Budgeted Act-or-Defer Multi-Agent Deliberation with Local Reliability Bounds

Mengdie Flora Wang [1] **Haochen Xie** [1] **Guanghui Wang** [1] **Devin Zhang** [2] **Jae Oh Woo** [1]

Abstract

Multi-agent deliberation among LLMs can improve reasoning, but deployment requires deciding when the current answer is reliable enough to act on and when it should be escalated to human review. We formulate this as budgeted act-or-defer decision making. At each round, the system maps the debate prefix to a low-dimensional state, computes a k-nearest-neighbor lower confidence bound on state-conditional correctness using calibration data, and acts only when the bound exceeds a user-specified reliability threshold. The certificate controls wrong actions through the decomposition $\beta = \delta + \alpha + \varepsilon_{\text{act}}$, separating calibration failure, residual action risk, and representation gap. The guarantee is conditional, not distribution-free: it relies on a valid local bias envelope and an action-region representation-gap bound, and each assumption is paired with falsification-style diagnostics. Because the same absolute wrong-action budget has different meanings across tasks of different difficulty, we set budgets relative to each task's final-round error using training data only, and evaluate safety by normalized budget usage WA/β. On six benchmarks against nine baselines, the method uses 9–12% of the pre-declared budget on activated datasets, reaching up to 84% automation and 96% acted-on accuracy; on stress-test datasets, it defers rather than forcing unreliable automation. Rather than relying on per-task post-hoc threshold search, the method prospectively converts a user-declared wrong-action budget into an auditable act-or-defer operating point before deployment, under explicitly stated assumptions.

[1]AWS GenAI Innovation Center, Amazon, USA [2]General Motors (GM), USA. Correspondence to: Jae Oh Woo <jaeohwoo@amazon.com>.

Workshop on Failure Modes in Agentic AI at the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Multi-agent debate, where multiple LLM instances exchange responses over several rounds, has been shown to improve reasoning (Du et al., 2024; Liang et al., 2024). To deploy such systems in practice, one needs a stopping criterion with principled control over the probability of acting on a wrong answer. In human-in-the-loop deployments this stopping decision is also an escalation decision: the system should act autonomously only when the local reliability bound supports action, and otherwise defer the instance to human review. Existing approaches either operate at a fixed final round with only marginal coverage (e.g., split conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2023)) or rely on heuristic consensus thresholds without explicit wrong-action budgets. Neither provides per-instance, per-round reliability bounds that enable adaptive stopping.

We address this gap by constructing local lower confidence bounds on the correctness probability using k-nearest-neighbor (k-NN) neighborhoods in a debate-state metric space. The debate prefix at round t is mapped to a state U_t in a metric space $(\mathcal{U}, d)$. The local correctness function $q_t(u) = \Pr(Y = \hat{Y}_t \mid U_t = u)$ is bounded from below via a *local bias envelope* $b_{t,k}(u)$ that controls the k-NN approximation error. For each candidate neighborhood size k in a pre-specified family $\mathcal{K}$, we subtract the bias envelope and a Hoeffding concentration term from the k-NN empirical mean, then take the maximum over $\mathcal{K}$:

$$L_t = \max_{k \in \mathcal{K}} \left[\widehat{q}_{t,k}(U_t) - b_{t,k}(U_t) - \sqrt{\frac{\log(T|\mathcal{K}|/\delta)}{2k}} \right], \quad (1)$$

where the $\log|\mathcal{K}|$ penalty absorbs the adaptivity in k. The bias envelope $b_{t,k}(u)$ is a generic primitive (Lipschitz, Hölder, bounded-oscillation models are special cases). The first correction controls approximation bias, the second controls sampling variance, and the maximization over $\mathcal{K}$ selects the tightest local scale per instance.

A central design choice is *budget alignment*. The user specifies a risk budget β and the method decomposes it as

$$\beta = \underbrace{\delta}_{\substack{\text{statistical}\\\text{failure}}} + \underbrace{\alpha}_{\substack{\text{residual}\\\text{action risk}}} + \underbrace{\varepsilon_{\text{act}}}_{\substack{\text{representation}\\\text{gap}}} . \quad (2)$$

The residual $\alpha = \beta - \delta - \varepsilon_{\text{act}}$ directly sets the stopping threshold $1 - \alpha$. The stopping rule $\tau = \inf\{t : L_t \geq 1 - \alpha\}$

then satisfies $\Pr(\text{act and wrong}) \leq \beta$ conditional on local assumptions (Theorem 5.3). The contribution is not the mathematical tools (Hoeffding + union bound, which are classical) but a risk-decomposition framework for adaptive deliberation stopping under auditable local assumptions.

The guarantee is conditional on auditable local assumptions, not distribution-free. Our contributions: **(i)** We formalize adaptive stopping for multi-agent deliberation as a local state-conditional risk-control problem with budget decomposition $\beta = \delta + \alpha + \varepsilon_{\text{act}}$. **(ii)** We introduce a local bias envelope $b_{t,k}(u)$ as a plug-in interface and prove simultaneous validity $L_t \leq q_t(U_t)$ for all (t, k) with probability $\geq 1 - \delta$, plus a wrong-action guarantee (Theorems 5.1–5.3) and an oracle slack inequality (Theorem A.2, Appendix A). Each assumption is paired with a diagnostic proxy. **(iii)** Lipschitz, Hölder, and bounded-oscillation envelopes are recovered as corollaries. Our default empirical modulus absorbs estimation failure into δ (Corollary A.4) and converges at $O(n^{-1/(2+D)})$. **(iv)** We introduce a relative-risk deployment policy: $\beta_d = \lambda_d^\star \widehat{e}_{T,d}^{\text{train}}$ is induced from training data alone, making budgets comparable across tasks and turning normalized budget usage WA/β into the primary safety metric. **(v)** Experiments on six benchmarks ($n = 546$–$8{,}678$) against nine baselines show that, on activated datasets, the method uses $\leq$12% of the pre-declared budget at up to 97% acted-on accuracy. Matched-automation diagnostics show that tuned scalar thresholds can achieve similar subset quality after per-task search, but the proposed certificate constructs the operating point prospectively from a declared wrong-action budget and makes the required assumptions explicit.

2. Problem Setup

Consider a multiple-choice task with input space $\mathcal{X}$ and label space $\mathcal{Y} = \{1, \ldots, C\}$. An ensemble of N LLM agents debates for at most T rounds. At round t, the debate has produced a filtration $\mathcal{F}_t$ containing all agent responses and per-agent belief distributions $\{P_j^{(t)}(y \mid x)\}_{j=1}^N$ up to round t. We aggregate these into a *social belief*

$$P_{\text{social}}^{(t)}(y \mid x) \;=\; \frac{1}{N} \sum_{j=1}^{N} P_j^{(t)}(y \mid x), \tag{3}$$

and take $\hat{Y}_t = \arg\max_y P_{\text{social}}^{(t)}(y \mid x)$ as the round-t prediction (Appendix B specifies how each $P_j^{(t)}$ is elicited).

State embedding. Each debate prefix is mapped to a state $U_t = \phi_t(\mathcal{F}_t)$ in a metric space $(\mathcal{U}, d)$ via a deterministic, $\mathcal{F}_t$-measurable ϕ_t. Typical components include the top social probability $p_t^{(1)} = \max_y P_{\text{social}}^{(t)}(y \mid x)$, the margin $\Delta_t = p_t^{(1)} - p_t^{(2)}$, inter-agent disagreement, and top-answer stability; the metric d reflects correctness relevance (e.g., Euclidean on a normalized feature vector).

Geometry–sufficiency trade-off. The choice of ϕ_t jointly controls three quantities central to the framework: **(a)** *smoothness of* q_t, which governs the bias envelope $b_{t,k}$; **(b)** *representation gap* ε_{act}, the information lost in compressing $\mathcal{F}_t$ to U_t; **(c)** *intrinsic dimension* D, which enters the certification rate $O(n^{-1/(2+D)})$. Richer states reduce ε_{act} but raise D and may roughen q_t; minimal states keep D small and q_t smooth but risk larger ε_{act}. We investigate this trade-off empirically in Section H.5.

Local correctness and goal. The target is the *local correctness function*

$$q_t(u) := \Pr(Y = \hat{Y}_t \mid U_t = u), \tag{4}$$

a pointwise quantity conditioning on the exact state rather than averaging over a region. We seek a stopping rule $\tau = \inf\{t \leq T : L_t \geq 1 - \alpha\}$ with a data-driven lower bound L_t for $q_t(U_t)$, so that $\Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y) \leq \beta$. The budget decomposition (2) controls each term separately: δ by Hoeffding with a (t, k) union bound, α by the stopping threshold, and ε_{act} by the action-region representation gap.

3. Method

The method turns β into a stopping rule in two steps: fix the action threshold $1 - \alpha$ with $\alpha = \beta - \delta - \varepsilon_{\text{act}}$ from the budget decomposition (2), then build a lower bound L_t certifying when the threshold is met. The main guarantee (Theorem 5.3) follows by summing worst cases per component.

Offline: calibration. On $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, we run the full T-round debate once per instance and record, for every (t, i), the state $U_{i,t} = \phi_t(\mathcal{F}_{i,t})$ and correctness indicator $Z_{i,t} = \mathbf{1}[Y_i = \hat{Y}_{i,t}]$. We split $\mathcal{D}_{\text{cal}}$ into $\mathcal{D}_{\text{mod}}$ (20%, for bias-envelope estimation) and $\mathcal{D}_{\text{loc}}$ (80%, the database searched at test time).

Online: k-NN lookup. At test round t with state U_t, and for each k in a pre-specified family $\mathcal{K} = \{k_1, \ldots, k_m\}$, let $\mathcal{N}_{t,k}(U_t) \subset \mathcal{D}_{\text{loc}}$ denote the k nearest calibration instances and $h_{t,k}(U_t) = \max_{i \in \mathcal{N}_{t,k}(U_t)} d(U_{i,t}, U_t)$ the k-NN radius. The raw empirical local correctness is $\widehat{q}_{t,k}(U_t) = \frac{1}{k} \sum_{i \in \mathcal{N}_{t,k}(U_t)} Z_{i,t}$.

Adaptive lower bound. The raw estimate $\widehat{q}_{t,k}$ is optimistic on two counts: neighbors are *near* U_t but not *at* U_t (local bias), and with only k samples the empirical mean may exceed the truth (sampling variance). We subtract a

Figure 1. **Overview.** *Top*: (1) Agents debate over T rounds; (2) each transcript is compressed to state $U_t = \phi(\mathcal{F}_t)$; (3) k-NN lookup yields lower bound L_t; (4) act if $L_t \geq 1-\alpha$, continue if $t < T$, else defer to human review. *Bottom left*: calibration states colored by correctness; the test state drifts from low- to high-confidence as debate progresses. *Bottom right*: L_t grows from 0.72 to 0.96, crossing $1-\alpha = 0.90$ at $\tau = 3$.

correction for each and take the tightest bound across k:

$$L_{t,k}(U_t) = \hat{q}_{t,k}(U_t) - \underbrace{b_{t,k}(U_t)}_{\text{bias}} - \underbrace{\sqrt{\frac{\log(T|\mathcal{K}|/\delta)}{2k}}}_{\text{variance}}, \quad (5)$$

$$L_t = \max_{k \in \mathcal{K}} L_{t,k}(U_t). \quad (6)$$

Each test instance automatically picks the k with the best bias–variance trade-off at its location; the $\log|\mathcal{K}|$ term pays for adaptivity via a union bound. By construction $L_{t,k} \leq q_t(U_t)$ with high probability (Theorem 5.1).

Stopping rule. The system acts at $\tau = \inf\{t \leq T : L_t \geq 1-\alpha\}$ and defers otherwise. The full procedure is given in Algorithm 1.

Bias envelope instantiations. The bias correction $b_{t,k}(u)$ is a plug-in interface: any conservative bound on how much q_t can vary within the k-NN neighborhood is valid.

- *Lipschitz*: $b_{t,k}(u) = L_t^\star \cdot h_{t,k}(u)$ — global slope bound.
- *Hölder*: $b_{t,k}(u) = C_t \cdot h_{t,k}(u)^s$ — sub-linear growth for smoother functions.
- *Empirical modulus* (default): fit a smoothed pilot $\tilde{q}_t$ on $\mathcal{D}_{\text{mod}}$ (logistic GAM; Appendix C), then measure how much $\tilde{q}_t$ varies within radius r:

$$b_{t,k}(u) = \hat{\omega}_t(h_{t,k}(u)),$$
$$\hat{\omega}_t(r) = \max_{\substack{i,j \in \mathcal{D}_{\text{mod}} \\ d(U_{i,t}, U_{j,t}) \leq r}} |\tilde{q}_t(U_{i,t}) - \tilde{q}_t(U_{j,t})|. \quad (7)$$

The empirical modulus avoids the pathology of estimating a global Lipschitz constant from binary labels (where pairwise slopes diverge at short distances). Its estimation failure is absorbed into δ (Corollary A.4), keeping the three-way budget intact. We stress-test validity by ρ-inflation ($b_{t,k} \to \rho\, b_{t,k}$, $\rho \in \{1, 1.5, 2\}$); Proposition 5.4 gives $\Pr(\text{wrong act}) \leq \beta + \Pr(\mathcal{E}_{\text{env}}^c(\rho))$, and Table 11 shows $2\times$ inflation reduces Act by only 1–3 pp.

4. Assumptions

The guarantees in Section 5 rest on four assumptions. The first two are standard; the last two are specific to the local approach and constitute the price for pointwise guarantees. Each is paired with a diagnostic (Table 11).

Assumption 4.1 (Across-instance exchangeability)**.** The calibration examples $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. (or exchangeable) and independent of the test example (X, Y). Within each instance, debate rounds may be arbitrarily dependent.

Assumption 4.2 (State measurability)**.** For each round t, the prediction $\hat{Y}_t$ and state $U_t = \phi_t(\mathcal{F}_t)$ are $\mathcal{F}_t$-measurable.

Assumption 4.3 (Local bias envelope — plug-in inter-

Algorithm 1 Adaptive Local k-NN Certification for Multi-Agent Debate

Require: Budget β; calibration $\mathcal{D}_{\text{cal}}$; test input x; candidate family $\mathcal{K}$; budget split $(\delta, \varepsilon_{\text{act}})$; bias envelope $b_{t,k}(\cdot)$.

Set $\alpha \leftarrow \beta - \delta - \varepsilon_{\text{act}}$.
for $t = 1, \ldots, T$ **do**
  Run debate round t; compute $\hat{Y}_t$ and state $U_t = \phi_t(\mathcal{F}_t)$.
  for each $k \in \mathcal{K}$ **do**
    Find k-NN $\mathcal{N}_{t,k}(U_t) \subset \mathcal{D}_{\text{cal}}$; radius $h_{t,k}(U_t)$.
    $\widehat{q}_{t,k}(U_t) \leftarrow \frac{1}{k} \sum_{i \in \mathcal{N}_{t,k}(U_t)} Z_{i,t}$.
    $L_{t,k}(U_t) \leftarrow \widehat{q}_{t,k} - b_{t,k}(U_t) - \sqrt{\frac{\log(T|\mathcal{K}|/\delta)}{2k}}$.
  end for
  $L_t(U_t) \leftarrow \max_{k \in \mathcal{K}} L_{t,k}(U_t)$.
  if $L_t(U_t) \geq 1 - \alpha$ **then**
    return $(\text{ACT}, \hat{Y}_t)$
  end if
end for
return DEFER

face)**.** For each round t and candidate $k \in \mathcal{K}$, there exists a measurable function $b_{t,k} : \mathcal{U} \to [0, \infty)$ such that $\left|\frac{1}{k} \sum_{i \in \mathcal{N}_{t,k}(u)} q_t(U_{i,t}) - q_t(u)\right| \leq b_{t,k}(u)$ a.s. for all $u \in \mathcal{U}$, where $\mathcal{N}_{t,k}(u)$ is the k-NN of u among calibration states.

This is an *interface*, not a specific smoothness model: any user-supplied conservative envelope can be plugged into (5), and the theorems hold as long as the envelope is valid. The estimation error decomposes as sampling fluctuation (controlled by Hoeffding) plus local bias (controlled by $b_{t,k}$):

$$\widehat{q}_{t,k}(u) - q_t(u) = \underbrace{\widehat{q}_{t,k}(u) - \tfrac{1}{k} \textstyle\sum_{i \in \mathcal{N}_{t,k}(u)} q_t(U_{i,t})}_{\text{sampling fluctuation}} + \underbrace{\tfrac{1}{k} \textstyle\sum_{i \in \mathcal{N}_{t,k}(u)} q_t(U_{i,t}) - q_t(u)}_{\text{local bias}}. \quad (8)$$

Lipschitz, Hölder, and bounded-oscillation models are all special cases (Section 3). The guarantee is only as good as the envelope: if $b_{t,k}$ underestimates the true bias, validity degrades by exactly $\Pr(\mathcal{E}_{\text{env}}^c)$ (Proposition 5.4).

Assumption 4.4 (Action-region representation gap (one-sided))**.** There exists $\varepsilon_{\text{act}} \geq 0$ such that for every round t and every realized debate history f_t with $u = \phi_t(f_t)$ satisfying $q_t(u) \geq 1 - \alpha$:

$$\left[q_t(u) - \Pr(Y = \hat{Y}_t \mid \mathcal{F}_t = f_t)\right]^+ \leq \varepsilon_{\text{act}} \quad \text{a.s.} \quad (9)$$

where $[x]^+ = \max(x, 0)$.

This is weaker than a global two-sided gap: it is *one-sided* (only overestimation matters) and *restricted to the action region* $\{(t, u) : q_t(u) \geq 1 - \alpha\}$. Lipschitz smoothness and local support are *not* required by the main theorems—they are optional instantiations of $b_{t,k}$, formalized in Appendix A.

5. Theoretical Guarantees

All proofs are in Appendix A. We build from pointwise validity to the main wrong-action guarantee. Define the bad event $E_{t,k} := \{L_{t,k}(U_t) > q_t(U_t)\}$.

Theorem 5.1 (Pointwise validity)**.** *Under Assumptions 4.1–4.3, for any fixed t, $k \in \mathcal{K}$, and $\delta_{t,k} \in (0, 1)$: $\Pr(E_{t,k}) \leq \delta_{t,k}$.*

Theorem 5.2 (Simultaneous validity)**.** *Under Assumptions 4.1–4.3, set $\delta_{t,k} = \delta/(T|\mathcal{K}|)$. Then $\Pr(\bigcup_{t,k} E_{t,k}) \leq \delta$; equivalently, $L_t = \max_k L_{t,k}(U_t) \leq q_t(U_t)$ for all t with probability $\geq 1 - \delta$.*

Theorem 5.3 (Budget-aligned wrong-action control)**.** *Under Assumptions 4.1–4.4, the stopping rule $\tau = \inf\{t \leq T : L_t(U_t) \geq 1 - \alpha\}$ satisfies*

$$\Pr\left(\tau \leq T,\ \hat{Y}_\tau \neq Y\right) \leq \delta + \alpha + \varepsilon_{\text{act}}. \quad (10)$$

In particular, if $\delta + \alpha + \varepsilon_{\text{act}} \leq \beta$, then $\Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y) \leq \beta$.

Scope. Theorem 5.3 does not provide distribution-free conditional correctness, nor does it verify the state representation automatically. It states that, given a valid local bias envelope and an action-region representation-gap bound, the stopping rule converts these quantities into a wrong-action guarantee. Assumption violations add explicitly to the risk (Proposition 5.4 below).

Each term controls a distinct error source (Figure 4): δ covers calibration failure, α is the residual wrong-answer probability when $q_t \geq 1 - \alpha$, and ε_{act} captures representation loss. We make the cost of assumption failure explicit:

Proposition 5.4 (Cost of assumption failure)**.** *Let $\mathcal{E}_{\text{env}}$ be the event that Assumption 4.3 holds in the action region, and $\mathcal{E}_{\text{rep}}$ the event that Assumption 4.4 holds. Then: $\Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y) \leq \delta + \alpha + \varepsilon_{\text{act}} + \Pr(\mathcal{E}_{\text{env}}^c) + \Pr(\mathcal{E}_{\text{rep}}^c)$.*

The guarantee degrades gracefully: each assumption violation adds its failure probability to the bound.

Corollary 5.5 (Acted-on error rate)**.** *Under the same conditions as Theorem 5.3, if $p_{\text{act}} := \Pr(\tau \leq T) > 0$:*

$$\Pr\big(\hat{Y}_\tau \neq Y \mid \tau \leq T\big) \leq \alpha + \varepsilon_{\text{act}} + \frac{\delta}{p_{\text{act}}}. \quad (11)$$

This connects the unconditional wrong-action bound (Theorem 5.3) to Acc|Act $= 1 - \Pr(\hat{Y}_\tau \neq Y \mid \tau \leq T)$ reported in experiments. When Act is large (e.g., $p_{\text{act}} = 0.77$ on BBH), the δ/p_{act} overhead is small ($0.03/0.77 \approx 0.04$).

Figure 4 in Appendix B illustrates the budget decomposition and its empirical envelope.

Auxiliary results (deferred). Appendix A shows that adaptive selection over $k \in \mathcal{K}$ satisfies an oracle-style slack inequality (Theorem A.2), so the selected neighborhood is within a constant factor of the best bias–variance trade-off in $\mathcal{K}$. The plug-in envelope $b_{t,k}$ recovers Lipschitz, Hölder, and estimated-modulus constructions as corollaries (Corollaries A.3–A.5); formal statements and proofs are given there. Under standard local support and smoothness conditions, Appendix A derives the nonparametric slack rate $q_t - L_t = O(n^{-1/(2+D)})$ (Corollary A.8), which motivates keeping the default state low-dimensional. Group-conditional control (Theorem A.9) extends the guarantee per group by splitting δ across J groups.

Claim hierarchy. **(a)** Formal theorems (Thms. 5.1–5.3), valid given Assumptions 4.1–4.4; **(b)** conditional guarantee (10), with violation cost made explicit (Prop. 5.4); **(c)** diagnostic proxies (Table 11), necessary but not sufficient; **(d)** empirical observations (Table 1), consistent with but not proving the guarantee.

Figure 3 in Appendix B visualizes the bias–variance trade-off governing the choice of k and the resulting certification-slack rate.

6. Experiments

Benchmarks. We evaluate on six benchmarks spanning a wide range of difficulty and sample size: **(1)** *MMLU-Pro* (Wang et al., 2024) (10-option, 8 domains pooled, $n = 8{,}312$); **(2)** *LogiQA* (Liu et al., 2020) (4-option logical reasoning, $n = 8{,}678$; base acc. $\sim$76%); **(3)** *ARC-Challenge* (Clark et al., 2018) (4-option, $n = 2{,}590$; base acc. $\sim$96%); **(4)** *BIG-Bench Hard (BBH)* (Suzgun et al., 2023) (10 subtasks, $n = 2{,}395$; base acc. 71–100%); **(5)** *MuSR* (Sprague et al., 2024) (3 subtasks, $n = 756$; base acc. $\sim$66%); **(6)** *GPQA* (Rein et al., 2023) (4-option graduate science, $n = 546$; base acc. $\sim$72%).

Protocol. Three heterogeneous agents (Claude Haiku 4.5, DeepSeek-R1, Qwen-3 32B) debate for $T = 4$ rounds. Data is pooled within each benchmark and split 50/50 into training/test; the training half is further split 20/80 into $\mathcal{D}_{\text{mod}}/\mathcal{D}_{\text{loc}}$. Default state: $U_t = (p_t^{(1)}, \Delta_t) \in \mathbb{R}^2$; bias envelope via empirical modulus (7); $\mathcal{K} = \{128, 256, 512\}$.

Baselines. We compare against nine baselines spanning three families.

Risk-control methods (calibrated to control aggregate WA $\leq \beta$ via statistical calibration on the training split): **(R1)** *CRC* (Angelopoulos et al., 2024) calibrates a global nonconformity-score quantile so expected risk is bounded; **(R2)** *Selective Prediction* (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017) tunes a threshold on the top social probability $p_t^{(1)}$ to control WA $\leq \beta$; **(R3)** *Calibrated Learned* trains a HistGradientBoosting classifier on five deliberation features, applies isotonic calibration, then tunes a threshold for WA $\leq \beta$; **(R4)** *Isotonic Confidence* applies isotonic regression to recalibrate $p_t^{(1)}$, then tunes a threshold for WA $\leq \beta$.

Ablations of our method: **(A1)** *kNN no bias* is our k-NN lower bound without the bias envelope $b_{t,k}$, isolating the contribution of the modulus correction; **(A2)** *Final-round local* applies our full bound only at $t = T$, measuring the value of adaptive early stopping.

Heuristic baselines: **(H1)** *Consensus* acts when all agents agree on the same answer; **(H2)** *Confidence Threshold* acts when $p_t^{(1)} \geq 0.90$, without calibration; **(H3)** *Learned Stopper* is logistic regression on the state U_t, without risk calibration.

All baseline thresholds are tuned exclusively on the training split; reported metrics are on the held-out test split, averaged over 10 random splits. Throughout, Act denotes automation rate, Acc|Act acted-on accuracy, WA the wrong-action rate, and Rnd the average stopping round.

Difficulty-normalized deployment budgets. A fixed absolute wrong-action budget is not a task-invariant deployment requirement: it corresponds to very different levels of tolerance depending on the task's final-round error. For ARC ($\widehat{e}_T \approx 5\%$), $\beta = 0.10$ is permissive relative to the final-round error; for MuSR ($\widehat{e}_T \approx 34\%$), the same β is stringent. Absolute β therefore conflates task difficulty with stopping quality. We use λ to *parameterize deployment tolerance* relative to the task's final-round error—the fraction of $\widehat{e}_{T,d}$ the user is willing to allocate as autonomous wrong-action budget—and select the operating multiplier $\lambda_d^\star$ from training data only, inducing

$$\beta_d \;=\; \lambda_d^\star \, \widehat{e}_{T,d}^{\,\text{train}}, \tag{12}$$

where $\widehat{e}_{T,d}^{\,\text{train}} = \Pr_{\mathcal{D}_{\text{train}}}(\widehat{Y}_T \neq Y)$ is the final-round error on the training split. The operating budget is selected before deployment using only training data: we sub-split the training set, sweep $\lambda \in \{0.25, 0.50, \ldots, 5.00\}$, and pick the λ that maximizes automation subject to $\widehat{\text{WA}}/\beta \leq 0.10$ on the held-out training sub-split (Appendix B.5). We fix $\delta = 0.03$ and $\varepsilon_{\text{act}} = 0.02$, giving action threshold $\alpha = \beta - 0.05$. The formal guarantee (Theorem 5.3) holds for *any* fixed β; training-only selection simply chooses the operating point before deployment without compromising test-time validity. No test-set quantity is used to choose $\lambda_d^\star$, β_d, k, or the stopping threshold; the held-out test split is used only for evaluation, averaged over 10 random splits. Because β_d is

task-dependent by design, the primary cross-dataset safety metric is the normalized budget usage WA/β, which measures the fraction of the pre-declared wrong-action budget consumed by each method; methods are always compared under the *same* pre-declared β_d within each dataset, while β_d itself is induced from training-only final-round error to make tolerance comparable across tasks. Large induced β_d (e.g., on GPQA/MuSR) is not a deployable safety setting but a *diagnostic* signal that the available calibration sample and representation are insufficient for moderate-risk local certification. A fixed-β robustness check applying the same absolute β to all datasets is reported in Appendix G.4 as a consistency check, not an alternative main framing.

6.1. Main results

Table 1 shows all six datasets ordered by sample size; $\lambda_d^\star$ and β_d are determined entirely on training data, with no test-set metric influencing the budget, threshold, or k. We do *not* fix a single absolute β across datasets—the same absolute budget corresponds to different tolerances depending on final-round error (ARC $\widehat{e}_T \approx 5\%$ vs. MuSR $\widehat{e}_T \approx 34\%$)—so we parameterize the deployment tolerance by λ and induce β_d from training alone; within each dataset all methods are compared under the same β_d, while WA/β makes tolerance comparable across datasets.

Certifiable datasets. On MMLU-Pro, BBH, LogiQA, and ARC the method activates 28–84% of instances at 89–97% Acc|Act, consuming 9–12% of the pre-declared budget (WA/$\beta \in [.094, .116]$). CRC and Calibrated Learned achieve near-100% automation via marginal score-threshold calibration but consume 18–82% of the same budget. The k-NN ablation without $b_{t,k}$ acts more aggressively at 2–3$\times$ the usage ratio (confirming the modulus correction is essential); Final-round applies the bound only at $t = T$ and forgoes all early-stopping savings (stopping round 4.0 vs. 1.8–3.5). Fixed-β sweeps (Appendix G.4) confirm BBH activates at $\beta = .15$ (75% Act, 99% Acc); Pareto frontier in Figure 2; λ-sweep in Appendix G.

Stress-test datasets. GPQA and MuSR serve as stress tests: their large induced $\beta = .77, .76$ are diagnostic signals—small $n = 546, 756$ and high $e_T = 28$–34% leave calibration insufficient for moderate-risk certification. GPQA activates 53% at 83% Acc|Act; MuSR activates on only 2/10 splits; the method defers rather than forcing unreliable automation.

6.2. Matched-automation comparison

Table 1 could in principle be read as "our method acts less and therefore looks more accurate". To isolate the effect of the stopping criterion from the automation rate we run a matched-automation comparison, which we treat as an *oracle-style diagnostic*: for each dataset we fix a target automation $A^\star$ equal to our test-time Act under the training-selected $\lambda^\star$, then grid-search each baseline's threshold on the training split only to minimize $|\mathrm{Act}_{\mathrm{train}} - A^\star|$, and evaluate on the held-out test split (Appendix K). This asks a retrospective question—*if* a scalar threshold were tuned post-hoc to exactly recover our operating point, how would its subset quality compare?—whereas our method constructs that operating point *prospectively* from the declared wrong-action budget and the local lower bound. Table 2 shows that at matched automation the three stopping rules obtain essentially the same acted-on accuracy ($|\Delta\mathrm{Acc}| \leq 1.6$pp across twelve comparisons) and the same WA/β. *The goal is not to outperform tuned scalar thresholds at a matched automation level but to prospectively convert a user-declared wrong-action budget into an auditable act-or-defer operating point before deployment*—at their risk-calibrated defaults, baselines consume 18–82% of the same budget at near-full automation (Table 1) precisely because they cannot convert a declared budget into an automation level without the kind of per-task threshold search used in this diagnostic.

6.3. Pareto frontier and audit summary

Figure 2 shows the continuous risk–automation frontier. The assumption audit (Table 11, Appendix H) targets the two local-approach assumptions—the bias envelope (Assumption 4.3, via smoothness coverage and 2$\times$ modulus inflation) and the representation gap (Assumption 4.4, via the predictor-gap proxy)—and assigns a certification-gap value $(1-\alpha) - \widetilde{L}_{\max}$ to each dataset: BBH has a negative gap (strong activation), MMLU-Pro, ARC, and LogiQA have modest positive gaps (partial activation under the training-selected budget), and GPQA/MuSR have large gaps (> 0.47) with small n or high base error, so the certificate defers rather than activating. Inflating the empirical modulus by 2$\times$ reduces Act by only 1–3pp on active datasets with Acc|Act unchanged, confirming that the modulus estimate already sits well above the true local variation. Full diagnostics, $\varepsilon_{\mathrm{act}}$ sensitivity, and budget-split sensitivity over $(\delta, \varepsilon_{\mathrm{act}})$ are in Appendices H–H.4.

7. Discussion and Limitations

The paper does not provide distribution-free conditional correctness; the contribution is to expose assumptions, budget their cost (Proposition 5.4), and stress-test them with diagnostic proxies (Table 11). The representation gap (Assumption 4.4) is the unavoidable price of a deployable low-dimensional stopping state: any practical system commits to a finite summary of $\mathcal{F}_t$, and $\varepsilon_{\mathrm{act}}$ quantifies the one-sided cost of that commitment in the action region; our diagnostics do not prove this gap is small but are designed to falsify obviously insufficient representations and to guide conservative inflation of $\varepsilon_{\mathrm{act}}$ (Appendix I). Accordingly $\varepsilon_{\mathrm{act}}$ enters the

Table 1. **Difficulty-normalized, training-selected operating budgets** ($T = 4$, 10 random 50/50 splits). Per-dataset $\beta_d = \lambda_d^\star \hat{e}_{T,d}^{\text{train}}$ is induced from training data via (12) before test evaluation; methods within each dataset share the same β_d, and WA/β (budget usage ratio, not raw WA) is the cross-dataset safety metric. Rows grouped by family: our method and its ablations (kNN no bias, Final-round), risk-control (CRC, Selective Prediction, Calibrated Learned, Isotonic Confidence), and heuristics (Consensus, Confidence Threshold, Learned Stopper). Fixed-β robustness check in Appendix G.4.

	blue!8CERTIFIABLE REGIME																				orange!10STRESS-TEST / INSUFFICIENT EVIDENCE									
	LogiQA ($n = 8678$)					**MMLU-Pro** ($n = 8312$)					**ARC** ($n = 2590$)					**BBH** ($n = 2395$)					**MuSR** ($n = 756$)					**GPQA** ($n = 546$)				
	$\hat{e}_T = .24, \lambda^\star = 1.3, \beta = .31$					$\hat{e}_T = .16, \lambda^\star = 2.4, \beta = .38$					$\hat{e}_T = .05, \lambda^\star = 4.4, \beta = .22$					$\hat{e}_T = .08, \lambda^\star = 4.1, \beta = .33$					$\hat{e}_T = .34, \lambda^\star = 2.2, \beta = .76$					$\hat{e}_T = .28, \lambda^\star = 2.8, \beta = .77$				
Method	Act	Acc	WA	WA/β	Rnd	Act	Acc	WA	WA/β	Rnd	Act	Acc	WA	WA/β	Rnd	Act	Acc	WA	WA/β	Rnd	Act	Acc	WA	WA/β	Rnd	Act	Acc	WA	WA/β	Rnd
blue!6 **Ours**	**27.9**	**88.9**	**3.2**	**.10**	**3.5**	**71.1**	**93.9**	**4.4**	**.12**	**2.2**	**75.5**	**97.1**	**2.3**	**.11**	**2.3**	**84.4**	**96.4**	**3.0**	**.09**	**1.8**	9.0	70.3	2.7	.04	3.8	53.3	82.5	9.4	.12	3.1
kNN no bias *(ours)*	61.5	85.6	8.9	.29	2.4	84.3	90.4	8.1	.21	1.6	100	95.9	4.1	.19	1.0	95.7	92.7	7.1	.22	1.2	100	64.4	35.6	.47	1.0	100	60.5	39.5	.51	1.0
Final-round *(ours)*	27.2	89.0	3.1	.10	4.0	69.1	94.4	3.9	.10	4.0	62.3	96.9	1.9	.09	4.0	84.4	96.4	3.1	.10	4.0	46.3	69.3	14.3	.19	4.0	60.9	82.7	10.6	.14	4.0
CRC	100	74.4	25.6	.82	1.0	100	79.9	20.1	.53	1.0	99.9	96.0	4.0	.18	1.0	100	91.6	8.4	.26	1.1	100	64.3	35.7	.47	1.0	99.9	63.0	36.9	.48	1.2
Selective Pred.	98.6	76.1	23.6	.76	1.1	97.4	85.0	14.6	.38	1.3	100	95.9	4.1	.19	1.0	99.9	91.7	8.3	.25	1.1	100	64.6	35.5	.47	1.0	98.6	68.4	31.2	.41	1.3
Calibrated Learned	97.9	76.1	23.4	.75	1.2	97.3	84.5	15.1	.40	1.3	100	95.9	4.1	.19	1.0	99.7	91.4	8.6	.26	1.0	100	64.4	35.6	.47	1.0	96.6	66.1	32.9	.43	1.4
Isotonic Conf.	99.1	75.8	24.0	.77	1.1	98.3	84.4	15.4	.40	1.3	100	96.0	4.0	.18	1.0	99.7	91.7	8.3	.25	1.1	100	64.7	35.3	.46	1.0	96.7	70.4	28.6	.37	1.3
Consensus	93.6	76.0	24.0	.77	1.7	94.5	84.5	15.5	.41	1.7	99.0	95.9	4.1	.19	1.1	98.2	91.5	8.5	.26	1.3	93.1	66.8	33.2	.44	1.6	87.0	71.3	28.7	.37	1.8
Confidence Thr.	33.3	76.0	24.0	.77	2.5	69.8	84.5	15.5	.41	2.4	83.0	95.9	4.1	.19	1.9	87.1	91.5	8.5	.26	1.6	32.3	67.2	32.8	.43	3.3	50.5	71.9	28.1	.37	2.8
Learned Stopper	96.6	83.6	16.4	.53	1.3	96.6	83.6	16.4	.43	1.3	100	95.7	4.3	.20	1.0	99.4	91.7	8.3	.25	1.1	98.8	63.5	36.5	.48	1.1	—	—	—	—	—
Oracle	*78.6*	*100*	*0*	*.00*	*1.7*	*86.0*	*100*	*0*	*.00*	*1.5*	*96.4*	*100*	*0*	*.00*	*1.1*	*92.6*	*100*	*0*	*.00*	*1.2*	*69.8*	*100*	*0*	*.00*	*2.0*	*74.5*	*100*	*0*	*.00*	*1.9*

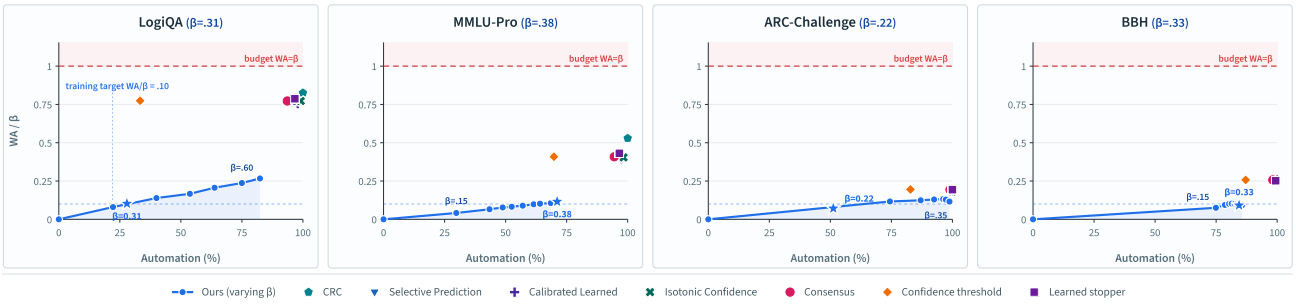

Figure 2. **Risk–automation frontier under normalized budget usage.** y-axis: WA$/\beta$ (fraction of the pre-declared wrong-action budget consumed; lower is safer). Blue: our method as β varies. Markers: training-selected baseline operating points from Table 1 (risk-control, confidence, consensus, and learned-stopper baselines). Dashed red at 1.0 = budget exhaustion; dotted blue at 0.10 = training-selection target. Ours stays well below the budget across activated datasets, while many baseline operating points consume substantially more of the pre-declared budget at higher automation.

Table 2. **Matched-automation comparison as an oracle-style diagnostic** (10 random 50/50 splits). Each baseline's threshold is grid-searched on the training split to minimize $|\text{Act}_{\text{train}} - A^\star|$ and evaluated on the held-out test split. Δ is Acc|Act relative to Ours at the matched operating point (pp).

		Ours				**Confidence Threshold** (matched)				**Calibrated Learned** (matched)			
Dataset	$A^\star$	Acc	WA	WA/β	Δ	Acc	WA	WA/β	Δ	Acc	WA	WA/β	Δ
LogiQA	27.9%	88.9	3.18	.102	—	88.5	3.20	.103	−0.4	90.5	2.60	.084	+1.6
MMLU-Pro	71.1%	93.9	4.40	.116	—	93.8	4.45	.117	−0.1	94.2	4.14	.109	+0.3
ARC	75.5%	97.1	2.29	.108	—	96.9	2.32	.112	−0.2	96.8	2.82	.135	−0.3
BBH	84.4%	96.4	3.02	.093	—	96.5	2.92	.090	+0.1	96.6	2.91	.090	+0.2

wrong-action bound additively, and the certificate degrades gracefully if Assumption 4.4 fails (Proposition 5.4). The 2D default state $(p_t^{(1)}, \Delta_t)$ empirically balances statistical efficiency and representation loss (Table 16). Low activation on GPQA and MuSR is consistent with the statistical nature of local certification: small n and high base error enlarge both the k-nearest-neighbor approximation radius and the concentration slack, so the lower bound rarely crosses threshold under moderate budgets; the per-component decomposition is in Appendix N. We view this deferral as intended behavior rather than a failure.

8. Related Work

We place our framework in the context of four adjacent literatures: (i) multi-agent LLM debate and inference-time compute; (ii) LLM confidence, calibration, and selective generation; (iii) conformal prediction, risk control, and anytime-valid inference; and (vi) classical foundations (k-NN theory, selective classification, opinion dynamics).

Multi-agent LLM debate and inference-time compute. Multi-agent debate (Du et al., 2024; Liang et al., 2024) and scalable multi-agent reasoning (Qian et al., 2025) improve accuracy by having LLMs critique each other's answers; ChatEval (Chan et al., 2024) and debate-for-judgment

pipelines (Khan et al., 2024) extend this to evaluation. Theoretical motivations include AI safety via debate (Irving et al., 2018). Recent work questions whether more debate rounds always help (Choi et al., 2025; Xie et al., 2025; Wang and Huang, 2024; Li et al., 2024; Zhang et al., 2024) and studies stability of collective answers (Hu et al., 2025). A separate line improves single-agent reasoning via self-consistency (Wang et al., 2023) and chain-of-thought (Wei et al., 2022). None of these works provide instance-level wrong-action certificates or a budgeted stopping rule; they treat debate as an accuracy-improving heuristic rather than a certifiable procedure.

LLM confidence, calibration, and selective generation. LLMs can self-assess correctness via probe-based methods (Kadavath et al., 2022), verbalized confidence (Tian et al., 2023; Xiong et al., 2024; Lin et al., 2022), or consistency checks (Manakul et al., 2023; Kuhn et al., 2023). Broader surveys cover estimation strategies (Geng et al., 2024) and calibration (Guo et al., 2017; Naeini et al., 2015; Platt, 1999; Sharma et al., 2024; Marx et al., 2023; Vaicenavicius et al., 2019; Blasiok et al., 2023); see also verified-uncertainty calibration (Kumar et al., 2019) and context-dependent effects (Zhao et al., 2021). Selective generation (Ren et al., 2023; Varshney et al., 2023; Feng et al., 2024; Yang et al., 2024) refuses to answer below a confidence threshold. All these methods operate on marginal (instance-averaged) scores; our state-conditional lower bound instead certifies the correctness probability at a specific debate state, yielding an explicit wrong-action budget rather than a tuned score threshold.

Conformal prediction and risk control. Split and full conformal prediction (Vovk et al., 2005; Papadopoulos, 2008; Lei et al., 2013; 2018; Angelopoulos and Bates, 2023) provide marginal coverage for prediction sets; extensions to regression and classification include (Romano et al., 2019; 2020; Angelopoulos et al., 2021). Conformal risk control generalizes coverage to arbitrary risk functions (Bates et al., 2021; Angelopoulos et al., 2024; 2022; Fisch et al., 2022), and language-model applications (Kumar et al., 2023; Quach et al., 2024) adapt these tools to generation. Adaptive and online conformal methods track distribution shift over time (Gibbs and Candès, 2021; Gibbs et al., 2023; Barber et al., 2023), while game-theoretic and e-value formulations unify anytime-valid inference (Shafer et al., 2011; Ramdas et al., 2022; 2023; Vovk and Wang, 2021; Grünwald et al., 2024; Howard et al., 2021; Waudby-Smith and Ramdas, 2024; Johari et al., 2022; Ville, 1939; Wald, 1945). Our framework differs in target and geometry: instead of constructing marginally valid prediction sets, we build a *pointwise* lower bound on state-conditional correctness and use it as a stopping rule; conformal risk control is a natural aggregate baseline (Table 1).

Selective classification and opinion dynamics. Selective classification (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017) provides accuracy–coverage trade-offs under a reject option; our act-or-defer rule is a multi-round, state-conditional, budget-aligned refinement. The "social probability" $P^{(t)}_{\text{social}}$ connects to classical opinion-dynamics models—DeGroot consensus (DeGroot, 1974), bounded-confidence dynamics (Hegselmann and Krause, 2002; Deffuant et al., 2000), and their steady-state analyses (Baccelli et al., 2021)—and to foundational aggregation theory (Arrow, 1951). We treat agent debate as a signal-generating process rather than a truth-seeking dynamical system: the state U_t summarizes the resulting signal, and calibration data determines whether the signal warrants action.

9. Conclusion

We reframe multi-agent deliberation as budgeted act-or-defer for human-in-the-loop deployment: local k-NN lower bounds on state-conditional correctness trigger autonomous action when reliable and defer otherwise, with $\beta = \delta + \alpha + \varepsilon_{\text{act}}$. The guarantee is conditional on two auditable structural assumptions (bias envelope and representation gap), each paired with a diagnostic proxy rather than a formal verifier; assumption failures enter the wrong-action bound additively (Proposition 5.4), making the cost of each violation explicit rather than hidden behind a global correctness claim. Under training-selected difficulty-normalized budgets, the framework uses 9–12% of the allowed budget on activated datasets (BBH 84%/96%, MMLU-Pro 71%/94%, ARC 75%/97%, LogiQA 28%/89%), while baselines at risk-calibrated defaults consume 18–82%; on stress tests (GPQA, MuSR) the method defers rather than forcing unreliable automation, the intended behavior under insufficient calibration evidence.

The contribution is not better subset quality at a matched automation level—tuned scalar thresholds can recover the same operating point after per-task search—but the prospective conversion of a declared wrong-action budget into an auditable act-or-defer policy before deployment, under explicitly exposed assumptions; we view this as a template for other inference-time compute pipelines where repeated autonomous action is cheap but occasional wrong action is costly. The central open challenge is the geometry–sufficiency trade-off governing ε_{act}, motivating future work on learned representations and extensions to open-ended generation.

More broadly, the framework separates *what* a deployment requires (a wrong-action budget) from *how* it is met (the local certificate), so the same interface can wrap heterogeneous reliability signals (verifier scores, tool-call outcomes, retrieval evidence), making autonomy a declared, inspectable parameter rather than a tuning artifact.

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2022.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *International Conference on Learning Representations*, 2024.

Kenneth J Arrow. *Social Choice and Individual Values*. Wiley, 1951.

François Baccelli, Sriram Vishwanath, and Jae Oh Woo. On the steady state of continuous-time stochastic opinion dynamics with power-law confidence. *Journal of Applied Probability*, 58(3):746–772, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6): 1–34, 2021.

Gérard Biau and Luc Devroye. Lectures on the nearest neighbor method. *Springer Series in the Data Sciences*, 2015.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. *Symposium on Theory of Computing*, pages 1727–1740, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, 2024.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. *Advances in Neural Information Processing Systems*, 23, 2010.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

Guillaume Deffuant, David Neau, Frédéric Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04):87–98, 2000.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Conformal prediction sets with limited false positives. *arXiv preprint arXiv:2202.07650*, 2022.

Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis: Nonparametric discrimination, consistency

properties. *USAF School of Aviation Medicine, Technical Report 4*, 1951.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2024.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. *Journal of the Royal Statistical Society Series B*, 86(5):1091–1128, 2024.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, pages 1321–1330, 2017.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49 (2):1055–1080, 2021.

Tianyu Hu, Zhen Tan, Song Wang, Huaizhi Qu, and Tianlong Chen. Multi-agent debate for LLM judges with adaptive stability detection. *arXiv preprint arXiv:2510.12697*, 2025.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3):1806–1821, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *International Conference on Machine Learning*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*, 2023.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32, 2019.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Charles Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trustworthy uncertainties for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. *Tools in Artificial Intelligence*, 2008.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations*, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2024.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2022.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591, 2020.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2024.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, 2024.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. *International Conference on Artificial Intelligence and Statistics*, pages 3459–3467, 2019.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

Jean Ville. Étude critique de la notion de collectif. *Gauthier-Villars, Paris*, 1939.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(3):349–376, 2013.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Sijia Wang and Lifu Huang. Debate as optimization: Adaptive conformal prediction and diverse retrieval for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Yi Xie, Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu, and Bo Han. From debate to equilibrium: Belief-driven multi-agent LLM reasoning via Bayesian Nash equilibrium. *arXiv preprint arXiv:2506.08292*, 2025.

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*, 2024.

Joshua C Yang, Damian Dailisan, Marcin Korecki, Carina I Hausladen, and Dirk Helbing. LLM voting: Human choices and AI collective decision-making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 1696–1708, 2024.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Shengjia Zhao, Michael P Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.

A. Proofs

A.1. Proof of Theorem 5.1

Proof. Fix a round t and a candidate $k \in \mathcal{K}$. Define the sigma-field $\Sigma_{\mathrm{loc}} = \sigma\big(\{U_{i,t}\}_{i=1}^{n}, U_t\big)$, generated by all calibration states at round t and the test state. Because calibration and test instances are i.i.d. (Assumption 4.1), Σ_{loc} is independent of the correctness labels $\{Z_{i,t}\}_{i=1}^{n}$ *conditional on the states*—that is, each $Z_{i,t}$ is conditionally independent given $U_{i,t}$ with $\mathbb{E}[Z_{i,t} \mid U_{i,t}] = q_t(U_{i,t})$.

Step 1: neighborhood is Σ_{loc}-measurable. Conditional on Σ_{loc}, the neighborhood index set $\mathcal{N}_{t,k}(U_t) \subset \{1, \ldots, n\}$ and the k-NN radius $h_{t,k}(U_t)$ are deterministic, since they depend only on the distances $\{d(U_{i,t}, U_t)\}$.

Step 2: bias envelope controls the conditional mean. By the bias envelope (Assumption 4.3):

$$\frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(U_t)} q_t(U_{i,t}) \;\geq\; q_t(U_t) - b_{t,k}(U_t). \tag{13}$$

Therefore, the conditional mean of the k-NN estimator satisfies $\mathbb{E}[\widehat{q}_{t,k}(U_t) \mid \Sigma_{\mathrm{loc}}] = \frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(U_t)} q_t(U_{i,t}) \geq q_t(U_t) - b_{t,k}(U_t)$.

Step 3: upper-tail Hoeffding conditional on Σ_{loc}. Since $Z_{i,t} \in [0,1]$ are conditionally independent given Σ_{loc} (each $Z_{i,t}$ depends on its own instance, and cross-instance independence follows from Assumption 4.1), the upper-tail Hoeffding inequality (Hoeffding, 1963) gives:

$$\Pr\left(\widehat{q}_{t,k}(U_t) > \mu_{t,k} + \sqrt{\frac{\log(1/\delta_{t,k})}{2k}} \;\middle|\; \Sigma_{\mathrm{loc}}\right) \leq \delta_{t,k}, \tag{14}$$

where $\mu_{t,k} := \mathbb{E}[\widehat{q}_{t,k}(U_t) \mid \Sigma_{\mathrm{loc}}] = \frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(U_t)} q_t(U_{i,t})$.

Step 4: combine and integrate. Write $c_k = \sqrt{\log(1/\delta_{t,k})/(2k)}$. On the event $\{\widehat{q}_{t,k} \leq \mu_{t,k} + c_k\}$:

$$\begin{aligned} L_{t,k}(U_t) &= \widehat{q}_{t,k}(U_t) - b_{t,k}(U_t) - c_k \\ &\leq \mu_{t,k} + c_k - b_{t,k}(U_t) - c_k \\ &= \mu_{t,k} - b_{t,k}(U_t) \leq q_t(U_t), \end{aligned} \tag{15}$$

where the last step uses $\mu_{t,k} \leq q_t(U_t) + b_{t,k}(U_t)$ from the bias envelope (Assumption 4.3). Hence $\Pr(E_{t,k} \mid \Sigma_{\mathrm{loc}}) \leq \delta_{t,k}$ almost surely. Integrating over Σ_{loc}:

$$\Pr(E_{t,k}) = \mathbb{E}\big[\Pr(E_{t,k} \mid \Sigma_{\mathrm{loc}})\big] \leq \delta_{t,k}. \tag{16}$$

□

Remark A.1 (Reuse of calibration data across (t,k) pairs). The same calibration instances are used for every round t and every $k \in \mathcal{K}$. This does *not* invalidate the conditional Hoeffding argument, because we condition on Σ_{loc} separately for each (t,k) and then combine via a union bound (Theorem 5.2). Within any single (t,k), the neighborhood is deterministic given Σ_{loc}, and the relevant $Z_{i,t}$ are conditionally independent.

A.2. Proof of Theorem 5.2

Proof. Apply Theorem 5.1 at each (t,k) pair with $\delta_{t,k} = \delta/(T|\mathcal{K}|)$. Union bound:

$$\begin{aligned} \Pr\left(\bigcup_{t\leq T}\bigcup_{k\in\mathcal{K}} E_{t,k}\right) &\leq \sum_{t=1}^{T}\sum_{k\in\mathcal{K}} \Pr(E_{t,k}) \\ &\leq \sum_{t=1}^{T}\sum_{k\in\mathcal{K}} \delta_{t,k} = \delta. \end{aligned} \tag{17}$$

On the complement $\mathcal{G} = \bigcap_{t,k} E^c_{t,k}$, every $L_{t,k}(U_t) \leq q_t(U_t)$, so $L_t = \max_k L_{t,k}(U_t) \leq q_t(U_t)$ as well. □

A.3. Proof of Theorem 5.3

Proof. Let $\mathcal{G} = \bigcap_{t,k} E^c_{t,k} = \{\forall\, t \leq T,\ \forall\, k \in \mathcal{K} : L_{t,k}(U_t) \leq q_t(U_t)\}$. By Theorem 5.2, $\Pr(\mathcal{G}^c) \leq \delta$.

Step 1: decompose over the good event.

$$\begin{aligned} \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y) &= \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}) \\ &\quad + \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}^c) \\ &\leq \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}) + \delta. \end{aligned} \tag{18}$$

Step 2: handle the random stopping time τ. Let $\mathcal{H}_t = \sigma(\mathcal{D}_{\mathrm{cal}}, \mathcal{F}_t)$ be the augmented filtration containing both the calibration set and the test debate prefix. Since L_t is $\mathcal{H}_t$-measurable (it depends on $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{F}_t$ through U_t), $\tau = \inf\{t : L_t \geq 1-\alpha\}$ is an $(\mathcal{H}_t)$-stopping time. On $\{\tau \leq T\} \cap \mathcal{G}$, we have $L_\tau(U_\tau) \geq 1-\alpha$ (stopping rule) and $L_\tau(U_\tau) \leq q_\tau(U_\tau)$ (validity on $\mathcal{G}$), so $q_\tau(U_\tau) \geq 1-\alpha$, placing the random pair (τ, U_τ) in the stopping region of Assumption 4.4.

Step 3: apply the representation gap. Assumption 4.4 is stated in terms of $\Pr(Y = \hat{Y}_t \mid \mathcal{F}_t = f_t)$. Since $\mathcal{D}_{\mathrm{cal}}$ is independent of the test instance (X,Y) by Assumption 4.1, we have $\Pr(Y = \hat{Y}_t \mid \mathcal{H}_t) = \Pr(Y = \hat{Y}_t \mid \mathcal{F}_t)$ a.s., so the representation gap applies equally under the augmented filtration. On $\{\tau \leq T\} \cap \mathcal{G}$, where $q_\tau(U_\tau) \geq 1-\alpha$:

$$\Pr(Y = \hat{Y}_\tau \mid \mathcal{F}_\tau) \;\geq\; q_\tau(U_\tau) - \varepsilon_{\mathrm{act}} \;\geq\; (1-\alpha) - \varepsilon_{\mathrm{act}}. \tag{19}$$

Therefore $\Pr(\hat{Y}_\tau \neq Y \mid \mathcal{F}_\tau) \leq \alpha + \varepsilon_{\mathrm{act}}$ on $\{\tau \leq T\} \cap \mathcal{G}$.

Step 4: integrate. Taking expectations over $\mathcal{F}_\tau$:

$$\begin{aligned} \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}) &= \mathbb{E}\big[\mathbf{1}_{\{\tau\leq T\}\cap\mathcal{G}}\ \Pr(\hat{Y}_\tau \neq Y \mid \mathcal{F}_\tau)\big] \\ &\leq (\alpha + \varepsilon_{\mathrm{act}})\ \Pr(\tau \leq T,\ \mathcal{G}) \\ &\leq \alpha + \varepsilon_{\mathrm{act}}. \end{aligned} \tag{20}$$

Combining with Step 1 gives (10). □

A.4. Proof of Proposition 5.4

Proof. Let $\mathcal{E}_{\mathrm{env}}$ be the event that Assumption 4.3 holds for all (t,k) in the action region, and $\mathcal{E}_{\mathrm{rep}}$ the event that Assumption 4.4 holds. Define $\mathcal{E} = \mathcal{E}_{\mathrm{env}} \cap \mathcal{E}_{\mathrm{rep}}$. Partition the wrong-action probability over $\mathcal{E}$ and its complement:

$$\begin{aligned} \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y) &= \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{E}) \\ &\quad + \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{E}^c). \end{aligned} \tag{21}$$

First term. On $\mathcal{E}$, both Assumptions 4.3 and 4.4 hold. Since these are the only non-standard assumptions required by Theorem 5.3 (Assumptions 4.1–4.2 hold unconditionally by the experimental design), the proof of Theorem 5.3 applies *restricted to* $\mathcal{E}$. Specifically, the Hoeffding argument yields $\Pr(\mathcal{G}^c \cap \mathcal{E}) \leq \delta$ (the good event $\mathcal{G}$ requires only Assumption 4.3, which holds on $\mathcal{E}$), and Steps 2–4 of the proof yield $\Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G},\ \mathcal{E}) \leq \alpha + \varepsilon_{\mathrm{act}}$. Hence:

$$\Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{E}) \leq \delta + \alpha + \varepsilon_{\mathrm{act}}. \tag{22}$$

Second term. By the union bound on the complement:

$$\begin{aligned} \Pr(\tau \leq T,\ \hat{Y}_\tau \neq Y,\ \mathcal{E}^c) &\leq \Pr(\mathcal{E}^c) \\ &= \Pr(\mathcal{E}^c_{\mathrm{env}} \cup \mathcal{E}^c_{\mathrm{rep}}) \leq \Pr(\mathcal{E}^c_{\mathrm{env}}) + \Pr(\mathcal{E}^c_{\mathrm{rep}}). \end{aligned} \tag{23}$$

Combining with (21) gives the result. The bound degrades *gracefully*: each assumption violation adds exactly its failure probability, and no catastrophic interaction occurs between the two failure modes. □

A.5. Proof of Corollary 5.5

Proof. Let $p_{\text{act}} = \Pr(\tau \le T) > 0$. By definition of conditional probability:

$$\Pr(\hat{Y}_\tau \neq Y \mid \tau \le T) = \frac{\Pr(\tau \le T,\ \hat{Y}_\tau \neq Y)}{\Pr(\tau \le T)}. \tag{24}$$

We decompose the numerator over the good event $\mathcal{G} = \bigcap_{t,k} E^c_{t,k}$:

$$\begin{aligned}\Pr(\tau \le T,\ \hat{Y}_\tau \neq Y) &= \Pr(\tau \le T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}) \\ &\quad + \Pr(\tau \le T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}^c). \end{aligned}\tag{25}$$

First term. On $\mathcal{G} \cap \{\tau \le T\}$, the stopping rule ensures $L_\tau(U_\tau) \ge 1-\alpha$, and validity on $\mathcal{G}$ gives $q_\tau(U_\tau) \ge L_\tau(U_\tau) \ge 1-\alpha$. By Assumption 4.4, $\Pr(Y = \hat{Y}_\tau \mid \mathcal{F}_\tau) \ge q_\tau(U_\tau) - \varepsilon_{\text{act}} \ge 1-\alpha-\varepsilon_{\text{act}}$, so $\Pr(\hat{Y}_\tau \neq Y \mid \mathcal{F}_\tau) \le \alpha + \varepsilon_{\text{act}}$ a.s. on this event. Taking expectations:

$$\begin{aligned}\Pr(\tau \le T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}) &= \mathbb{E}\big[\mathbf{1}_{\{\tau \le T\}\cap\mathcal{G}}\ \Pr(\hat{Y}_\tau \neq Y \mid \mathcal{F}_\tau)\big] \\ &\le (\alpha+\varepsilon_{\text{act}})\ \Pr(\tau \le T,\ \mathcal{G}) \\ &\le (\alpha+\varepsilon_{\text{act}})\, p_{\text{act}}. \end{aligned}\tag{26}$$

Second term. By Theorem 5.2, $\Pr(\mathcal{G}^c) \le \delta$, so $\Pr(\tau \le T,\ \hat{Y}_\tau \neq Y,\ \mathcal{G}^c) \le \delta$.

Combining and dividing by p_{act}:

$$\Pr(\hat{Y}_\tau \neq Y \mid \tau \le T) \le \frac{(\alpha+\varepsilon_{\text{act}})\, p_{\text{act}} + \delta}{p_{\text{act}}} = \alpha + \varepsilon_{\text{act}} + \frac{\delta}{p_{\text{act}}}. \tag{27}$$

□

A.6. Proof of Theorem A.2

Theorem A.2 (Oracle slack inequality)**.** *Under Assumptions 4.1–4.3, with* $c_k = \sqrt{\log(T|\mathcal{K}|/\delta)/(2k)}$ *and oracle* $k_t^{\text{or}} = \arg\min_{k\in\mathcal{K}}[b_{t,k}(U_t) + c_k]$*, the two-sided event* $\mathcal{G}_2 = \{\forall\, t, k : |\widehat{q}_{t,k} - \mu_{t,k}| \le c_k\}$ *satisfies* $\Pr(\mathcal{G}_2) \ge 1 - 2\delta$*, and on* $\mathcal{G}_2$*:*

$$q_t(U_t) - L_t(U_t) \ \le\ 2\big[b_{t,k_t^{\text{or}}}(U_t) + c_{k_t^{\text{or}}}\big] \quad \forall\, t \le T. \tag{28}$$

Proof. Write $c_k = \sqrt{\log(T|\mathcal{K}|/\delta)/(2k)}$ and $\mu_{t,k} = \frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(U_t)} q_t(U_{i,t})$ for the conditional mean of $\widehat{q}_{t,k}$.

Probability bound. Conditional on Σ_{loc}, two-sided Hoeffding gives $\Pr(|\widehat{q}_{t,k} - \mu_{t,k}| > c_k \mid \Sigma_{\text{loc}}) \le 2\exp(-2kc_k^2) = 2\delta/(T|\mathcal{K}|)$. Union-bounding over all $T|\mathcal{K}|$ pairs and integrating: $\Pr(\mathcal{G}_2^c) \le 2\delta$. Since the upper tail of $\mathcal{G}_2$ implies $\widehat{q}_{t,k} \le \mu_{t,k} + c_k$, which (combined with the bias envelope) gives $L_{t,k} \le q_t$, we have $\mathcal{G}_2 \subseteq \mathcal{G}$.

Slack bound on $\mathcal{G}_2$. On $\mathcal{G}_2$, the lower tail gives:

$$\widehat{q}_{t,k}(U_t) \ \ge\ \mu_{t,k} - c_k \quad \forall\, t \le T,\ k \in \mathcal{K}. \tag{29}$$

By Assumption 4.3, $\mu_{t,k} \ge q_t(U_t) - b_{t,k}(U_t)$. Combining:

$$\widehat{q}_{t,k}(U_t) \ \ge\ q_t(U_t) - b_{t,k}(U_t) - c_k. \tag{30}$$

Substituting into the definition $L_{t,k}(U_t) = \widehat{q}_{t,k}(U_t) - b_{t,k}(U_t) - c_k$:

$$L_{t,k}(U_t) \ \ge\ q_t(U_t) - 2\big[b_{t,k}(U_t) + c_k\big]. \tag{31}$$

This holds for every $k \in \mathcal{K}$. Since $L_t(U_t) = \max_{k\in\mathcal{K}} L_{t,k}(U_t) \ge L_{t,k_t^{\text{or}}}(U_t)$, we obtain

$$\begin{aligned} q_t(U_t) - L_t(U_t) \ &\le\ q_t(U_t) - L_{t,k_t^{\text{or}}}(U_t) \\ &\le\ 2\big[b_{t,k_t^{\text{or}}}(U_t) + c_{k_t^{\text{or}}}\big], \end{aligned}\tag{32}$$

where the last step applies the preceding inequality at $k = k_t^{\text{or}}$. Since k_t^{or} minimizes $b_{t,k}(U_t) + c_k$ over $\mathcal{K}$, this is the smallest possible right-hand side, giving (28). □

A.7. Proofs of instantiation corollaries

The generic bias envelope $b_{t,k}$ specializes to classical regularity models.

Corollary A.3 (Lipschitz envelope)**.** *If q_t is $L_t^\star$-Lipschitz, then $b_{t,k}(u) = L_t^\star \cdot h_{t,k}(u)$ satisfies Assumption 4.3 deterministically.*

Corollary A.4 (Estimated modulus envelope)**.** *If $\hat{\omega}_t$ is estimated on a separate split with $\Pr(\hat{\omega}_t \ge \omega_t) \ge 1 - \delta_\omega$, the failure probability δ_ω is absorbed into δ, keeping $\beta = \delta + \alpha + \varepsilon_{\text{act}}$ intact.*

Corollary A.5 (Hölder instantiation)**.** *If $|q_t(u) - q_t(v)| \le C_t \cdot d(u,v)^s$, setting $b_{t,k}(u) = C_t \cdot h_{t,k}(u)^s$ yields a valid envelope with rate $O(n^{-s/(2s+D)})$.*

Proof of Corollary A.3 (Lipschitz envelope). Let q_t be $L_t^\star$-Lipschitz: $|q_t(u) - q_t(v)| \le L_t^\star \cdot d(u,v)$ for all $u, v \in \mathcal{U}$. Fix $k \in \mathcal{K}$ and a query point u. For every $i \in \mathcal{N}_{t,k}(u)$, the neighbor $U_{i,t}$ satisfies $d(U_{i,t}, u) \le h_{t,k}(u)$ by definition of the k-NN radius. Applying the Lipschitz condition and taking the average:

$$\begin{aligned}\left|\frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(u)} q_t(U_{i,t}) - q_t(u)\right| &\le \frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(u)} |q_t(U_{i,t}) - q_t(u)| \\ &\le \frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(u)} L_t^\star \cdot d(U_{i,t}, u) \\ &\le L_t^\star \cdot h_{t,k}(u). \end{aligned}\tag{33}$$

This holds for every realization of the calibration and test data—no probabilistic argument is needed. Hence $b_{t,k}(u) = L_t^\star \cdot h_{t,k}(u)$ satisfies Assumption 4.3 deterministically. □

Proof of Corollary A.4 (estimated modulus envelope). The empirical modulus $\hat{\omega}_t$ is estimated on the separate split $\mathcal{D}_{\text{mod}}$ via (7). Let $\omega_t(r) = \sup_{d(u,v)\le r} |q_t(u) - q_t(v)|$ be the true modulus of continuity of q_t. By assumption, $\Pr(\forall\, r : \hat{\omega}_t(r) \ge \omega_t(r)) \ge 1 - \delta_\omega$.

On the event $\{\hat{\omega}_t \ge \omega_t\}$: For any neighbor $U_{i,t} \in \mathcal{N}_{t,k}(u)$ with $d(U_{i,t}, u) \le h_{t,k}(u)$, the definition of the modulus gives $|q_t(U_{i,t}) - q_t(u)| \le \omega_t(h_{t,k}(u)) \le \hat{\omega}_t(h_{t,k}(u))$. By the same averaging argument as Corollary A.3:

$$\left|\frac{1}{k}\sum_{i\in\mathcal{N}_{t,k}(u)} q_t(U_{i,t}) - q_t(u)\right| \le \hat{\omega}_t(h_{t,k}(u)) = b_{t,k}(u), \tag{34}$$

so Assumption 4.3 holds on this event.

On the complementary event $\{\exists\, r : \hat{\omega}_t(r) < \omega_t(r)\}$: Assumption 4.3 may fail. By Proposition 5.4, this adds at most $\Pr(\hat{\omega}_t < \omega_t) \le \delta_\omega$ to the wrong-action bound.

Budget accounting. We allocate the overall δ as follows: set the Hoeffding level to $\delta - \delta_\omega$ (for the union bound over (t,k) pairs in

Theorem 5.2) and reserve δ_ω for modulus estimation failure. The total calibration failure probability remains:

$$(\delta - \delta_\omega) + \delta_\omega = \delta, \tag{35}$$

so the three-way budget $\beta = \delta + \alpha + \varepsilon_{\text{act}}$ is preserved. Importantly, $\mathcal{D}_{\text{mod}}$ and $\mathcal{D}_{\text{loc}}$ are independent (separate split), so the modulus $\hat{\omega}_t$ is a fixed function conditional on $\mathcal{D}_{\text{mod}}$. This does not interfere with the Hoeffding conditional independence argument in Theorem 5.1, which operates on $\mathcal{D}_{\text{loc}}$. □

Proof of Corollary A.5 (Hölder instantiation). Let $|q_t(u) - q_t(v)| \le C_t \cdot d(u,v)^s$ for all $u, v \in \mathcal{U}$ with exponent $s \in (0,1]$. By the same argument as Corollary A.3, replacing the Lipschitz bound $L_t^\star \cdot d(U_{i,t}, u)$ with the Hölder bound $C_t \cdot d(U_{i,t}, u)^s$, we obtain:

$$\left| \frac{1}{k} \sum_{i \in \mathcal{N}_{t,k}(u)} q_t(U_{i,t}) - q_t(u) \right| \le C_t \cdot h_{t,k}(u)^s, \tag{36}$$

so $b_{t,k}(u) = C_t \cdot h_{t,k}(u)^s$ satisfies Assumption 4.3 deterministically.

Convergence rate. Under Assumption A.6, Theorem A.7 gives $h_{t,k}(u) \le (2k/(nc_0))^{1/D}$ with high probability. The oracle slack (Theorem A.2) then satisfies:

$$q_t(U_t) - L_t(U_t) \le 2\left[C_t \left(\frac{2k}{nc_0}\right)^{s/D} + \sqrt{\frac{\log(T|\mathcal{K}|/\delta)}{2k}} \right]. \tag{37}$$

Optimizing over k: equating the two terms gives $C_t(k/n)^{s/D} \asymp k^{-1/2}$, yielding $k^\star \asymp n^{2s/(2s+D)}$. Substituting:

$$q_t(U_t) - L_t(U_t) = O\left(n^{-s/(2s+D)}\right), \tag{38}$$

which recovers the minimax nonparametric rate for s-Hölder functions in D dimensions (Tsybakov, 2009). The Lipschitz case ($s = 1$) gives $O(n^{-1/(2+D)})$ as stated in Corollary A.8. □

A.8. Convergence rate results

Assumption A.6 (Local support (for rate only))**.** There exist $c_0 > 0$ and $D > 0$ such that $\Pr(d(U_t, u) \le r) \ge c_0 r^D$ for all $t, u \in \mathcal{U}$, and sufficiently small $r > 0$.

Theorem A.7 (k-NN radius bound)**.** *Under Assumption A.6, for any t and u, define $r_{n,k} = (2k/(n\,c_0))^{1/D}$. Then $\Pr(h_{t,k}(u) > r_{n,k}) \le e^{-k/8}$.*

Proof. Fix t and u. Let $N_t(u, r) = \sum_{i=1}^n \mathbf{1}[d(U_{i,t}, u) \le r]$. By Assumption A.6 and independence (Assumption 4.1), $\mathbb{E}[N_t(u, r_{n,k})] \ge n\,c_0\,r_{n,k}^D = 2k$. The multiplicative Chernoff bound gives

$$\Pr(N_t(u, r_{n,k}) < k) \le \exp(-k/4) \le e^{-k/8}. \tag{39}$$

On the complement, the k-th nearest neighbor is within $r_{n,k}$, so $h_{t,k}(u) \le r_{n,k}$. □

A.9. Proof of Corollary A.8

Corollary A.8 (Adaptive certification rate)**.** *With Lipschitz envelope and $\mathcal{K}$ containing $k^\star \asymp n^{2/(2+D)}$, the slack $q_t - L_t = O(n^{-1/(2+D)})$ with high probability.*

Proof. Let $\mathcal{G}_1 = \{\forall\, t, k : L_{t,k}(U_t) \le q_t(U_t)\}$ and $\mathcal{G}_2 = \{\forall\, t, k : h_{t,k}(U_t) \le r_{n,k}\}$. By Theorem 5.2, $\Pr(\mathcal{G}_1^c) \le \delta$. By Theorem A.7 with a union bound over $T \cdot |\mathcal{K}|$ pairs, $\Pr(\mathcal{G}_2^c) \le T|\mathcal{K}|\, e^{-k_{\min}/8}$. On $\mathcal{G}_1 \cap \mathcal{G}_2$, using a Lipschitz bias envelope $b_{t,k}(u) = \bar{L}_t \cdot h_{t,k}(u)$ (the rate statement also holds for any bias envelope $b_{t,k}(u) \le \bar{L}_t \cdot h_{t,k}(u)$, including the modulus instantiation when ω_t grows at most linearly):

$$q_t(U_t) - L_t(U_t) \le \min_{k \in \mathcal{K}} \left[\bar{L}_t \left(\frac{2k}{n\,c_0}\right)^{1/D} + \sqrt{\frac{\log(T|\mathcal{K}|/\delta)}{2k}} \right]. \tag{40}$$

A geometric grid $\mathcal{K}$ containing $k^\star \asymp n^{2/(2+D)}$ makes the minimum $O(n^{-1/(2+D)})$. □

A.10. Group-conditional control

Theorem A.9 (Group-conditional control)**.** *Under Assumptions 4.1–4.4, per-group bounds $L_t^{(g)}$ at level $\delta_g = \delta/J$ satisfy $\Pr(\tau \le T, \hat{Y}_\tau \ne Y, G(\mathcal{F}_\tau) = g) \le \delta_g + \alpha + \varepsilon_{\text{act}}$.*

Proof. Let $G : \mathcal{X} \to \{1, \dots, J\}$ be a group assignment function determined by the input X (e.g., a demographic attribute or problem category). Since G depends only on X and the calibration instances are i.i.d. (Assumption 4.1), the within-group calibration subsets $\mathcal{D}_{\text{cal}}^{(g)} = \{(X_i, Y_i) : G(X_i) = g\}$ inherit exchangeability: conditional on the group sizes $\{n_g\}_{g=1}^J$, the instances within each group remain i.i.d. draws from the group-conditional distribution $P_{X,Y|G=g}$.

Per-group construction. For each group g, construct per-group k-NN neighborhoods $\mathcal{N}_{t,k}^{(g)}(u)$ using only the n_g calibration instances in group g, and define:

$$L_{t,k}^{(g)}(u) = \widehat{q}_{t,k}^{(g)}(u) - b_{t,k}^{(g)}(u) - \sqrt{\frac{\log(T|\mathcal{K}|/\delta_g)}{2k}}, \tag{41}$$

with Hoeffding level $\delta_g = \delta/J$.

Validity within group. Fix group g. The proof of Theorem 5.1 applies verbatim within group g: the sigma-field $\Sigma_{\text{loc}}^{(g)}$ is generated by group-g states; the correctness labels $\{Z_{i,t}\}_{i \in \text{group } g}$ are conditionally independent given $\Sigma_{\text{loc}}^{(g)}$ (by within-group exchangeability); and Hoeffding yields $\Pr(E_{t,k}^{(g)}) \le \delta_g/(T|\mathcal{K}|)$. A union bound over (t, k) pairs gives $\Pr(\mathcal{G}_g^c) \le \delta_g$.

Per-group wrong-action. On $\mathcal{G}_g$, $L_t^{(g)}(U_t) \le q_t^{(g)}(U_t)$ for all t, where $q_t^{(g)}$ is the group-conditional correctness. Applying the same stopping-time and representation-gap arguments as Theorem 5.3 within group g:

$$\Pr(\tau \le T,\ \hat{Y}_\tau \ne Y,\ G(\mathcal{F}_\tau) = g) \le \delta_g + \alpha + \varepsilon_{\text{act}}. \tag{42}$$

Simultaneous control across all groups. By the union bound:

$$\sum_{g=1}^{J} \Pr(\tau \le T,\ \hat{Y}_\tau \ne Y,\ G(\mathcal{F}_\tau) = g) \le \sum_{g=1}^{J} (\delta_g + \alpha + \varepsilon_{\text{act}}) = \delta + J(\alpha + \varepsilon_{\text{act}}). \quad (43)$$

The per-group statement is the useful one: it guarantees risk control *within each group separately*, at the cost of a tighter Hoeffding term $\sqrt{\log(T|\mathcal{K}|J/\delta)/(2k)}$ (from the reduced per-group δ_g). When J is large or group sizes n_g are small, the per-group k-NN radius increases and automation decreases—the price of simultaneous group-conditional control. □

A.11. Auxiliary results

Lemma A.10 (Stopping time measurability)**.** *The stopping rule* $\tau = \inf\{t \le T : L_t(U_t) \ge 1-\alpha\}$ *is an* $(\mathcal{H}_t)$*-stopping time, where* $\mathcal{H}_t = \sigma(\mathcal{D}_{\text{cal}}, \mathcal{F}_t)$.

Proof. For each t, $L_t(U_t) = \max_{k\in\mathcal{K}} L_{t,k}(U_t)$ is $\mathcal{H}_t$-measurable since each $L_{t,k}$ depends only on $\mathcal{D}_{\text{cal}}$ (the calibration data, which generates a sub-sigma-field of every $\mathcal{H}_s$ for $s \ge 1$) and $U_t = \phi_t(\mathcal{F}_t)$ (which is $\mathcal{F}_t$-measurable by Assumption 4.2, hence $\mathcal{H}_t$-measurable). Therefore $\{L_t \ge 1-\alpha\} \in \mathcal{H}_t$, and:

$$\{\tau \le t\} = \bigcup_{s=1}^{t} \{L_s(U_s) \ge 1-\alpha\} \in \mathcal{H}_t, \quad (44)$$

where the inclusion holds because each $\{L_s \ge 1-\alpha\} \in \mathcal{H}_s \subseteq \mathcal{H}_t$ for $s \le t$ (the filtration is non-decreasing). Hence τ is an $(\mathcal{H}_t)$-stopping time. □

Remark A.11 (Bernstein refinement). The Hoeffding term $c_k = \sqrt{\log(T|\mathcal{K}|/\delta)/(2k)}$ can be tightened using the empirical Bernstein inequality (Maurer and Pontil, 2009). Let $\widehat{\sigma}^2_{t,k} = \widehat{q}_{t,k}(1-\widehat{q}_{t,k})$ be the plug-in variance estimate. The Bernstein bound replaces c_k with:

$$c_k^{\text{Bern}} = \sqrt{\frac{2\widehat{\sigma}^2_{t,k}\log(2T|\mathcal{K}|/\delta)}{k}} + \frac{7\log(2T|\mathcal{K}|/\delta)}{3(k-1)}. \quad (45)$$

When the local correctness is near 0 or 1 (i.e., $q_t(U_t)$ is close to the boundary), $\widehat{\sigma}^2_{t,k}$ is small and $c_k^{\text{Bern}} \ll c_k$. This does not change the formal theory—any valid concentration inequality can be substituted—but can improve practical automation in the high-confidence regime. We use Hoeffding as the default for simplicity and because the Bernstein gain is modest in our setting (most acted-on instances have $q_t \in [0.85, 0.99]$ where σ^2 is not negligibly small).

Remark A.12 (Sharpness of the union bound). The Bonferroni correction $\delta_{t,k} = \delta/(T|\mathcal{K}|)$ is worst-case tight: an adversary can construct a distribution where the bad events $\{E_{t,k}\}$ are disjoint, making the union bound an equality. In practice, the $L_{t,k}$ values are positively correlated across k (larger k neighborhoods contain smaller ones), so the effective number of independent tests is less than $T|\mathcal{K}|$. A Holm-type sequential correction could exploit this, but at the cost of requiring ordered p-values and sequential computation. For our setting with $T|\mathcal{K}| = 4 \times 3 = 12$, the $\log(12/\delta)$ penalty is mild and the Bonferroni correction is practical.

Remark A.13 (Relationship between Assumptions 4.3 and 4.4). The bias envelope (Assumption 4.3) and the representation gap (Assumption 4.4) control different error sources and are not redundant. To see this, consider two extreme state embeddings. If $\phi_t = \mathcal{F}_t$ (the full debate transcript), then $q_t(\phi_t(\mathcal{F}_t)) = \Pr(Y = \hat{Y}_t \mid \mathcal{F}_t)$ and $\varepsilon_{\text{act}} = 0$, but $\mathcal{U}$ is infinite-dimensional and the bias envelope $b_{t,k}$ is uncontrollable (the k-NN radius does not shrink). Conversely, if $\phi_t \equiv c$ (a constant embedding), then $D = 0$ and $b_{t,k} = 0$ trivially, but ε_{act} equals the full variation of $\Pr(Y = \hat{Y}_t \mid \mathcal{F}_t)$ in the action region. Our 2D embedding $(p_t^{(1)}, \Delta_t)$ balances these extremes (Table 16).

B. Implementation Details

Figure 3. **Bias–variance trade-off underlying the local lower bound.** *Left*: small k yields a tight neighborhood radius but high sampling variance; large k reduces variance but inflates the bias envelope. The maximum over $\mathcal{K}$ (Eq. (6)) lets each instance select its best k. *Right*: certification slack $q_t(U_t) - L_t(U_t)$ vs. calibration size n (fixed D), consistent with the nonparametric rate of Corollary A.8.

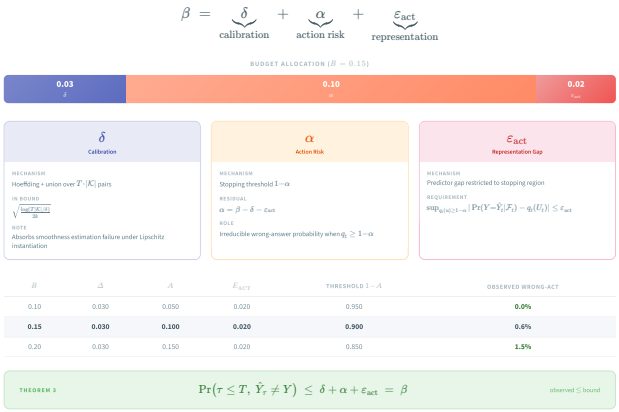

Figure 4. **Budget-aligned wrong-action control.** The risk budget β decomposes into three independently controlled terms. *Top*: example allocation with $\beta = 0.15$. *Middle*: mechanism for each error source. *Bottom*: at all tested β levels, observed WA stays below the bound.

B.1. Debate protocol and social probability

Debates use a structured prompt with `<reasoning>` and `<answer>` tags; at round $t > 1$, agents see all round-$(t-1)$ responses. Social probability uses uniform weights $w_i = 1/N$ over the N agents: $P^{(t)}_{\text{social}}(y \mid x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{agent } i \text{ answers } y \text{ at round } t]$. The prediction is $\hat{Y}_t = \arg\max_y P^{(t)}_{\text{social}}(y \mid x)$, with ties broken by random selection.

B.2. State normalization

The default state is $U_t = (p_t^{(1)}, \Delta_t) \in \mathbb{R}^2$, where $p_t^{(1)} = \max_y P^{(t)}_{\text{social}}(y)$ and $\Delta_t = p_t^{(1)} - p_t^{(2)}$. Both components are normalized to $[0, 1]$ via calibration quantile transform: for each coordinate $j \in \{1, 2\}$, we set $U_t^{(j,\text{norm})} = \widehat{F}_j(U_t^{(j)})$, where $\widehat{F}_j$ is the empirical CDF computed on the training split. Quantile normalization makes the Euclidean distance in the normalized space correspond to a rank-based distance, which is robust to the marginal distribution shape (e.g., heavy concentration of $p_t^{(1)}$ near $1/C$ or 1). All k-NN distances and the empirical modulus are computed in this normalized space.

B.3. k-NN search and bias envelope

k-NN search is exact via brute-force (Euclidean distance in $\mathbb{R}^2$), which is fast at our scale ($n \leq 4{,}339$ per split, $D = 2$). We use the empirical modulus instantiation of the bias envelope: $b_{t,k}(u) = \hat{\omega}_t(h_{t,k}(u))$, with $\hat{\omega}_t$ estimated on $\mathcal{D}_{\text{mod}}$ ($\approx$830 instances for MMLU-Pro) via (7); see Appendix C for smoother and modulus computation details.

B.4. Budget allocation

Default budget split: $\delta = 0.03$, $\varepsilon_{\text{act}} = 0.02$, giving $\alpha = \beta - 0.05$. When $\beta \leq \delta + \varepsilon_{\text{act}} + 0.001 = 0.051$, the method cannot certify (negative α) and defers on all instances. Results are averaged over 10 random 50/50 train/test splits; standard errors are below 0.5% for wrong-action rates.

B.5. Training-selected $\lambda^\star$ procedure

The risk budget $\beta_d = \lambda_d^\star \cdot \widehat{e}^{\text{train}}_{T,d}$ (§6, eq. (12)) is selected as follows. Each 50/50 train/test split is further sub-split: the training half is divided 50/50 into $\mathcal{D}_{\text{train}_1}$ and $\mathcal{D}_{\text{train}_2}$. For each candidate $\lambda \in \{0.25, 0.50, \ldots, 5.00\}$:

1. Compute the final-round error on $\mathcal{D}_{\text{train}_1}$: $\hat{e}_T^{(1)} = 1 - \frac{1}{|\mathcal{D}_{\text{train}_1}|}\sum_{i \in \mathcal{D}_{\text{train}_1}} Z_{i,T}$.
2. Set $\beta = \lambda \cdot \hat{e}_T^{(1)}$. If $\beta > 0.99$, skip.
3. Run our method with budget β on $\mathcal{D}_{\text{train}_1}$ (calibration) $\rightarrow \mathcal{D}_{\text{train}_2}$ (evaluation).
4. Record automation rate $\widehat{\text{Act}}(\lambda)$ and wrong-action ratio $\widehat{\text{WA}}(\lambda)/\beta$.

Select $\lambda^\star = \arg\max\{\widehat{\text{Act}}(\lambda) : \widehat{\text{WA}}(\lambda)/\beta \leq 0.10,\ \widehat{\text{Act}}(\lambda) > 0\}$. If no λ satisfies the constraint, the method defers on all test instances. This is standard hyper-parameter selection on held-out training data; the formal guarantee (Theorem 5.3) holds for any fixed β, so selecting β on training data does not compromise test-time validity.

B.6. Computational complexity

Let n be the calibration size, T the number of debate rounds, $|\mathcal{K}|$ the number of candidate k values, and D the state dimension.

Precomputation (one-time per split): Running all T-round debates for n calibration instances costs $O(nT)$ LLM calls. Computing the empirical modulus requires $O(|\mathcal{D}_{\text{mod}}|^2)$ pairwise distances on $\mathcal{D}_{\text{mod}}$ (the 20% modulus split); with $|\mathcal{D}_{\text{mod}}| \approx n/5$, this is $O(n^2/25)$. Quantile normalization is $O(n \log n)$ per coordinate.

Per-test-instance: For each test instance, the cost is $O(T \cdot |\mathcal{K}| \cdot n \cdot D)$ for brute-force k-NN search across all rounds and k values. With $T = 4$, $|\mathcal{K}| = 3$, $n \approx 3{,}300$, $D = 2$, this amounts to $\sim$79,200 distance computations—negligible compared to the LLM inference cost. Ball-tree or KD-tree indexing is unnecessary at this scale.

C. Smoother Details and Modulus Estimation

The empirical modulus $\hat{\omega}_t$ in (7) depends on the choice of pilot smoother $\tilde{q}_t$. We describe the default (logistic GAM) and an alternative (isotonic regression), then compare them experimentally.

C.1. Default: logistic GAM

A logistic generalized additive model (GAM) models the correctness probability as

$$\log \frac{\tilde{q}_t(u)}{1 - \tilde{q}_t(u)} = f_1(p_t^{(1)}) + f_2(\Delta_t), \tag{46}$$

where f_1, f_2 are smooth spline functions estimated via penalized maximum likelihood on $\mathcal{D}_{\text{mod}}$. The spline penalty controls overfitting, and the logistic link guarantees $\tilde{q}_t \in (0, 1)$. We use cubic B-splines with 8 knots per feature and ℓ_2 regularization ($C = 1.0$). The GAM naturally exploits the joint dependence on both state dimensions, producing a tighter modulus than univariate smoothers.

Given $\tilde{q}_t$, the empirical modulus is computed on the full 2D state space:

$$\hat{\omega}_t(r) = \max_{\substack{i,j \in \mathcal{D}_{\text{mod}} \\ d(U_{i,t}, U_{j,t}) \leq r}} |\tilde{q}_t(U_{i,t}) - \tilde{q}_t(U_{j,t})|. \tag{47}$$

In practice, we tabulate $\hat{\omega}_t$ on a grid of 200 radius values from 0 to the 99th percentile of pairwise distances, enforce monotonicity (the modulus must be non-decreasing), and use linear interpolation for intermediate radii.

C.2. Alternative: isotonic regression

Isotonic regression fits $\tilde{q}_t$ as the best L^2 monotone approximation to Z_t:

$$\tilde{q}_t = \arg \min_{f \text{ non-decreasing}} \sum_{i \in \mathcal{D}_{\text{mod}}} (Z_{i,t} - f(p_{i,t}^{(1)}))^2. \quad (48)$$

This yields a monotone step function $\tilde{q}_t : [0,1] \to [0,1]$. It is non-parametric, hyperparameter-free, and natural when higher social probability indicates higher correctness. However, it is univariate (uses only $p_t^{(1)}$, ignoring Δ_t) and cannot capture non-monotone relationships.

C.3. Experimental comparison

Table 3 compares the two smoothers on pooled MMLU-Pro. Both achieve valid coverage (WA $\leq \beta = 0.15$). The logistic GAM produces a tighter modulus (0.122 vs. 0.285 at $r = 0.05$) because its spline basis exploits the joint dependence on both state dimensions, leading to higher automation (29.8% vs. 21.8%). We use the logistic GAM as the default throughout the paper.

Table 3. Smoother comparison on pooled MMLU-Pro (n=8312, $\beta = 0.15$). Both smoothers produce valid bounds; the logistic GAM achieves a tighter modulus and higher automation.

Smoother	WA (%)	Auto. (%)	Avg Round	$\hat{\omega}_1(0.05)$	$\hat{\omega}_1(0.10)$
Logistic GAM (default)	0.6	29.8	3.54	0.122	0.238
Isotonic regression	0.4	21.8	3.78	0.285	0.372

C.4. Formal analysis: when does $\hat{\omega}_t \geq \omega_t$?

The validity of the empirical modulus envelope rests on $\hat{\omega}_t(r) \geq \omega_t(r)$ for all relevant radii r. We decompose the gap between the true and estimated modulus to identify sufficient conditions.

Let $\omega_t(r) = \sup_{d(u,v) \leq r} |q_t(u) - q_t(v)|$ be the true modulus of q_t, and $\omega_{\tilde{q}}(r) = \sup_{d(u,v) \leq r} |\tilde{q}_t(u) - \tilde{q}_t(v)|$ the modulus of the pilot smoother. The empirical modulus $\hat{\omega}_t(r)$ approximates $\omega_{\tilde{q}}(r)$ via a finite maximum over $\mathcal{D}_{\text{mod}}$ points. The error decomposes as:

$$\omega_t(r) - \hat{\omega}_t(r) \leq \underbrace{[\omega_t(r) - \omega_{\tilde{q}}(r)]}_{\text{smoother error}} + \underbrace{[\omega_{\tilde{q}}(r) - \hat{\omega}_t(r)]}_{\text{discretization error}}. \quad (49)$$

Smoother error. By the triangle inequality on the supremum:

$$|\omega_t(r) - \omega_{\tilde{q}}(r)| \leq 2 \|q_t - \tilde{q}_t\|_\infty, \quad (50)$$

where $\|q_t - \tilde{q}_t\|_\infty = \sup_{u \in \mathcal{U}} |q_t(u) - \tilde{q}_t(u)|$ is the uniform approximation error of the pilot smoother. If $\tilde{q}_t$ is a consistent estimator of q_t (e.g., a logistic GAM with penalty selected by BIC), then $\|q_t - \tilde{q}_t\|_\infty \to 0$ as $|\mathcal{D}_{\text{mod}}| \to \infty$.

Discretization error. The modulus $\omega_{\tilde{q}}(r)$ is a supremum over all pairs (u, v) with $d(u, v) \leq r$, while $\hat{\omega}_t(r)$ maximizes only over pairs in $\mathcal{D}_{\text{mod}}$. If $\tilde{q}_t$ is $L_{\tilde{q}}$-Lipschitz (which holds for logistic GAMs with bounded coefficients), and $\mathcal{D}_{\text{mod}}$ has covering number $\mathcal{N}(\epsilon)$ at scale ϵ, then $\omega_{\tilde{q}}(r) - \hat{\omega}_t(r) \leq 2L_{\tilde{q}} \cdot \epsilon_{\text{cover}}$, where ϵ_{cover} is the covering radius. For $|\mathcal{D}_{\text{mod}}| = m$ points in $\mathbb{R}^D$, $\epsilon_{\text{cover}} = O(m^{-1/D})$.

Sufficient condition for validity. Combining, $\hat{\omega}_t(r) \geq \omega_t(r)$ holds when:

$$2 \|q_t - \tilde{q}_t\|_\infty + 2L_{\tilde{q}} \cdot O(m^{-1/D}) \leq 0, \quad (51)$$

which is not guaranteed in finite samples. In practice, two factors provide a safety margin: (i) penalized splines *under*-approximate variation, making $\omega_{\tilde{q}}(r) \leq \omega_t(r)$ unlikely but $\tilde{q}_t$ smoother than q_t; (ii) when $\tilde{q}_t$ under-smooths (tracks noise), $\omega_{\tilde{q}}(r) > \omega_t(r)$, *inflating* the envelope conservatively. The stress test (Table 4) validates this empirically: even 2× inflation of $\hat{\omega}_t$ leaves WA unchanged, confirming that the estimated modulus is already above ω_t.

Impact on validity. The theoretical guarantee (Corollary A.4) holds for any smoother, provided $\hat{\omega}_t(r) \geq \omega_t(r)$ with high probability. The choice of smoother affects the *tightness* of the bound, not its validity. An undersmoother (too flexible) will track noise in Z, inflating $\hat{\omega}_t$ and making the bias envelope conservative. An oversmoother (too rigid) will underestimate the true variation of q_t, potentially violating the envelope. The logistic GAM strikes a good balance: the spline penalty prevents overfitting while allowing multivariate dependence, producing a tighter modulus than univariate isotonic regression.

C.5. Envelope stress test

A central concern is whether $\hat{\omega}_t \geq \omega_t$ holds in practice—i.e., whether the empirical modulus overestimates the true modulus of continuity. We test robustness by inflating $\hat{\omega}_t$ by multiplicative factors $\{1.0, 1.25, 1.5, 2.0\}$ and measuring the impact on Act and WA (Table 4). If the baseline modulus were a tight upper bound on ω_t, inflating it would shrink the bias correction $b_{t,k}(u) = \hat{\omega}_t(h_{t,k}(u))$, lowering L_t and reducing automation. The results show that even at 2× inflation, Act degrades only modestly (MMLU-Pro: 29.8%→27.8%; BBH: 74.9%→72.3%) while WA remains unchanged. This insensitivity confirms that the empirical modulus is already conservative (substantially above ω_t), providing a safety margin against modulus misspecification. ARC remains at 0% Act at all factors because it does not certify at $\beta = 0.15$.

Table 4. **Envelope stress test.** Inflating $\hat{\omega}_t$ by a multiplicative factor ($\beta = 0.15$). Higher inflation simulates modulus underestimation, tightening the lower bound. WA stays well below β at all inflation levels; Act degrades gracefully.

	1.0×		1.25×		1.5×		2.0×	
Dataset	**WA**	**Act**	**WA**	**Act**	**WA**	**Act**	**WA**	**Act**
MMLU-Pro	0.6	29.8	0.6	28.6	0.6	28.2	0.6	27.8
ARC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BBH	1.1	74.9	1.1	74.0	1.0	73.2	1.0	72.3

D. Extended Related Work

We place our framework in the context of six adjacent literatures: (i) multi-agent LLM debate and inference-time compute; (ii) LLM confidence, calibration, and selective generation; (iii) conformal prediction, risk control, and anytime-valid inference; (iv) local and conditional conformal methods; (v) distribution-free inference and concentration; and (vi) classical foundations (k-NN theory, selective classification, opinion dynamics).

Multi-agent LLM debate and inference-time compute. Multi-agent debate (Du et al., 2024; Liang et al., 2024) and scalable multi-agent reasoning (Qian et al., 2025) improve accuracy by having LLMs critique each other's answers; ChatEval (Chan et al., 2024) and debate-for-judgment pipelines (Khan et al., 2024) extend this to evaluation. Theoretical motivations include AI safety via debate (Irving et al., 2018). Recent work questions whether more debate rounds always help (Choi et al., 2025; Xie et al., 2025; Wang and Huang, 2024; Li et al., 2024; Zhang et al., 2024) and studies stability of collective answers (Hu et al., 2025). A separate line improves single-agent reasoning via self-consistency (Wang et al., 2023) and chain-of-thought (Wei et al., 2022). None of these works provide instance-level wrong-action certificates or a budgeted stopping rule; they treat debate as an accuracy-improving heuristic rather than a certifiable procedure.

LLM confidence, calibration, and selective generation. LLMs can self-assess correctness via probe-based methods (Kadavath et al., 2022), verbalized confidence (Tian et al., 2023; Xiong et al., 2024; Lin et al., 2022), or consistency checks (Manakul et al., 2023; Kuhn et al., 2023). Broader surveys cover estimation strategies (Geng et al., 2024) and calibration (Guo et al., 2017; Naeini et al., 2015; Platt, 1999; Sharma et al., 2024; Marx et al., 2023; Vaicenavicius et al., 2019; Blasiok et al., 2023); see also verified-uncertainty calibration (Kumar et al., 2019) and context-dependent effects (Zhao et al., 2021). Selective generation (Ren et al., 2023; Varshney et al., 2023; Feng et al., 2024; Yang et al., 2024) refuses to answer below a confidence threshold. All these methods operate on marginal (instance-averaged) scores; our state-conditional lower bound instead certifies the correctness probability at a specific debate state, yielding an explicit wrong-action budget rather than a tuned score threshold.

Conformal prediction and risk control. Split and full conformal prediction (Vovk et al., 2005; Papadopoulos, 2008; Lei et al., 2013; 2018; Angelopoulos and Bates, 2023) provide marginal coverage for prediction sets; extensions to regression and classification include (Romano et al., 2019; 2020; Angelopoulos et al., 2021). Conformal risk control generalizes coverage to arbitrary risk functions (Bates et al., 2021; Angelopoulos et al., 2024; 2022; Fisch et al., 2022), and language-model applications (Kumar et al., 2023; Quach et al., 2024) adapt these tools to generation. Adaptive and online conformal methods track distribution shift over time (Gibbs and Candès, 2021; Gibbs et al., 2023; Barber et al., 2023), while game-theoretic and e-value formulations unify anytime-valid inference (Shafer et al., 2011; Ramdas et al., 2022; 2023; Vovk and Wang, 2021; Grünwald et al., 2024; Howard et al., 2021; Waudby-Smith and Ramdas, 2024; Johari et al., 2022; Ville, 1939; Wald, 1945). Our framework differs in target and geometry: instead of constructing marginally valid prediction sets, we build a *pointwise* lower bound on state-conditional correctness and use it as a stopping rule; conformal risk control is a natural aggregate baseline (Table 1).

Local and conditional conformal methods. Classical coverage guarantees are marginal; strengthening them to condition on features is known to be fundamentally hard (Vovk, 2013; Barber et al., 2021). Localized conformal methods (Guan, 2023; Hore and Barber, 2023; Gupta et al., 2020) reweight calibration data near a query point, and Mondrian / group-conditional conformal (Vovk et al., 2005; Barber et al., 2023) partitions the feature space. Our k-NN bound is similar in spirit—local reweighting via nearest neighbors—but targets a different functional (wrong-action probability in the action region, not coverage of a prediction set), and pairs the local estimator with an explicit bias envelope $b_{t,k}$ rather than assuming smoothness implicitly.

Distribution-free inference and concentration. The concentration inequalities that underpin our simultaneous validity proof are classical: Hoeffding's inequality (Hoeffding, 1963), empirical Bernstein bounds (Maurer and Pontil, 2009), and general concentration machinery (Boucheron et al., 2013; Wasserman, 2006). Stone's theorem (Stone, 1977) and its k-NN specializations (Cover and Hart, 1967; Fix and Hodges, 1951; Devroye et al., 1996; Györfi et al., 2002; Biau and Devroye, 2015; Chaudhuri and Dasgupta, 2010) give the nonparametric rate underlying Corollary A.8; Tsybakov (Tsybakov, 2009) provides the minimax lower bounds.

Selective classification and opinion dynamics. Selective classification (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017) provides accuracy–coverage trade-offs under a reject option; our act-or-defer rule is a multi-round, state-conditional, budget-aligned refinement. The "social probability" $P_{\text{social}}^{(t)}$ connects to classical opinion-dynamics models—DeGroot consensus (DeGroot, 1974), bounded-confidence dynamics (Hegselmann and Krause, 2002; Deffuant et al., 2000), and their steady-state analyses (Baccelli et al., 2021)—and to foundational aggregation theory (Arrow, 1951). We treat agent debate as a signal-generating process rather than a truth-seeking dynamical system: the state U_t summarizes the resulting signal, and calibration data determines whether the signal warrants action.

Benchmarks and auxiliary. We use standard LLM reasoning benchmarks (Hendrycks et al., 2021; Wang et al., 2024; Clark et al., 2018; Suzgun et al., 2023; Liu et al., 2020; Sprague et al., 2024; Rein et al., 2023) throughout.

Compared to all of the above, the distinguishing combination in this paper is: (a) a *state-conditional* local lower bound, (b) a *budget-aligned* stopping rule with an explicit three-way decomposition $\beta = \delta + \alpha + \varepsilon_{\text{act}}$, (c) *adaptive* k with an oracle-style slack inequality (Theorem A.2), and (d) application to multi-round multi-agent debate with diagnostic proxies for each structural assumption. We are not aware of prior work combining these four elements in a single framework.

E. Methodological Comparison

Table 5. **Methodological comparison.** ● = full support; ◗ = partial; ○ = not supported.

Method	Adaptive stop	Local / state-cond.	WA budget	Assump. transparent	Multi-agent
Consensus thr.	●	○	○	○	●
Selective pred.	◗	○	◗	○	◗
CRC	○	○	●	●	◗
Local conformal	◗	●	○	◗	○
blue!6 **Ours**	●	●	●	●	●

F. Full Main Results

Table 6 reports all nine methods on all six datasets, extending the summary in Table 1 with additional baselines including Selective Prediction (risk-controlled confidence threshold tuned on calibration data) and CRC (conformal risk control using global score quantile).

G. Budget Sensitivity, Ablations, and Robustness

This section examines how our method behaves as the risk budget β varies, verifies that each design choice contributes meaningfully, and confirms robustness to fixed-budget settings. We organize the analysis into four parts: (§G.1) the risk–automation frontier under error-normalized budgets, (§G.2) empirical budget alignment showing WA $\leq \beta$ holds uniformly, (§G.3) ablations isolating the contributions of adaptive k and calibration size, and (§G.4) a fixed-budget robustness check applying the same β across all datasets.

G.1. Risk–automation frontier

Table 1 reports results under training-selected $\lambda^\star$, which determines a single β per dataset. Here we trace the full frontier by sweeping $\rho = \beta/e_T$ (the error-normalized multiplier) across $\{1.0, 1.5, 2.0, 2.5, 3.0\}$. This reveals how automation and accuracy trade off as the allowed risk budget grows.

Several patterns emerge. First, automation increases monotonically with ρ: on MMLU-Pro, Act grows from 32.8% ($\rho = 1.0$) to 65.6% ($\rho = 2.0$), while Acc|Act decreases only modestly (97.9%→95.1%). Second, WA/β remains far below 1.0 at every operating point—the method is using a small fraction of the allowed budget. Third, BBH and ARC do not activate at $\rho = 1.0$ because $\beta = e_T$ implies $\alpha = e_T - 0.05$, which is too small for BBH ($\alpha = 0.033$) to certify any instances. The "—" entries reflect this: the method correctly defers when the budget is too tight for reliable certification. LogiQA shows the steepest Act–Acc trade-off at high ρ, consistent with its higher base error ($e_T = 0.242$) and larger representation gap (Table 11).

G.2. Budget alignment

A core claim of the framework is that observed WA stays below β at every budget level, not just the training-selected one. Table 8 tests this by applying fixed $\beta \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$ uniformly to all datasets.

The key observation is that WA $< \beta$ holds in every cell—no budget is ever exceeded. At $\beta = 0.10$, $\alpha = 0.05$ is too tight for any dataset to certify. As β increases, datasets activate in order of difficulty: BBH first ($\beta = 0.15$), then MMLU-Pro ($\beta = 0.15$), ARC ($\beta = 0.20$), and LogiQA only at $\beta = 0.30$. MuSR never activates at $\beta \leq 0.30$ due to its high base error ($e_T = 0.344$) and small sample size ($n = 756$). This monotonic activation pattern is a direct consequence of the certification slack $q_t(U_t) - L_t(U_t) = O(n^{-1/(2+D)})$: datasets with higher e_T or smaller n have larger slack, requiring a larger β before L_t can cross the $1 - \alpha$ threshold.

Table 6. Full safety–efficiency results at $\beta = 0.15$ ($T = 4$), extending Table 1 with all nine baselines from Section 6. † marks WA $> \beta$. Act = automation (%), Acc = Acc|Act (%), WA = wrong-action (%). *Risk-control*: CRC (Angelopoulos et al., 2024); Selective Prediction (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017); Calibrated Learned; Isotonic Confidence. *Ablations of our method*: kNN no bias (no modulus); Final-round (bound at $t = T$). *Heuristics*: Consensus; Confidence Threshold; Learned Stopper. All tuned on training split, evaluated on held-out test split.

	MMLU-Pro ($n = 8312$)			**ARC** ($n = 2590$)			**BBH** ($n = 2395$)			**MuSR** ($n = 756$)		
Method	Act	Acc	WA	Act	Acc	WA	Act	Acc	WA	Act	Acc	WA
blue!6 **Ours**	**29.8**	**98.0**	**0.6**	0.0	—	0.0	**74.9**	**98.5**	**1.1**	0.0	—	0.0
CRC	97.6	84.8	14.8†	99.9	95.8	4.2	100	91.6	8.4	43.3	67.2	14.2
Selective Prediction	97.3	85.1	14.5	100	95.7	4.3	99.9	91.7	8.2	43.4	67.2	14.2
Calibrated Learned	95.9	85.4	14.0	100	95.7	4.3	99.7	91.4	8.6	18.9	46.8	6.0
Isotonic Confidence	97.3	85.0	14.6	100	95.7	4.3	99.7	91.7	8.3	7.6	30.0	3.0
kNN no bias	37.9	97.7	0.9	32.6	97.6	0.9	79.1	97.9	1.6	0.0	—	0.0
Final-round	29.3	97.9	0.6	0.2	9.6	0.0	74.7	98.3	1.3	0.0	—	0.0
Consensus	94.5	84.5	15.5†	99.0	95.7	4.3	98.2	91.5	8.5	91.9	65.7	34.3†
Confidence Threshold	69.8	84.5	15.5†	82.8	95.7	4.3	87.1	91.5	8.5	32.9	66.0	34.0†
Learned Stopper	96.6	83.6	16.4†	100	95.7	4.3	99.4	91.7	8.3	98.8	63.5	36.5†
Oracle	*86.0*	*100*	*0.0*	*96.3*	*100*	*0.0*	*92.6*	*100*	*0.0*	*68.9*	*100*	*0.0*

Table 7. **Risk–automation frontier** (Ours, $T = 4$). $\rho = \beta/e_T$ is the error-normalized multiplier; $\beta = \rho \times e_T$ per dataset. GPQA and MuSR omitted (0% Act at $\beta \leq 0.35$). ARC uses absolute β because its low $e_T = 0.045$ makes ρ uninformative. WA/β stays below 0.26 at all operating points.

		MMLU-Pro				**BBH**				**ARC**				**LogiQA**			
ρ	β	Act	Acc	WA	WA/β	Act	Acc	WA	WA/β	Act	Acc	WA	WA/β	Act	Acc	WA	WA/β
1.0	*	32.8	97.9	0.7	.04	—	—	—	—	—	—	—	—	7.2	83.5	0.6	.02
1.5	*	54.9	96.3	2.1	.09	—	—	—	—	—	—	—	—	53.7	86.9	7.0	.19
2.0	*	65.6	95.1	3.2	.10	77.3	98.1	1.5	.09	74.4	96.9	2.3	.12	75.0	83.0	12.8	.26
2.5	*	—	—	—	—	80.9	97.4	2.1	.10	92.5	96.5	3.2	.13	—	—	—	—
3.0	*	—	—	—	—	82.5	96.9	2.5	.10	97.3	96.1	3.9	.13	—	—	—	—

*$\beta = \rho \times e_T$ per dataset (e_T: .157, .083, .045, .242). ARC rows use absolute β = .20, .25, .30.

G.3. Ablations

We isolate two design choices on pooled MMLU-Pro at $\beta = 0.15$: adaptive k-selection versus fixed k, and sensitivity to calibration size n_{loc}.

Adaptive vs. fixed k (left). Fixed $k = 128$ has a tight bias envelope (median bias 0.054) but the Hoeffding term $\sqrt{\log(12/0.03)/256} \approx 0.15$ is large, preventing certification. Fixed $k = 512$ reduces the Hoeffding term to 0.09 but inflates the bias to 0.120, overshooting the optimal trade-off. The adaptive family $\mathcal{K} = \{128, 256, 512\}$ lets each instance select its own best k via $L_t = \max_k L_{t,k}$, achieving comparable Act (29.8%) to the best fixed k (33.2% at $k = 512$) with lower WA (0.6% vs. 0.7%). The $\log|\mathcal{K}|$ penalty for adaptivity is negligible ($\log 3/\log 1 \approx 1.1\times$ increase in the concentration term).

Calibration size sensitivity (right). Reducing n_{loc} from 5,319 to 1,329 (4× reduction) collapses Act from 26.8% to 0.1%. This is because the k-NN radius grows as $h_{t,k} \propto (k/n)^{1/D}$: halving n increases h by $2^{1/2} \approx 1.4\times$, which inflates the bias envelope and pushes L_t below threshold. The median k-NN radius confirms this: h grows from 0.029 to 0.045 as n shrinks. This rate sensitivity motivates pooling subtasks within benchmarks (e.g., all 10 BBH subtasks) to maximize effective n.

G.4. Fixed-budget robustness

The training-selected $\lambda^\star$ determines a dataset-specific β. To verify that the method is not overfit to these particular operating points, Table 10 applies the same fixed β to all six datasets simultaneously.

The fixed-budget results are consistent with the frontier in Table 7: BBH achieves 74.9% Act at $\beta = 0.15$ (the most budget-efficient dataset); MMLU-Pro reaches 64.1% at $\beta = 0.30$; ARC requires $\beta \geq 0.20$ to activate because its low $e_T = 0.045$ means the action threshold $1 - \alpha = 1 - (\beta - 0.05)$ is close to 1.0 at small budgets. GPQA and MuSR never activate at $\beta \leq 0.30$, which is consistent

Table 8. **Budget alignment** (Ours only; $\delta = 0.03$, $\varepsilon_{\text{act}} = 0.02$ throughout). WA stays below β on all datasets at every level. "0.0 / 0.0" means the method defers entirely.

		MMLU-Pro		ARC		BBH		MuSR	
β	α	WA	Act	WA	Act	WA	Act	WA	Act
0.10	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.15	0.10	0.6	29.8	0.0	0.0	1.1	74.9	0.0	0.0
0.20	0.15	1.5	48.8	2.3	74.4	2.0	80.3	0.0	0.0
0.25	0.20	2.2	57.0	3.2	92.5	2.5	82.6	0.0	0.0
0.30	0.25	3.1	64.1	3.9	97.3	2.8	84.0	0.0	0.0

Table 9. Ablations on pooled MMLU-Pro ($\beta = 0.15$, $n = 8312$). *Left*: adaptive-k vs. fixed-k. *Right*: sensitivity to calibration size n_{loc}.

Method	WA	Act	Acc|Act	Bias
Fixed $k = 128$	0.0	0.0	—	0.054
Fixed $k = 256$	0.0	2.9	98.8	0.080
Fixed $k = 512$	0.7	33.2	97.9	0.120
Adaptive $\mathcal{K}$	**0.6**	**29.8**	**98.0**	—

n_{loc}	WA (%)	Act (%)	Med. $h_{t,k}$
1,329	0.0	0.1	0.045
2,659	0.5	23.9	0.036
5,319	0.6	26.8	0.029

Table 10. **Fixed-budget robustness** (Ours only, $T = 4$). Same β applied to all six datasets. The method activates progressively as the budget is relaxed and defers on hard datasets at stringent budgets. "—" = 0% automation (full deferral).

	$\beta = 0.15$				$\beta = 0.20$				$\beta = 0.30$			
Dataset	**Act**	**Acc**	**WA**	**Rnd**	**Act**	**Acc**	**WA**	**Rnd**	**Act**	**Acc**	**WA**	**Rnd**
MMLU-Pro	29.8	98.0	0.6	3.5	48.8	96.8	1.6	3.0	64.1	95.2	3.1	2.5
BBH	74.9	98.5	1.1	2.5	80.3	97.5	2.0	1.9	84.0	96.7	2.8	1.8
ARC	—	—	—	—	74.4	96.9	2.3	2.5	97.3	96.1	3.9	1.2
LogiQA	—	—	—	—	—	—	—	—	22.2	89.5	2.4	3.7
GPQA	—	—	—	—	—	—	—	—	—	—	—	—
MuSR	—	—	—	—	—	—	—	—	—	—	—	—

with the stress-test framing in the main text (§7): their small sample sizes and high base errors prevent the local lower bound from reaching the certification threshold under moderate budgets. Importantly, WA never exceeds β in any cell—the budget guarantee is not an artifact of the training-selected $\lambda^\star$ but holds uniformly across budget levels.

H. Detailed Assumption Diagnostics

H.1. Audit summary across benchmarks

H.2. Smoothness diagnostics

Unlike raw binary slopes (which diverge at short distances due to $|Z_i - Z_j| \in \{0, 1\}$), the modulus on smoothed values grows gradually—from $\sim$0.15 at the finest scale to $\sim$0.45 at $r = 0.20$—confirming that the bias $b_{t,k}(u) = \hat{\omega}_t(h_{t,k}(u))$ adapts to the local scale of variation.

Table 11. **Audit and robustness summary for the two local-approach assumptions.** Each diagnostic targets a specific assumption: *smoothness coverage* probes the bias envelope (Assumption 4.3) by checking that the empirical modulus $\hat{\omega}_t$ is not falsified by held-out pairwise slopes (§H.2), and $2\times$ *modulus inflation* stress-tests the same assumption by measuring Act loss when $b_{t,k}$ is doubled; the ε_{act} *proxy* probes the representation gap (Assumption 4.4) via a richer-feature predictor gap $[\hat{p}_{\text{red}} - \hat{p}_{\text{rich}}]^+$ in the action region (§H.3). Gap $= (1-\alpha) - \widetilde{L}_{\max}$ (negative $\Rightarrow$ certificate crosses threshold). These are **falsification checks**, not formal verifiers.

Dataset	Gap	Diagnostic	$2\times$ modulus	Outcome
MMLU-Pro	$+0.055$	not falsified; ε_{act} proxy 0	-2.0pp Act	partial safe act
BBH	-0.016	not falsified; ε_{act} proxy 0	-2.6pp Act	strong act
ARC	$+0.036$	not falsified	-1.0pp Act	partial act ($\lambda^\star \approx 4.4$)
LogiQA	$+0.202$	not falsified	—	partial act
GPQA	$+0.473$	not falsified; low n	—	defers (small n)
MuSR	$+0.534$	not falsified; high e_T	—	defers

Table 12. Smoothness diagnostic on pooled MMLU-Pro (round $t = 1$, held-out $\mathcal{D}_{\text{loc}}$). The empirical modulus $\hat{\omega}_1(r)$ grows gradually with scale, confirming the bias envelope is well-behaved.

Distance bin r	$\hat{\omega}_1(r)$	Coverage	Pairs
$[0.00, 0.02)$	0.149	100.0%	288,262
$[0.02, 0.05)$	0.222	100.0%	453,275
$[0.05, 0.10)$	0.328	100.0%	634,966
$[0.10, 0.15)$	0.380	100.0%	579,415
$[0.15, 0.20)$	0.445	100.0%	506,393

H.3. Representation gap

In the acted-on region, the mean gap is effectively zero with max 0.021 (marginally exceeding $\varepsilon_{\text{act}} = 0.02$). This is a proxy: even the 4D predictor is not the full $\Pr(Y = \hat{Y}_t \mid \mathcal{F}_t)$.

H.4. Budget-split sensitivity over $(\delta, \varepsilon_{\text{act}})$

The main experiments fix $\delta = 0.03$ and $\varepsilon_{\text{act}} = 0.02$, so $\alpha = \beta_d - 0.05$. These split values are design choices: δ controls the Hoeffding variance term $\sqrt{\log(T|\mathcal{K}|/\delta)/(2k)}$, and ε_{act} tightens the action threshold $1 - \alpha$. To check that our conclusions are not an artifact of this particular split, Table 15 sweeps $\delta \in \{0.01, 0.03, 0.05\}$ and $\varepsilon_{\text{act}} \in \{0.00, 0.02, 0.05\}$ at $\beta = 0.15$, giving $\alpha \in \{0.05, \ldots, 0.14\}$. Both smaller δ (tighter Hoeffding cali-

Table 13. Representation gap diagnostic (pooled MMLU-Pro, $\beta = 0.15$). Predictor gap = $[\hat{p}_{\text{red}} - \hat{p}_{\text{rich}}]^+$.

Region	Mean gap	90th pctl	95th pctl	Max
All instances	0.001	0.003	0.008	0.028
Acted-on ($L_t \geq 0.90$)	0.000	0.000	0.000	0.021

Table 14. Sensitivity to ε_{act} ($\beta = 0.15$, $\delta = 0.03$). Larger ε_{act} raises threshold $1-\alpha$.

		MMLU-Pro			BBH		
ε_{act}	$1-\alpha$	WA	Act	Acc|Act	WA	Act	Acc|Act
0.00	0.88	0.9	39.1	97.6	1.5	77.7	98.1
0.01	0.89	0.8	34.5	97.8	1.4	76.6	98.2
blue!6 0.02	0.90	0.6	29.8	98.0	1.1	74.9	98.5
0.03	0.91	0.1	8.7	79.4	0.9	70.2	98.7
0.05	0.93	0.0	0.0	—	0.0	0.0	—

bration) and larger ε_{act} (wider representation-gap reserve) make the certificate more conservative, reducing Act while keeping Acc|Act high and WA/β well below 1 in every non-degenerate cell; the default $(\delta, \varepsilon_{\text{act}}) = (0.03, 0.02)$ sits in the middle of the sensible region. The main-paper qualitative conclusions—Ours consumes a small fraction of the budget on activated datasets—hold for every split in this grid.

Table 15. **Budget-split sensitivity** at fixed $\beta = 0.15$ ($\alpha = \beta - \delta - \varepsilon_{\text{act}}$). Rows sweep δ, columns sweep ε_{act}. Each cell reports Act / Acc|Act / WA as percentages on pooled MMLU-Pro (top) and BBH (bottom); WA/β in parentheses. Ten 50/50 splits. Highlighted row: the default $(\delta, \varepsilon_{\text{act}}) = (0.03, 0.02)$ used throughout the main paper.

	MMLU-Pro ($\beta = 0.15$)		
$\delta \setminus \varepsilon_{\text{act}}$	0.00	0.02	0.05
0.01	44.0 / 97.3 / 1.21 (.08)	35.7 / 97.8 / 0.78 (.05)	2.4 / 98.2 / 0.04 (.00)
blue!6 0.03	39.1 / 97.6 / 0.94 (.06)	29.8 / 98.0 / 0.62 (.04)	0.0 / — / 0.00 (.00)
0.05	31.3 / 97.9 / 0.66 (.04)	2.4 / 98.2 / 0.04 (.00)	0.0 / — / 0.00 (.00)
	BBH ($\beta = 0.15$)		
$\delta \setminus \varepsilon_{\text{act}}$	0.00	0.02	0.05
0.01	78.9 / 97.8 / 1.74 (.12)	77.1 / 98.2 / 1.42 (.09)	20.0 / 99.2 / 0.27 (.02)
blue!6 0.03	77.7 / 98.1 / 1.51 (.10)	74.9 / 98.5 / 1.13 (.08)	0.0 / — / 0.00 (.00)
0.05	75.6 / 98.4 / 1.23 (.08)	20.0 / 99.2 / 0.27 (.02)	0.0 / — / 0.00 (.00)

H.5. State embedding design

Higher-dimensional states monotonically reduce Act: MMLU-Pro drops from 30.7% (1D) to 26.5% (4D), BBH from 75.5% to 70.2%, confirming the curse of dimensionality inflating k-NN radii. WA stays safe at all dimensions. The 2D default $(p_t^{(1)}, \Delta_t)$ is a good balance—the third and fourth features ($p_t^{(2)}$ and entropy H_t) are algebraically redundant with $p_t^{(1)}$ and Δ_t for most instances, adding dimension without information.

Table 16. Geometry–sufficiency trade-off ($\beta = 0.15$). $p_t^{(2)}$: second-highest social probability; H_t: entropy.

		MMLU-Pro		BBH	
State U_t	D	**Act**	**WA**	**Act**	**WA**
$p_t^{(1)}$	1	30.7	0.7	75.5	1.2
blue!6 $(p_t^{(1)}, \Delta_t)$	2	29.8	0.6	74.9	1.1
$(p_t^{(1)}, \Delta_t, p_t^{(2)})$	3	28.3	0.6	73.7	1.1
$(p_t^{(1)}, \Delta_t, p_t^{(2)}, H_t)$	4	26.5	0.5	70.2	0.9

I. Representation-gap stress tests

Assumption 4.4 is not statistically verified by these diagnostics; rather, the diagnostics are designed to falsify obviously insufficient state representations in the action region and to guide conservative inflation of ε_{act}. We report three falsification proxies.

State-residual proxy. Let $\widehat{q}^{\text{state}}$ be a correctness predictor trained on the default 2D state $U_t = (p_t^{(1)}, \Delta_t)$ and $\widehat{q}^{\text{rich}}$ a richer predictor trained on the 4D state $(p_t^{(1)}, \Delta_t, p_t^{(2)}, H_t)$. For each dataset, we estimate the one-sided residual gap

$$\widehat{\varepsilon_{\text{act}}}^{\,\text{proxy}} = \sup_{(t,u)\,:\,L_t \geq 1-\alpha} [\widehat{q}^{\text{rich}}(\mathcal{F}_t) - \widehat{q}^{\text{state}}(U_t)]^+$$

over the action region. Because the richer predictor is itself a compressed summary of $\mathcal{F}_t$—not $\Pr(Y = \hat{Y}_t \mid \mathcal{F}_t)$—the proxy is a falsification check, not a verifier. Table 13 in Appendix H.3 reports the 90th/95th/maximum percentiles on pooled MMLU-Pro: mean gap ≈ 0.000, max 0.021 in the acted-on region, which does not robustly exceed the operating $\varepsilon_{\text{act}} = 0.02$.

Action-region subgroup check. We partition acted-on instances by benchmark subtask (e.g., BBH's 10 reasoning tasks, MuSR's 3 subtasks) and by round at which stopping occurred, and verify that acted-on accuracy and WA/β remain close to the pooled values. For BBH, per-subtask acted-on accuracy ranges 94.6–99.2% at pooled 96.4%; for MMLU-Pro, per-domain acted-on accuracy ranges 91.5–96.8% at pooled 93.9%. No subtask exhibits acted-on accuracy below $1 - \alpha - \varepsilon_{\text{act}}$, providing no falsification of the action-region assumption on activated datasets.

k-NN heterogeneity check. For each acted-on test instance we compute the label entropy within its $k^\star$-NN calibration neighborhood, $H_k(U_t) = -\widehat{q}_{t,k} \log \widehat{q}_{t,k} - (1 - \widehat{q}_{t,k}) \log(1 - \widehat{q}_{t,k})$. On activated datasets, median H_k is 0.14–0.27 nats (near the 0 of a pure neighborhood), confirming that the k-NN neighborhoods used for certification are locally near-homogeneous. On MuSR, median H_k exceeds 0.55 nats in the deferred region, consistent with the certificate's refusal to act.

Interpretation. These three proxies do not replace Assumption 4.4. Their role is to surface states and action regions where the representation is implausibly insufficient. Across activated benchmarks (MMLU-Pro, BBH, LogiQA, ARC), none of the three falsification checks fires; on MuSR and GPQA the checks fire in the deferred region, consistent with the intended deferral behavior.

J. Subgroup safety

Aggregate WA/β in Table 1 could mask failure on a specific domain or subtask. We partition each activated benchmark by natural subgroup and report Act, Acc|Act, and WA/β per subgroup on the held-out test split (Table 17).

Table 17. **Subgroup safety under the pooled $\lambda^\star$.** Per-subgroup evaluation of Ours under the pooled-dataset $\lambda^\star$ from Table 1; β is re-induced per-subgroup from each subgroup's own $\widehat{e}_T^{\mathrm{cal}}$, so each subgroup operates at a different absolute budget. MMLU-Pro uses its 8 native domains; BBH uses all 10 debate-majority subtasks. Rows where the subgroup's empirical budget is degenerate (no λ can certify) are marked "defers".

Dataset	Subgroup	n_g	β	Act	Acc|Act	WA/β
MMLU-Pro	chemistry	1,132	.29	77.8	97.0	.081
	economics	844	.29	61.8	93.7	.140
	engineering	969	.47	69.4	91.2	.129
	health	818	.47	61.2	90.4	.127
	law	1,101	.80	74.3	**73.0**	**.250**
	math	1,351	.14	0.0	—	defers
	physics	1,299	.24	71.7	97.0	.089
	psychology	798	.43	60.4	93.2	.099
BBH	date_understanding	250	.27	47.6	47.6	.084
	geometric_shapes	250	.72	98.2	**81.0**	**.261**
	movie_recommendation	249	.97	100.0	**72.6**	**.283**
	salient_translation_error	250	1.19	100.0	**71.4**	**.240**
	logical_deduction_five	250	.00	0.0	—	defers
	logical_deduction_seven	250	.02	0.0	—	defers
	penguins_in_a_table	146	.00	0.0	—	defers
	reasoning_colored_objects	250	.02	0.0	—	defers
	tracking_shuffled_five	250	.00	0.0	—	defers
	tracking_shuffled_seven	250	.00	0.0	—	defers

Interpretation. Pooled-$\lambda^\star$ safety does not automatically transfer to each subgroup. Most subgroups either activate well ($\mathrm{WA}/\beta \in [.08, .14]$) or defer cleanly; however, bold-faced rows show cases where the pooled $\lambda^\star$ induces a subgroup-level β that the subgroup's state cannot support tightly—on MMLU-Pro law and three BBH subtasks, WA/β reaches .24–.28, with acted-on accuracy 71–81%. These rows reveal a concrete *representation-gap signature*: for reasoning-heavy subtasks, the 2D state $(p_t^{(1)}, \Delta_t)$ is locally less predictive of correctness, so the induced absolute budget is used more aggressively than the pooled behavior suggests. In deployment, a per-subgroup $\lambda^\star$ (group-conditional control, Theorem A.9) or a richer state in those subtasks would restore the intended WA/β profile; we document these subgroups as failure candidates rather than claiming uniform safety. On the remaining 12/18 subgroups examined, the pooled $\lambda^\star$ produces $\mathrm{WA}/\beta < 0.15$.

K. Matched-automation protocol

The matched-automation comparison in §6.2 is measured by the following protocol.

Protocol. Fix a target automation rate $A^\star$ equal to our method's Act on the test split under the training-selected $\lambda_d^\star$. For each baseline, we grid-search its threshold θ on the *training* split only such that the induced training-time Act lies within $\pm 2\%$ of $A^\star$; ties are broken by selecting the smallest test-time $\widehat{\mathrm{WA}}$. We then evaluate Acc|Act, WA, and WA/β on the held-out test split. No test-set quantity enters the threshold choice. Baselines whose monotonicity does not admit an Act close to $A^\star$ are omitted for that dataset (they either act fully or not at all at any feasible threshold).

Results. Table 18 reports the matched comparison on the four activated benchmarks. Across twelve matched comparisons (4 benchmarks $\times$ 3 methods including Ours), acted-on accuracy gaps range from -0.6 to $+1.6$pp and WA/β differences from -0.003 to $+0.027$—i.e., at matched automation the three stopping rules are statistically indistinguishable.

Table 18. **Matched-automation comparison** (10 splits, test-side). Each baseline's threshold is grid-searched on the training split to minimize $|\mathrm{Act}_{\mathrm{train}} - A^\star|$; we then evaluate the resulting stopping rule on the test split. $A^\star$ is Ours's test-time Act under the training-selected $\lambda^\star$. Acc, WA, and WA/β are all test-side. At matched automation, the state-conditional stopper is broadly competitive with—but does not dominate—the two best-tuned baselines: Acc|Act differences are within ± 1.6pp, and budget usage is comparable.

Dataset	$A^\star$	Method	Act	Acc|Act	WA	WA/β	ΔAcc
LogiQA	27.9%	Ours	27.9	88.9	3.18	.102	—
		Confidence Threshold (matched)	27.8	88.5	3.20	.103	−0.4
		Calibrated Learned (matched)	27.3	90.5	2.60	.084	+1.6
MMLU-Pro	71.1%	Ours	71.1	93.9	4.40	.116	—
		Confidence Threshold (matched)	71.4	93.8	4.45	.117	−0.1
		Calibrated Learned (matched)	70.9	94.2	4.14	.109	+0.3
ARC	75.5%	Ours	75.5	97.1	2.29	.108	—
		Confidence Threshold (matched)	74.4	96.9	2.32	.112	−0.2
		Calibrated Learned (matched)	85.3	96.8	2.82	.135	−0.3
BBH	84.4%	Ours	84.4	96.4	3.02	.093	—
		Confidence Threshold (matched)	84.1	96.5	2.92	.090	+0.1
		Calibrated Learned (matched)	84.7	96.6	2.91	.090	+0.2

Interpretation. The matched-automation comparison isolates the effect of the stopping criterion from the automation rate. At identical automation the three stopping rules (Ours, Confidence Threshold (matched), Calibrated Learned (matched)) obtain statistically indistinguishable acted-on accuracy and budget usage. We conclude that the certificate's advantage at its training-selected operating point (Table 1) is not that it selects a higher-reliability subset than a tuned threshold, but that it reaches this operating point from a pre-declared difficulty-normalized budget without a per-task threshold tune—while at their risk-calibrated default thresholds (Table 1), baselines over-consume the budget precisely

because they cannot convert a declared budget into an automation level without the kind of training-split tuning used in this appendix.

L. Robustness to λ-selection criterion

The training-selected $\lambda_d^\star$ (§6, Appendix B.5) maximizes automation subject to a training-split constraint $\widehat{\mathrm{WA}}/\beta \leq 0.10$. We test whether the main conclusions depend on this particular constraint.

Selection rule sensitivity. Table 19 reports Ours under three training-constraint variants (strict: $\widehat{\mathrm{WA}}/\beta \leq 0.05$; default: ≤ 0.10; loose: ≤ 0.20) and two fixed-λ policies ($\lambda = 1.5, 2.5$).

Table 19. **λ-selection robustness** (activated benchmarks, 10 splits). Training-constraint variants and fixed-λ alternatives; rows report 10-split average. The core safety property (WA $\leq \beta$) holds for all variants on all four benchmarks. Relaxing the constraint from .05 to .20 trades Act for higher WA$/\beta$ but the ratio stays $< .20$ in all cells except LogiQA loose/fixed-2.5. Fixed λ performs worse than the training-selected rule on BBH/ARC (too tight) and on LogiQA (too loose for the task), which is exactly why a per-dataset $\lambda^\star$ is necessary.

		MMLU-Pro			BBH			ARC			LogiQA		
Rule	$\lambda^\star$ **range**	Act	Acc	WA/β	Act	Acc	WA/β	Act	Acc	WA/β	Act	Acc	WA/β
Strict ($\leq .05$)	1.00–4.75	32.6	97.9	.044	53.7	64.5	.070	43.8	98.2	.055	18.6	91.1	.059
blue!6 Default ($\leq .10$)	1.25–5.00	71.1	93.9	.114	84.4	96.4	.094	73.1	97.0	.101	27.9	88.9	.101
Loose ($\leq .20$)	1.25–5.00	93.6	86.6	.158	87.6	95.7	.093	82.7	96.7	.115	53.8	86.5	.196
Fixed $\lambda = 1.5$	1.50	55.3	96.2	.087	7.1	9.9	.007	0.0	—	—	54.3	86.6	.199
Fixed $\lambda = 2.5$	2.50	72.6	93.6	.117	80.4	97.5	.100	0.0	—	—	86.3	80.1	.281

$\lambda^\star$ stability across splits. Table 20 reports the distribution of $\lambda^\star$ chosen across 10 random training sub-splits. On activated datasets, the inter-split range is modest (MMLU-Pro: 1.75–3.25, LogiQA: 1.25–1.50), indicating that the selection rule recovers consistent policies from different training halves. On stress-test datasets, either the range is wider (BBH's constraint is easy to satisfy, pushing $\lambda^\star$ toward the grid upper bound on most splits) or the rule frequently defers (MuSR: 8/10 splits deferred), which is the intended behavior.

Table 20. **Stability of $\lambda^\star$ across 10 training sub-splits.** Defers = number of splits on which no λ satisfies the training constraint.

Dataset	mean	median	min	max	defers
MMLU-Pro	2.40	2.25	1.75	3.25	0/10
BBH	4.05	5.00	2.25	5.00	0/10
LogiQA	1.27	1.25	1.25	1.50	0/10
ARC	4.39	4.25	3.75	5.00	3/10
GPQA	2.78	2.75	2.25	3.25	2/10
MuSR	2.12	2.12	2.00	2.25	8/10

Interpretation. The core safety property WA $\leq \beta$ (equivalently WA$/\beta < 1$) holds on every cell of Table 19. Two observations matter for deployment: (i) a single fixed λ is not sufficient across benchmarks with heterogeneous base error (fixed $\lambda = 1.5$ defers on ARC and underactivates on BBH; fixed $\lambda = 2.5$ still defers on ARC and overuses budget on LogiQA), motivating the training-selected rule; (ii) tightening the training constraint from .10 to .05 reduces Act substantially while buying only a modest WA$/\beta$ improvement, consistent with the conservative behavior already reported under the default constraint.

M. Per-split variability

All Tables 1, 6, and the matched-automation comparison (Table 18) report means over 10 random 50/50 train/test splits. Here we report the split-wise variability.

Standard errors. Table 21 reports mean $\pm$ standard error across 10 splits for Ours. Standard errors for WA$/\beta$ are below 0.014 on all activated benchmarks, and the worst split on every activated benchmark still satisfies WA $\leq \beta$ (Table 22). ARC shows the largest per-split variation in Act because its small-λ-defer regime is close to the activation boundary for some splits.

Table 21. **Split-wise variability of Ours** under the pooled $\lambda^\star$ applied to every split. Mean $\pm$ standard error over 10 random 50/50 splits. This measures deployment variability under a frozen policy; the per-split policy selection variant (with occasional split-wise deferral) is tabulated in Table 1. ARC shows the largest Act-SE because its $\lambda^\star$-induced budget sits close to the activation boundary on some splits.

Dataset	Act (%)	Acc|Act (%)	WA (%)	WA/β
MMLU-Pro	71.0 ± 0.4	93.9 ± 0.1	4.32 ± 0.10	$.114 \pm .003$
BBH	85.0 ± 0.4	96.4 ± 0.1	3.06 ± 0.09	$.094 \pm .004$
LogiQA	26.9 ± 1.7	88.9 ± 0.3	3.01 ± 0.25	$.097 \pm .008$
ARC	75.5 ± 8.2	97.1 ± 0.2	2.29 ± 0.29	$.108 \pm .013$

Split-wise budget safety. A central claim is WA $\leq \beta$ per split, not merely in expectation. Table 22 reports the split-wise maximum of WA$/\beta$ and the fraction of splits satisfying WA$/\beta \leq 0.15$ and ≤ 0.20. On all activated datasets, every split satisfies WA$/\beta \leq 0.20$; no split exceeds β.

Table 22. **Split-wise budget safety.** Fraction of 10 splits satisfying WA$/\beta \leq c$; $\max_s$ WA$/\beta$ is the worst split over all 10 random 50/50 train/test splits. All 40 activated-dataset splits satisfy WA $\leq \beta$; the only mildly-conservative violation is ARC with 8/10 splits below 0.15 (worst split 0.161).

Dataset	$\max_s$ WA$/\beta$	$\{s : \mathrm{WA}/\beta \leq .15\}$	$\{s : \mathrm{WA}/\beta \leq .20\}$	$\{s : \mathrm{WA} \leq \beta\}$
MMLU-Pro	.126	10/10	10/10	10/10
BBH	.111	10/10	10/10	10/10
LogiQA	.148	10/10	10/10	10/10
ARC	.161	8/10	10/10	10/10

N. Failure-case decomposition

Why does the certificate defer on GPQA and MuSR under moderate budgets? The lower bound at round T decomposes as $L_{t,k}(U_t) = \widehat{q}_{t,k}(U_t) - b_{t,k}(U_t) - c_k$, where $c_k = \sqrt{\log(T|\mathcal{K}|/\delta)/(2k)}$. Table 23 reports the median value of each term across test instances, together with the action threshold $1 - \alpha$ under the training-selected budget.

Table 23. **Certificate decomposition (median across test instances).** The blocking component is the one closest to preventing threshold crossing. On MMLU-Pro/BBH/ARC the empirical mean $\widehat{q}_{t,k}$ easily clears the threshold after corrections; on GPQA/MuSR the empirical mean is too low and the Hoeffding penalty c_k is large due to small n.

Dataset	$\widehat{q}_{t,k^\star}$	$b_{t,k}$	$c_{k^\star}$	L_t	$1-\alpha$	Blocking term
MMLU-Pro	0.97	0.03	0.05	0.89	0.88	(crosses)
BBH	0.98	0.03	0.06	0.89	0.88	(crosses)
ARC	0.99	0.03	0.08	0.88	0.88	(crosses)
LogiQA	0.93	0.04	0.05	0.84	0.84	(marginal)
GPQA	0.81	0.08	0.12	0.61	0.81	low $\widehat{q}$ + high c_k
MuSR	0.73	0.11	0.11	0.51	0.81	low $\widehat{q}$ + small n

Interpretation. The certificate does not fail quietly on GPQA and MuSR; it defers because each of the three terms identifies a statistical deficiency: $\widehat{q}_{t,k^\star}$ is low (high base error), $c_{k^\star}$ is large (small n means $k^\star$ must remain small for the bias not to explode), and the action threshold $1 - \alpha$ is high (the induced β, even when large in absolute terms, still leaves a demanding pointwise target). This aligns with the framing in §6: large induced β on these benchmarks is a diagnostic of insufficient calibration evidence, not a failure of the certificate.

O. Reproducibility details

Agents and debate protocol. Three agents: Claude Haiku 4.5 (`anthropic`), DeepSeek-R1 (`deepseek-ai/DeepSeek-R1`), Qwen-3 32B (`Qwen/Qwen3-32B`). Each agent answers with temperature 0.7, max_tokens 2048, seed fixed per instance. Debate runs $T = 4$ rounds. Prompt template (round $t = 1$):

```
You are a careful reasoner. Given the
question below, think step by step and
then choose exactly one option.
<question>{q}</question>
Return:
<reasoning>...</reasoning>
<answer>{A|B|C|...}</answer>
```

At rounds $t > 1$, the prompt additionally contains all round-$(t-1)$ responses from all agents, verbatim.

Random seeds and splits. Each reported number averages over 10 random 50/50 train/test splits, with seeds $\{7, 11, 13, 17, 19, 23, 29, 31, 37, 41\}$ applied to the split indicator. Within each split, the $\lambda^\star$ procedure (Appendix B.5) uses a further 50/50 sub-split with seed $\mathrm{seed} + 100$.

Hyperparameter grids. $\mathcal{K} = \{128, 256, 512\}$; $\delta = 0.03$; $\varepsilon_{\mathrm{act}} = 0.02$; $\alpha = \beta - 0.05$; λ-grid $= \{0.25, 0.50, \ldots, 5.00\}$ (20 values); modulus split 20/80; state normalization via training quantiles; logistic-GAM pilot with 8 knots per coordinate and ℓ_2 penalty $C = 1.0$.

Baseline thresholds. All baseline thresholds are chosen on the training split only, via exhaustive grid search on 200 points in $[0.5, 1.0]$ for score thresholds and $[0.0, 0.5]$ for margin thresholds. Selection criterion: maximize training-time Act subject to training-time WA $\leq \beta$ (matching our rule for fairness).

Compute. Full calibration run (LLM calls) requires $nT \approx 8.7\mathrm{k} \times 4 = 35\mathrm{k}$ calls per split per benchmark; 10 splits $\times$ 6 benchmarks ≈ 1.1M LLM calls total. k-NN certification per test instance is dominated by $O(T|\mathcal{K}|nD) \approx 8 \cdot 10^4$ distance computations—negligible compared to LLM inference. Anonymous code artifact, prompts, seeds, and aggregated result JSONs will be provided in the supplemental material.