process.

Obs.3. The effect of report checking depends on the lead model. Lead agent checking improves BaRe-Mem's task completion over ***No check*** for five of the six lead models, with gains of 1.5%–3.0% points. Its effect is not uniform: with Ministral-8B, task completion changes from 41.6% without checking to 40.8% with lead agent checking, whereas dataset verifier checking raises it to 55.0%. Across all six models, replacing lead agent checking with dataset verifier checking improves BaRe-Mem's task completion by 11.6%–14.2% points. Together, these results distinguish the com-

plementary roles of routing and checking: reliability memory orders candidate workers, while report checking determines when to proceed to the next candidate.

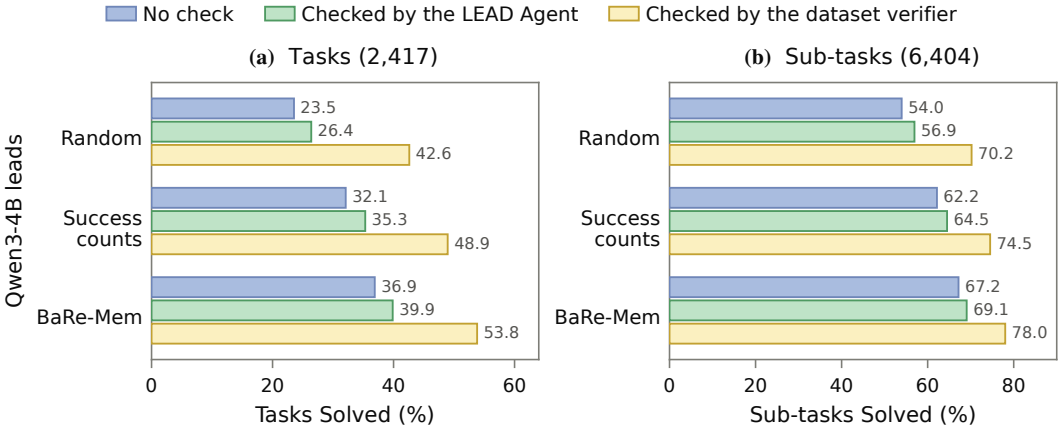


Figure 23: Agent team performance on MuSiQue with Qwen3-4B as the lead agent. **(a)** Task completion; **(b)** sub-task completion. We compare random routing, routing by historical success counts, and BaRe-Mem under three report checking settings: no check, lead agent check, and dataset verifier check. All settings provide gold feedback for queried reports after each sub-task is committed.

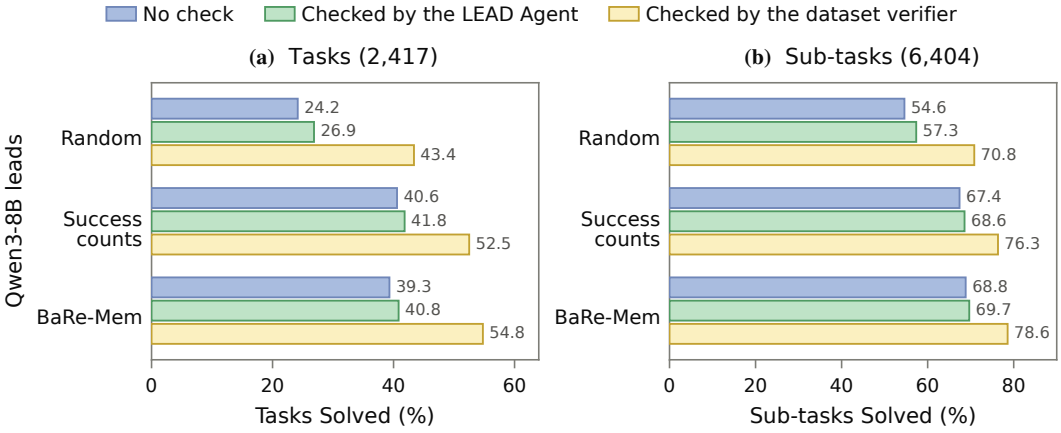


Figure 24: Agent team performance on MuSiQue with Qwen3-8B as the lead agent. **(a)** Task completion; **(b)** sub-task completion. We compare random routing, routing by historical success counts, and BaRe-Mem under three report checking settings: no check, lead agent check, and dataset verifier check. All settings provide gold feedback for queried reports after each sub-task is committed.

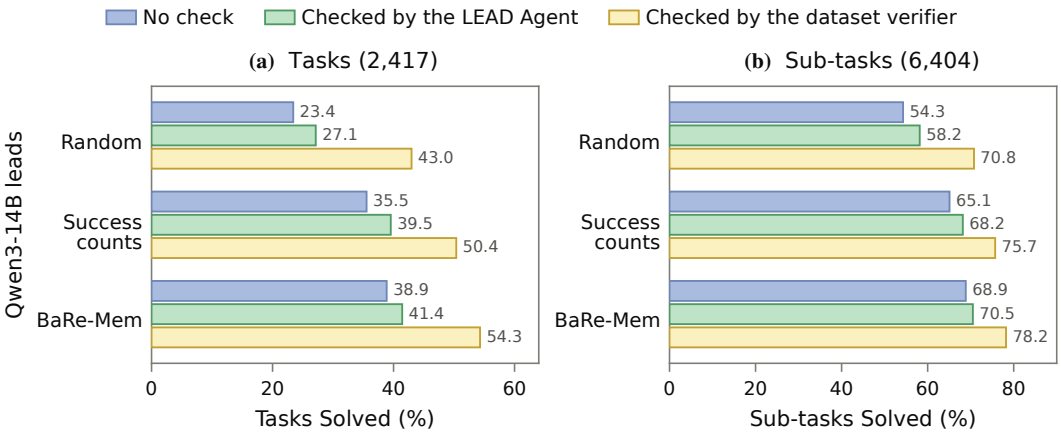


Figure 25: Agent team performance on MuSiQue with Qwen3-14B as the lead agent. **(a)** Task completion; **(b)** sub-task completion. We compare random routing, routing by historical success counts, and BaRe-Mem under three report checking settings: no check, lead agent check, and dataset verifier check. All settings provide gold feedback for queried reports after each sub-task is committed.

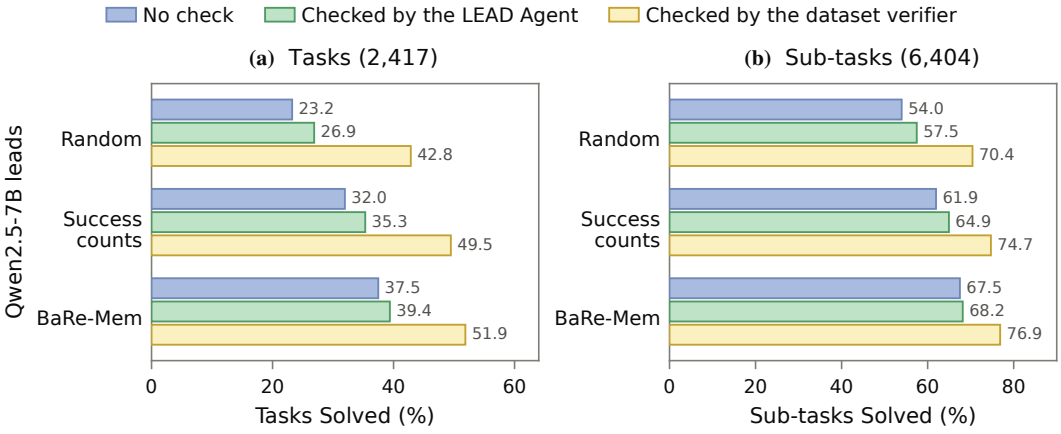


Figure 26: Agent team performance on MuSiQue with Qwen2.5-7B as the lead agent. **(a)** Task completion; **(b)** sub-task completion. We compare random routing, routing by historical success counts, and BaRe-Mem under three report checking settings: no check, lead agent check, and dataset verifier check. All settings provide gold feedback for queried reports after each sub-task is committed.

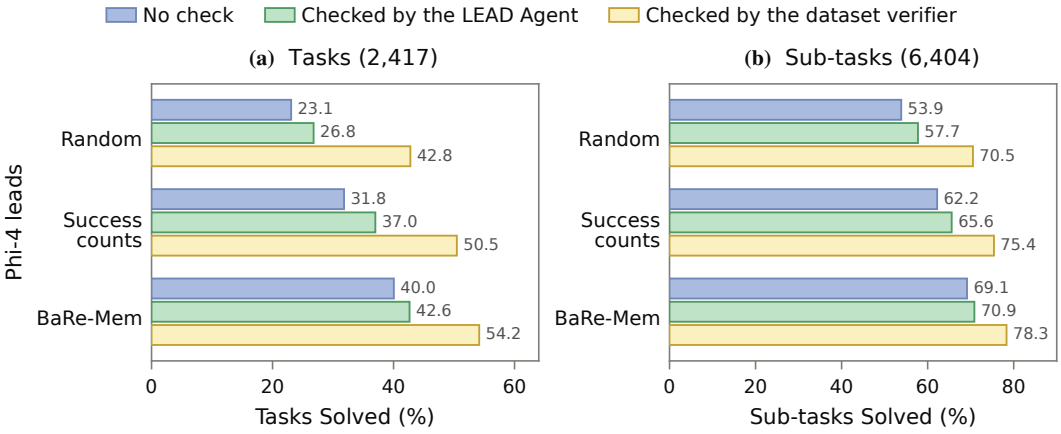


Figure 27: Agent team performance on MuSiQue with Phi-4 as the lead agent. **(a)** Task completion; **(b)** sub-task completion. We compare random routing, routing by historical success counts, and BaRe-Mem under three report checking settings: no check, lead agent check, and dataset verifier check. All settings provide gold feedback for queried reports after each sub-task is committed.

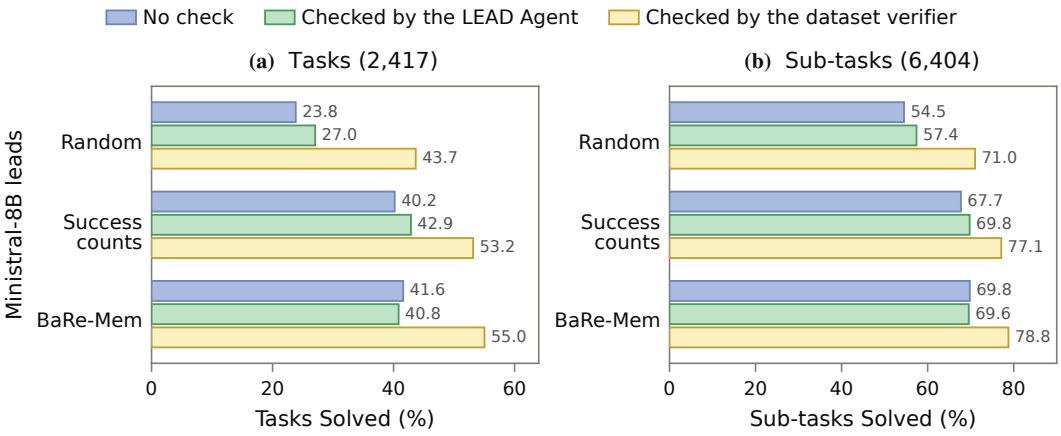


Figure 28: Agent team performance on MuSiQue with Ministral-8B as the lead agent. **(a)** Task completion; **(b)** sub-task completion. We compare random routing, routing by historical success counts, and BaRe-Mem under three report checking settings: no check, lead agent check, and dataset verifier check. All settings provide gold feedback for queried reports after each sub-task is committed.

C.3 Agent Team Sweep

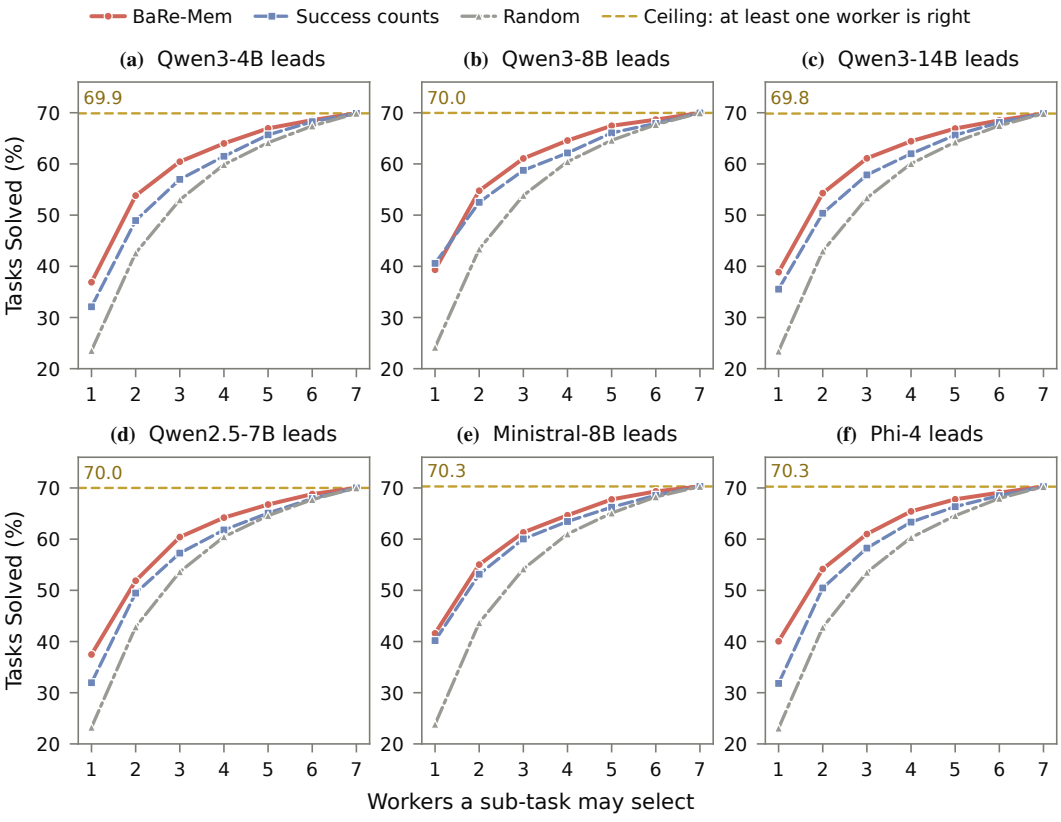


Figure 29: Task completion on MuSiQue as the maximum number of workers tried per sub-task increases. We compare BaRe-Mem, routing by historical success counts, and random routing. The dashed line marks the ceiling obtained when at least one worker can solve the sub-task. Panels (a)–(f) use Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen2.5-7B, Ministral-8B, and Phi-4 as the lead agent, respectively.

We examine how quickly each routing strategy can identify a worker capable of solving the current sub-task for all the central models following the setting in the main text Sec. **4.4**. As the worker

budget increases, all methods gradually approach the same ceiling once all seven workers have been tried. This is expected, because the candidate pool is identical across routing strategies. The remaining difference is therefore not *whether* a solvable worker exists, but *how early* that worker is ranked and selected.

Obs. 1. BaRe-Mem identifies capable workers earlier. Figures 29 and 30 show that **BaRe-Mem consistently achieves higher task and sub-task completion with fewer worker calls across all six lead models**. The advantage is largest when only a small number of workers may be tried. With a single worker per sub-task, BaRe-Mem already outperforms routing by historical success counts for every lead model. As the worker budget increases, all methods gradually approach the same ceiling because they eventually access the same candidate pool. Therefore, the gap at smaller budgets reflects how early a capable worker is ranked and selected. BaRe-Mem reaches a larger fraction of the achievable performance with the same worker budget.

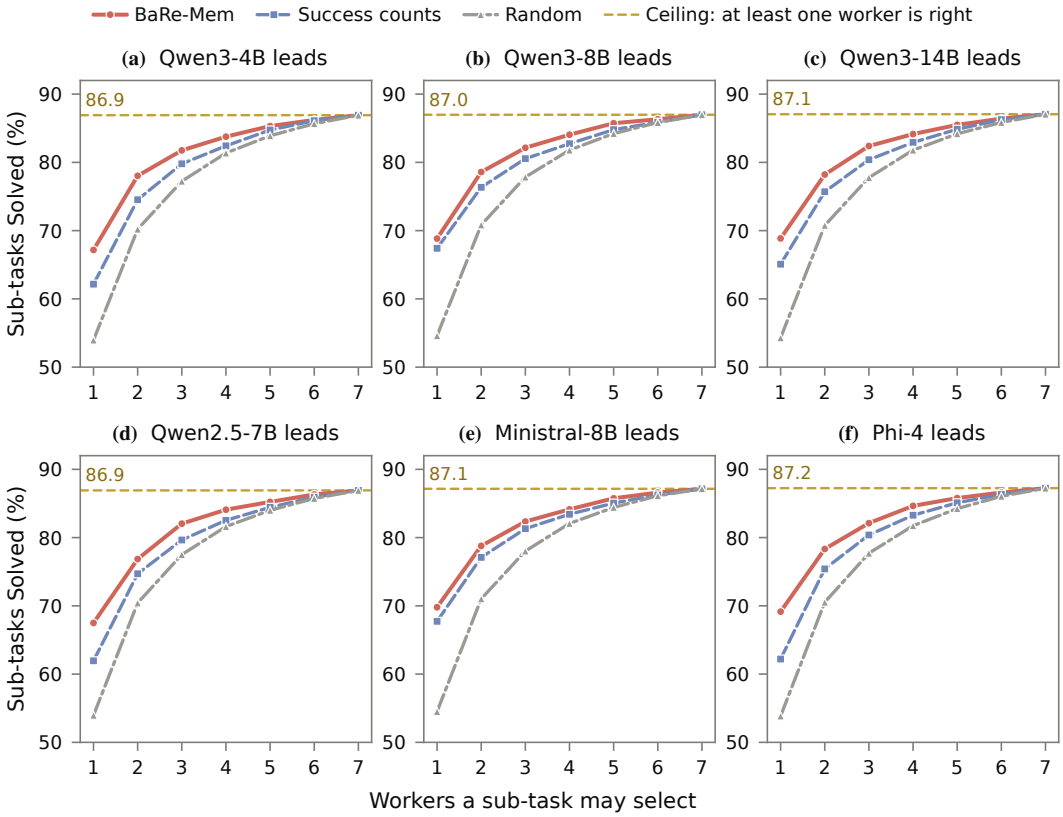


Figure 30: Sub-task completion on MuSiQue as the maximum number of workers tried per sub-task increases. We compare BaRe-Mem, routing by historical success counts, and random routing. The dashed line marks the ceiling given by the fraction of sub-tasks for which at least one worker is correct. Panels (a)–(f) use Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen2.5-7B, Ministral-8B, and Phi-4 as the lead agent, respectively.

Obs. 2. Contextual reliability improves worker ranking. The comparison with historical success counts isolates the value of contextual reliability. Success counts assign each worker a reliability based only on its aggregate past performance, whereas BaRe-Mem conditions worker reliability on the current sub-task. **The consistent advantage of BaRe-Mem across all six lead models** therefore shows that worker competence cannot be adequately represented by a single source-level success rate. Conditioning reliability on the current sub-task yields a more informative worker ordering, supporting the same instance-specific reliability principle used in advisor consultation.

C.4 Adapt Speed

We finally examine how quickly BaRe-Mem learns useful worker reliability from the online feedback stream. Following the original sub-task order, we partition the 6,404 MuSiQue sub-tasks into eight consecutive blocks of approximately 800 sub-tasks each and measure the accuracy of the **initial worker** selected in each block. This directly measures how the quality of the learned worker ranking evolves as verified interactions accumulate along the stream.

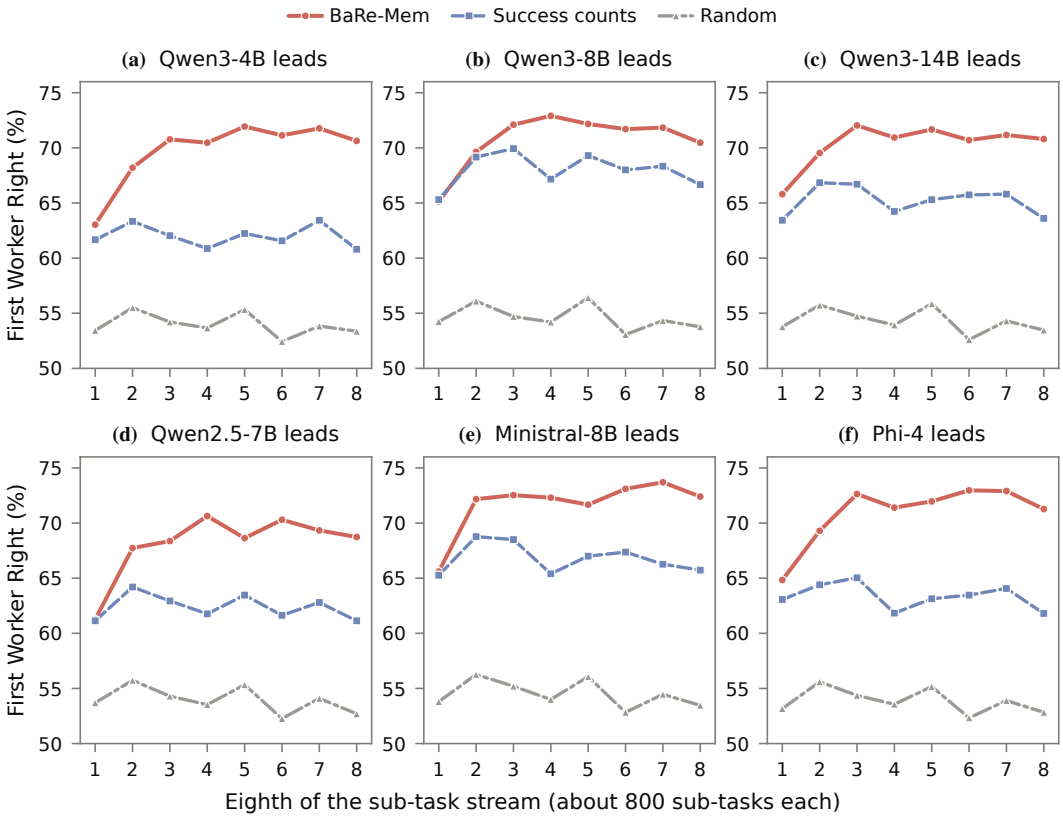


Figure 31: Initial worker accuracy over the MuSiQue sub-task stream. The x-axis denotes position in the sub-task stream, grouped into eight consecutive blocks of approximately 800 sub-tasks each. We compare BaRe-Mem, routing by historical success counts, and random routing. Panels (a)–(f) use Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen2.5-7B, Ministral-8B, and Phi-4 as the lead agent, respectively.

BaRe-Mem adapts rapidly from early feedback. Figure 31 shows that initial worker accuracy rises quickly during the early part of the stream for all six lead models and then remains relatively stable. Historical success counts also improve as feedback accumulates, but **BaRe-Mem separates from this baseline early and maintains higher initial worker accuracy throughout the stream**. In contrast, random routing remains nearly unchanged. These results show that **BaRe-Mem** can rapidly extract useful reliability information from early interactions.

This adaptation pattern is also consistent with the Bayesian update in Section A.2: when posterior uncertainty is high, the Kalman gain gives early verified outcomes greater influence, while accumulated evidence reduces uncertainty and stabilizes subsequent updates. The resulting behavior allows **BaRe-Mem** to adapt quickly without requiring a separate retraining stage.