# CITE: Anytime-Valid Statistical Inference in LLM Self-Consistency

**Hirofumi Ota**
Komaba Institute of Science
The University of Tokyo
Tokyo, Japan
hirofumi-ota@g.ecc.u-tokyo.ac.jp

**Naoto Iwase**
School of Medicine
Nagoya University
Nagoya, Japan
naoto@iwase.dev

**Yuki Ichihara**
Mohamed bin Zayed University
of Artificial Intelligence
Abu Dhabi, United Arab Emirates
y.ichihara3406@gmail.com

**Junpei Komiyama**
Mohamed bin Zayed University
of Artificial Intelligence
Abu Dhabi, United Arab Emirates
junpei@komiyama.info

**Masaaki Imaizumi**
The University of Tokyo / RIKEN AIP / Kyoto University
Tokyo, Japan
imaizumi@g.ecc.u-tokyo.ac.jp

## Abstract

Large language models often improve reasoning by sampling multiple outputs and aggregating their final answers, but precise and efficient control of error levels remains a challenging task. In particular, deciding when to stop sampling remains difficult when the stopping rule is data-dependent and the set of possible response labels is not known in advance. We study anytime-valid certification of a pre-specified target answer as the unique mode of the model's response distribution, a guarantee distinct from answer correctness. We propose the Certification by Intersection-union Testing with E-processes (CITE) algorithm, which provably controls false certification at any prescribed level under arbitrary data-driven stopping, without requiring prior knowledge of the answer category set. We also prove a category-set-size-free stopping-time rate, establish matching minimax lower bounds up to constants in the main regime, and extend the construction to confidence-weighted voting. Simulations and LLM self-consistency experiments show empirical error control and improved certification in diffuse-tail settings.

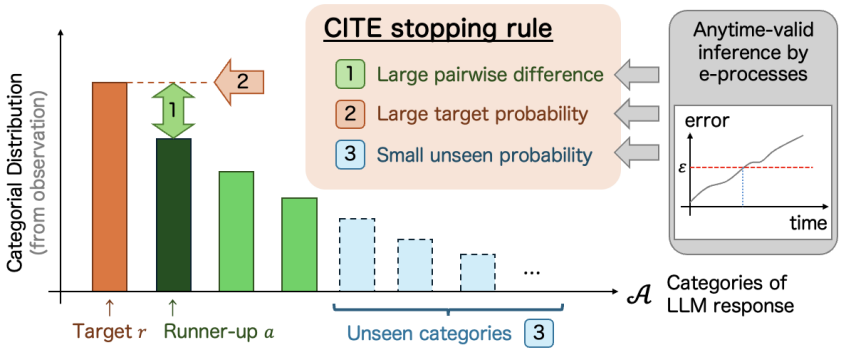
## 1 Introduction

### 1.1 Background and Motivation

Large language models (LLMs) have demonstrated increasingly strong performance on a wide range of tasks Brown et al. [2020], Chowdhery et al. [2023], Achiam et al. [2023], Touvron et al. [2023], especially from the use of *test-time compute*. Prompting methods such as chain-of-thought and least-to-most decomposition encourage models to expose intermediate reasoning steps, while inference-time strategies such as self-consistency and tree-of-thought style search improve robustness by sampling, exploring, and aggregating multiple candidate reasoning trajectories [Wei et al., 2022, Zhou et al., 2023, Wang et al., 2023, Yao et al., 2023, Brown et al., 2024, Snell et al., 2025].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Figure 1: Overview of the stopping rule by CITE. With the countably infinite set of categories $\mathcal{A}$, CITE tests the hypothesis of whether the target category $r \in \mathcal{A}$ is a mode or not.




A notion of *adaptive self-consistency* has gained attention to reduce computational cost of sampling growing approximately linearly with the number of sampled trajectories. A seminal approach is *majority voting* [Wang et al., 2023], which is simple, model-agnostic, and does not require a separate verifier, reward model, or additional training. In particular, the method adapts to the case that easy questions often stabilize after only a few samples, whereas hard questions remain ambiguous and may benefit from continued sampling. As extensions, Aggarwal et al. [2023] studies a stopping method based on posterior agreement among sampled answers ; Li et al. [2024] uses low-entropy windows as a practical stopping signal; Wang et al. [2025b] and Wang et al. [2025a] allocate budgets using difficulty estimates or bandit-style uncertainty across sets of queries ; and confidence-aware variants use response-level confidence to improve aggregation and stopping Taubenfeld et al. [2025], Aghazadeh et al. [2025].

One challenge of the literature is to handle the situation where we must precisely control error level $\varepsilon$ of certifying a modal response, those that arise when the risk of failure is high. A typical real example is the emergency-room triage on data such as ER-REASON [Mehandru et al., 2025], where each LLM-output acuity score determines the time-to-physician for an arriving patient and each disposition determines admission-versus-discharge. In this case, if a modal-response certificate is used operationally, the false-certification probability must be bounded at a prescribed level $\varepsilon$. At present, it is difficult to control the certification errors with precision; therefore, keeping errors below a certain threshold requires performing an excessive number of inferences, which directly leads to a massive increase in computational costs.

Based on the situation, we have the following question: *can we derive a data-driven stopping rule for adaptive self-consistency that controls the error level of certifying a modal response by LLMs?*

### 1.2 Our Study

In this study, we propose the *Certification by Intersection-union Testing with E-processes* (CITE), which is an *anytime-valid* and *category-agnostic* sequential test for controlling the Type-I error of falsely certifying a prespecified response label as the unique mode. We note the important features: (i) *Anytime validity*: CITE works with data-driven stopping time and the error guarantee must hold uniformly over the time, not at a single prespecified sample size. (ii) *Category-agnostic*, CITE works even when the response category is unknown a priori and grows as sampling proceeds, which can adapt to new responses from LLMs. CITE gives a stopping time based on the three conditions on empirical probabilities: (i) the target is larger than the runner-up, (ii) the target itself is large, and (iii) the unobserved category is small, using anytime-valid e-process machinery [Ville, 1939, Grünwald et al., 2024, Ramdas and Wang, 2025]. Figure 1 gives a graphical abstract.

Our contributions are as follows.

- **Anytime-valid category-agnostic error guarantee**: CITE controls Type-I error of certifying a label as the unique mode uniformly over all stopping times at any prescribed level $\varepsilon$ (Theorem 4.1). The error guarantee is valid without the knowledge of answer categories.
- **Category-size-free stopping-time**: We show that the stopping time by CITE has an order which is free from category size (Theorem 4.3), which is contrast to several baseline methods. We also derive a matching minimax lower bound (Theorem 4.4).

- **Extension to weighted case**: We extend CITE to handle a case with weighted observations and gaps. We also derive an anytime-valid Type-I control and yield the same rate of the stopping time (Section 5).
- **Empirical performance**: Simulations and LLM experiments (Section 6) confirm that CITE controls Type-I error empirically, matches or improves on baselines on three LLM benchmarks.

## 2 Problem Formulation

### 2.1 Setup and Observation

Let $\mathcal{A}$ be a countable (possibly infinite) set of response categories. We consider an unknown categorical distribution $P$ on $\mathcal{A}$, and write $p_a := P(X = a)$ for its probability mass function for $a \in \mathcal{A}$. Let $p_{(1)} \geq p_{(2)} \geq \cdots$ denote the decreasing rearrangement of $(p_a)_{a\in\mathcal{A}}$, and let $\mathcal{F}_t := \sigma(X_1, \ldots, X_t)$ be the natural filtration.

Suppose that we observe an i.i.d. sequence $X_1, X_2, \ldots$ from $P$:

**Assumption 2.1** (i.i.d. categorical data)**.** *$(X_t)_{t\geq 1}$ are i.i.d. draws from a categorical distribution $P$ on a countable set $\mathcal{A}$.*

We define some notations for statistics from the observations. For each $a \in \mathcal{A}$ and $t \geq 1$, let

$$N_t(a) := \sum_{i=1}^{t} \mathbf{1}\{X_i = a\}, \qquad \widehat{p}_t(a) := \frac{N_t(a)}{t}, \qquad \mathcal{A}_t := \{a \in \mathcal{A} : N_t(a) \geq 1\}$$

denote the count, empirical frequency, and set of observed categories by time $t$.

Our definition of $P$ has a distinctive feature: A may contain a countably infinite number of categories, and since $\mathcal{A}$ is unknown, information about the set of categories cannot be used in decision-making. This setup is effective for representing the complex response patterns of LLMs.

### 2.2 Statistical Testing

We study the fixed-target unique-mode testing problem. For a target category $r \in \mathcal{A}$ fixed *a priori*, i.e. independent of the data, we design a sequential test of the unique-mode hypothesis that $r$ strictly dominates every other category.

**Definition 2.2** (Null and alternative hypothesis)**.** *For a fixed target category $r \in \mathcal{A}$, set*

$$H_{0,r} := \big\{P : p_r \leq \sup_{a\neq r} p_a\big\}, \qquad H_{1,r} := \big\{P : p_r > \sup_{a\neq r} p_a\big\}.$$

*$H_{1,r}$ is the unique-mode alternative, i.e. $r$ strictly dominates every other category. Under any $P \in H_{0,r}$ there exists a deterministic witness $a^\star \neq r$ such that $p_{a^\star} \geq p_r$ holds. A procedure* certifies *$r$ whenever it rejects $H_{0,r}$ in favor of $H_{1,r}$.*

To handle this testing, we consider certifying the mode through a sequential procedure with the observations using the following criteria.

**Definition 2.3** (Sequential certification procedure)**.** *A sequential certification procedure for target $r$ is a sequence $(D_t)_{t\geq 1}$ of $\mathcal{F}_t$-measurable decision rules, where $D_t = 1$ means that, after observing $X_1, \ldots, X_t$, the procedure certifies $r$ as the unique mode. Its stopping time is*

$$\tau := \inf\{t \geq 1 : D_t = 1\},$$

*with $\inf \emptyset = \infty$. The procedure has anytime-valid Type-I error control at level $\varepsilon$ if it holds that*

$$\sup_{P\in H_{0,r}} \mathbb{P}_P(\tau < \infty) \leq \varepsilon. \tag{1}$$

*If $\tau < \infty$, the procedure certifies $r$ as the unique mode.*

Importantly, the sampling time $t$ is not fixed in advance, hence The bound (1) is time-uniform. At each time $t$, the procedure either continues sampling or stops and certifies, using only the data observed so far. The Type-I guarantee controls the probability that such a data-dependent rule ever falsely certifies $r$.

**Remark 2.4.** *We note the originality of our setting. While similar settings have been considered in statistics [Good, 1953, Painsky, 2025] and LLM test-time compute [Aggarwal et al., 2023, Li et al., 2024, Komiyama et al., 2026], to the best of our knowledge, an anytime valid sequential testing for the infinite and unknown category set has not been investigated. This setting describes with the nature of LLMs and represents a novel approach in terms of hypothesis testing.*

## 3 CITE: Certification by Intersection-union Testing with E-processes

We present our proposed method, CITE. Here, the target $r \in \mathcal{A}$ and the error level of certification $\varepsilon > 0$ is fixed before observing data as formalized in Kim and Ramdas [2025]. Our strategy involves using three data-driven stochastic processes to estimate the components of the distribution uniformly over time. In particular, we consider $\mathcal{F}_t$-measurable stochastic processes $E_t, L_t$ and $U_t$ with their forms will be derived in the following sections, which satisfies the following conditions under the null hypothesis $H_{0,r}$:

1. **Pairwise evaluation**: For $a \in \mathcal{A}$, $E_t$ evaluates an relative size of $p_r$ against $p_a$, that is, for any $\alpha \in (0,1)$, $\sup_{t\geq 1} E_t \leq 1/\alpha$ holds with probability at least $1-\alpha$.
2. **Lower bound of the target probability** $p_r$: the second process $L_t$ bounds $p_r$ below, that is, $\sup_{t\geq 1} L_t \leq p_r$ holds.
3. **Upper bound of unseen categories**: the third process $U_t$ is an upper bound of probability of unseen categories, that is, $\inf_{t\geq 1} U_t \geq \sup_{a\in\mathcal{A}:N_t(a)=0} p_a$ holds.

Note that these stochastic processes are calculated from observed values $X_1, ..., X_t$ without the knowledge of $P$.

Using these processes, we define our stopping rule: if any of the conditions above satisfied by these stochastic processes is violated at time $t$, we reject the null hypothesis $H_{0,r}$ and stop as $\tau = t$. Rigorously, the stopping rule, with $P \in H_{0,r}$ and $a^\star \neq r$ satisfying $p_{a^\star} \geq p_r$, provides the following inclusion:

$$\{\tau < \infty\} \subseteq \underbrace{\{\exists t : E_t \geq 3/\varepsilon\}}_{\text{(i) pairwise}} \cup \underbrace{\{\exists t : L_t(r) > p_r\}}_{\text{(ii) target probability}} \cup \underbrace{\{\exists a, t : N_t(a) = 0, p_a > U_t\}}_{\text{(iii) unseen categories}}. \tag{2}$$

In the following sections, we will define the stochastic processes $E_t$, $L_t$, and $U_t$ in detail and evaluate these three events.

### 3.1 Component (i): Pairwise E-Values

We give a specific form of the stochastic process $E_t$ by the notion of e-process in the testing-by-betting style of Ramdas et al. [2023]. For the fixed target $r \in \mathcal{A}$ and a competing category $a \neq r$, we consider a centered indicator $Z_i^{(r,a)} := \mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\} \in \{-1, 0, +1\}$ with a betting parameter $\lambda \in (0,1)$, and define the *pairwise e-process* as

$$E_t^{(r,a)}(\lambda) := \prod_{i=1}^{t} \big(1 + \lambda Z_i^{(r,a)}\big) = (1+\lambda)^{N_t(r)}(1-\lambda)^{N_t(a)}. \tag{3}$$

The closed form depends on the data only through $(N_t(r), N_t(a))$. Under the null hypothesis, i.e. $p_r \leq p_a$, it holds that $\sup_{t\geq 1} E_t^{(r,a)}(\lambda) \leq 1/\alpha$ with probability at least $1-\alpha$; namely, it is a Nonnegative Super-Martingale (NSM) process.

We define the stochastic process $E_t$ as the pairwise mixture e-process against the empirical runner-up. We extend $E_t^{(r,a)}(\lambda)$ by integrating $\lambda$ out. Since the optimal $\lambda$ is $(p_r - p_a)/(p_r + p_a)$ (Proposition C.6) depends on unknown values, we therefore mix over a finite grid $\Lambda_{\text{pw}} \subset (0,1)$ with positive weights $(w_\lambda)$ summing to at most 1. Namely, we define the pairwise mixture e-process as follows:

**Definition 3.1** (pairwise mixture e-process)**.** *For each competitor $a \neq r$, define*

$$E_t = E_t^{(r,\widehat{a}_t)} := \sum_{\lambda\in\Lambda_{\text{pw}}} w_\lambda E_t^{(r,\widehat{a}_t)}(\lambda), \quad \widehat{a}_t := \arg\max_{a\in\mathcal{A}_t\setminus\{r\}} N_t(a). \tag{4}$$

While $E_t^{(r,\widehat{a}_t)}$ is not a NSM process, $E_t^{(r,a)}$ is NSM since it is a convex combination of NSM processes:

**Corollary 3.2.** *Under $p_r \le p_a$, for any $\alpha \in (0,1)$, we have $\sup_{t\ge 1} E_t^{(r,a)} \le 1/\alpha$ with probability at least $1-\alpha$.*

**Remark 3.3** (Continuous pairwise mixtures)**.** *A uniform mixture is also valid:*

$$\int_0^1 E_t^{(r,a)}(\lambda)\, d\lambda = 2^{N_t(r)+N_t(a)+1} B_{1/2}(N_t(a)+1, N_t(r)+1)\,,$$

*where $B_z$ is the lower incomplete beta function. This avoids choosing a pairwise grid and mixture weights. We use finite mixtures 3 as a common construction, whose heterogeneous weights generally do not admit the same beta-function reduction.*

### 3.2 Component (ii): Lower Confidence Bound on the Target Probability $p_r$

We develop a time-uniform lower confidence bound (LCB) on the target probability $p_r$ to handle *hidden competitors*, i.e., categories $h \in \mathcal{A}$ with $p_h > p_r$ that have never appeared in the sample.

For each candidate value $q \in (0,1]$, we test a hypothesis "$p_r \le q$" by betting on the centered indicator $\mathbf{1}\{X_i = r\} - q$, which is an analogous of $\mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\}$ in Component (i). Fix a finite grid $\Lambda_r \subset (0,\infty)$ with positive weights $(v_\lambda)$ summing to at most 1, and for $q \in (0,1]$, we consider $\Lambda_r(q) := \{\lambda \in \Lambda_r : \lambda < 1/q\}$, thus each factor $1 + \lambda(\mathbf{1}\{X_i = r\} - q)$ is positive. Then, we define the mixture e-process

$$M_t(q) := \sum_{\lambda \in \Lambda_r(q)} v_\lambda \prod_{i=1}^{t} \big(1 + \lambda(\mathbf{1}\{X_i = r\} - q)\big), \tag{5}$$

and it follows $\mathbb{P}(\sup_t M_t(q) \ge 1/\alpha) \le \alpha$ for any $\alpha \in (0,1)$ under $p_r \le q$ (Corollary 3.2). Moreover, $q \mapsto M_t(q)$ is nonincreasing (Lemma C.3), so the set of $q$ at which the wealth exceeds $1/\alpha$ is a downset, making the following definition well-posed.

**Definition 3.4** (Lower confidence bound on $p_r$)**.** *Given $\varepsilon > 0$, at time $t$, we define the lower confidence bound on $p_r$ as*

$$L_t = L_t(r) := \sup\Big(\{0\} \cup \big\{q \in (0, \widehat{p}_t(r)] : M_t(q) \ge 3/\varepsilon\big\}\Big).$$

**Proposition 3.5** (LCB validity)**.** *Under Assumption 2.1, it holds that $\mathbb{P}\big(\exists t \ge 1 : L_t(r) > p_r\big) \le \varepsilon/3$.*

### 3.3 Component (iii): Unseen Upper Bound

We develop an upper bound for the probability of any currently unseen category, simultaneously over all categories and times. A category of mass $u \in (0,1)$ remains unseen after $t$ i.i.d. draws with probability $(1-u)^t$, and a *probability-weighted* union bound turns this fact into a time-uniform *per-category* upper bound, bypassing the need to estimate the aggregate missing mass.

**Definition 3.6** (Unseen upper bound)**.** *An upper bound of the unseen probabilities is defined as*

$$U_t := \min\big\{u \in (0,1] : u^{-1}(1-u)^t \le \varepsilon/3\big\}.$$

The threshold balances two forces: $(1-u)^t$ decays exponentially in $u$, while the $u^{-1}$ factor pays the union-bound cost over rare categories. As a result, we obtain the following guarantee:

**Proposition 3.7** (Simultaneous validity of the unseen bound)**.** *Under Assumption 2.1, we obtain*

$$\mathbb{P}\big(\exists a \in \mathcal{A},\ \exists t \ge 1 : N_t(a) = 0 \text{ and } p_a > U_t\big) \le \varepsilon/3.$$

### 3.4 Stopping Rule

We define the CITE stopping rule by combining these three stochastic processes above, which are uniform in time. In particular, we design the stopping rule to follow the inclusion of the events as (2). Algorithm 1 shows the pseudo-code of the CITE stopping rule.

**Definition 3.8** (CITE stopping rule)**.** *With given $\varepsilon > 0$, CITE stops at the first time both **(a) Pairwise rejection:** $E_t \ge 3/\varepsilon$ at the empirical runner-up, and **(b) LCB exceeds unseen bound:** $L_t > U_t$, hold simultaneously, i.e. we define the stopping time as*

$$\tau := \inf\big\{t \ge 1 : E_t \ge 3/\varepsilon \text{ and } L_t > U_t\big\}. \tag{6}$$

*If $\tau < \infty$, the procedure certifies $r$ as the unique mode.*

**Algorithm 1** Stopping rule by CITE

**Input:** target label $r \in \mathcal{A}$; parameter grid $\Lambda_{\mathrm{pw}} \subset (0,1)$ with weights $(w_\lambda)$; LCB grid $\Lambda_r \subset (0,\infty)$ with weights $(v_\lambda)$; error level $\varepsilon > 0$.

**Output:** stopping time $\tau$ and certification decision.

1. Initialise $N_0(a) \leftarrow 0$ for all $a \in \mathcal{A}$; $E_0^{(r,a)}(\lambda) \leftarrow 1$ for all $a \neq r$, $\lambda \in \Lambda_{\mathrm{pw}}$.
2. **For** $t = 1, 2, \ldots$ **do**

   2a. Draw $X_t$ and set $N_t(a) \leftarrow N_{t-1}(a) + \mathbf{1}\{X_t = a\}$ for all $a \in \mathcal{A}$.

   2b. STEP(I): PAIRWISE E-PROCESSES.
   If $\mathcal{A}_t \setminus \{r\} = \emptyset$, set $P_t^{\mathrm{pw}} \leftarrow \texttt{true}$ (the unseen condition $U_t$ then governs whether $\tau = t$) and skip to Step (II);
   otherwise, set $\widehat{a}_t \leftarrow \arg\max_{a \in \mathcal{A}_t \setminus \{r\}} N_t(a)$ and, for each $\lambda \in \Lambda_{\mathrm{pw}}$, compute

   $$E_t = E_t^{(r,\widehat{a}_t)} \leftarrow \textstyle\sum_\lambda w_\lambda E_t^{(r,\widehat{a}_t)}(\lambda), \qquad E_t^{(r,\widehat{a}_t)}(\lambda) \leftarrow (1+\lambda)^{N_t(r)}(1-\lambda)^{N_t(\widehat{a}_t)},$$

   and set $P_t^{\mathrm{pw}} \leftarrow \big\{E_t \geq 3/\varepsilon\big\}$.

   2c. STEP(II): LOWER CONFIDENCE BOUND ON $p_r$. Set $\widehat{p}_t(r) \leftarrow N_t(r)/t$ and

   $$L_t = L_t(r) \leftarrow \sup\big(\{0\} \cup \{q \in (0, \widehat{p}_t(r)] : M_t(q) \geq 3/\varepsilon\}\big).$$

   2d. STEP(III): UNSEEN UPPER BOUND. Set $U_t \leftarrow \min\{u \in (0,1] : u^{-1}(1-u)^t \leq \varepsilon/3\}$.

   2e. **If** $P_t^{\mathrm{pw}}$ **and** $L_t > U_t$ **then** return ($\tau \leftarrow t$, "certify $r$ as unique mode").

# 4 Theoretical Guarantees

We establish three theoretical guarantees for CITE. First, the false certification probability is controlled at the prescribed anytime-valid level for the fixed target $r$ (Theorem 4.1). Second, under the unique-mode alternative, the expected stopping time is bounded by a category-set-size-free rate depending on $(\delta, p_r)$ but not on $|\mathcal{A}|$ (Theorem 4.3). Third, a combined information-theoretic lower bound shows that this rate is optimal up to constants and logarithmic terms (Theorem 4.4).

## 4.1 Main Theorems

We first provide the sequential version of Type-I error control on the stopping time $\tau$ given in Definition 3.8.

**Theorem 4.1** (Type-I error control)**.** *Let $\varepsilon \in (0,1)$ be a prespecified significance level. Under Assumption 2.1, the stopping time $\tau$ of Definition 3.8 satisfies*

$$\sup_{P \in H_{0,r}} \mathbb{P}_P(\tau < \infty) \leq \varepsilon.$$

The proof, deferred to Appendix B.6, fixes a null witness $a^\star \neq r$ satisfying $p_{a^\star} \geq p_r$. If this witness has been observed by the stopping time, runner-up monotonicity implies that the fixed-competitor e-process $E_t^{(r,a^\star)}$ must have crossed $3/\varepsilon$. If the witness is still unseen, then the inequality $L_t > U_t$ can hold only if either the LCB or unseen bound has failed. This yields (2) and a union bound over the three $\varepsilon/3$ components.

The above result on the Type-I error control does not provide any guarantee regarding the time required to certify the true mode. Under the alternative, write the modal gap as $\delta := p_r - p_{(2)} > 0$. We now show that $\mathbb{E}[\tau]$ is bounded by a function of $(\delta, p_r)$ alone, independent of $|\mathcal{A}|$.

**Condition 4.2.** $\Lambda_{\mathrm{pw}}$ *contains some* $\lambda_{\mathrm{pw}} \in [\delta/8, \delta/4]$*, and* $\Lambda_r$ *contains some* $\lambda_r \in [1/32, 1/16]$.

Theorem F.5 supplies a geometric grid for $\delta \geq \delta_0$, explicitly accounting for its mixture-weight penalty. A continuous pairwise mixture is also valid; see Appendix C.

**Theorem 4.3** (Stopping-time bound)**.** *Suppose Assumption 2.1 holds and Condition 4.2 with $\delta > 0$, and the selected grid weights $w_{\lambda_{\mathrm{pw}}}, v_{\lambda_r}$ are bounded below by a constant independent of $\delta, p_r, \varepsilon$.*

*Then, we have* $\mathbb{P}(\tau < \infty) = 1$ *and*

$$\mathbb{E}[\tau] = O\left(\frac{\log(1/\varepsilon) + \log(1/\delta)}{\delta^2} + \frac{\log(1/\varepsilon) + \log(1/p_r)}{p_r}\right), \tag{7}$$

*with universal constants independent of* $\delta$, $p_r$, $|\mathcal{A}|$.

We next state the matching lower bound used to interpret Theorem 4.3. For $p \in (0, 1/2]$ and $\delta \in (0, p)$, define

$$\mathcal{P}(p, \delta) := \left\{P : p_r = p, \quad p_{(2)} \le p - \delta\right\}.$$

**Theorem 4.4** (Combined minimax lower bound)**.** *Fix* $p \in (0, 1/2]$ *and* $\delta \in (0, p)$. *Assume either that* $\mathcal{A}$ *is countably infinite, or that it contains at least* $\lceil (1-p)/(p-\delta) \rceil + 2$ *distinct non-target categories. Let* $\tau'$ *be any level-*$\varepsilon$ *sequential certification procedure for the fixed target* $r$, *i.e.,* $\sup_{Q \in H_{0,r}} \mathbb{P}_Q(\tau' < \infty) \le \varepsilon$. *Assume that, for every* $P \in \mathcal{P}(p, \delta)$, $\mathbb{P}_P(\tau' < \infty) = 1$ *and* $\mathbb{E}_P[\tau'] < \infty$ *hold. Then, there is a universal constant* $c > 0$ *such that*

$$\sup_{P \in \mathcal{P}(p,\delta)} \mathbb{E}_P[\tau'] \ge c \log(1/\varepsilon) \left(\frac{p}{\delta^2} + \frac{1}{p}\right).$$

Comparing the lower bound of Theorem 4.4 with the upper bound (7), we show that the CITE stopping time achieves the minimax optimal order up to logarithmic factors whenever $p_r$ is bounded away from zero. In the very small-$p_r$ regime, the current upper bound can be worse in the pairwise term; closing this gap is left open.

In Appendix F, we give extensions of the analysis: sharper bounds with refined constant (Theorem F.2), a growth rate of the e-process (Theorem F.8); a stopping time bound with an geometric grid (Theorem F.5); and a top-$k$ extension for the Type-I error and the stopping-time bound (Theorem F.9).

## 5 Extension to Confidence-Weighted Voting

### 5.1 Setup and Weighted CITE

We extend CITE to the confidence-weighted case: suppose we observe a pair $(X_i, W_i)$ with a confidence score $W_i \in [0, 1]$, e.g. the score from logits or a verifier; see [Taubenfeld et al., 2025, Fu et al., 2025]. In this case, our purpose is to certify the *weighted unique mode* $\arg\max_a \mu_a$ with $\mu_a := \mathbb{E}[W_i \mathbf{1}\{X_i = a\}]$ and weighted modal gap $\Delta_w := \mu_r - \mu_\star$, where $\mu_\star := \sup_{a \ne r} \mu_a$.

**Assumption 5.1** (i.i.d. weighted categories)**.** $(X_i, W_i)_{i \ge 1}$ *are i.i.d. with* $X_i \in \mathcal{A}$ *and* $W_i \in [0, 1]$.

We develop *Weighted CITE* (W-CITE) for this setting. The weighted null and alternative are

$$H_{0,r}^{(w)} := \{P : \mu_r \le \mu_\star\}, \quad \text{and} \quad H_{1,r}^{(w)} := \{P : \mu_r > \mu_\star\}. \tag{8}$$

W-CITE extends the three components of Section 3 to the weighted masses $(\mu_a)_{a \in \mathcal{A}}$.

**Component (i): Pairwise E-Values.** We consider the weighted indicator $\widetilde{Z}_i^{(r,a)} := W_i(\mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\})$ and a corresponding weighted pairwise e-process and its mixture as

$$\widetilde{E}_t^{(r,a)} := \sum_{\lambda \in \Lambda_{\mathrm{pw}}} w_\lambda \widetilde{E}_t^{(r,a)}(\lambda), \qquad \widetilde{E}_t^{(r,a)}(\lambda) := \prod_{i=1}^{t} \big(1 + \lambda \widetilde{Z}_i^{(r,a)}\big). \tag{9}$$

Because runner-up monotonicity need not hold with heterogeneous weights, W-CITE uses the all-competitors check $\min_{a \in \mathcal{A}_t \setminus \{r\}} \widetilde{E}_t^{(r,a)} \ge 3/\varepsilon$.

**Component (ii): LCB on Target Weighted Mass.** The LCB targets $\mu_r$ via the increment $W_i \mathbf{1}\{X_i = r\} - q$. Let $\widehat{\mu}_t(r) := \frac{1}{t} \sum_{i=1}^{t} W_i \mathbf{1}\{X_i = r\}$. For $q \in (0, 1]$, define

$$\widetilde{M}_t(q) := \sum_{\lambda \in \Lambda_r(q)} v_\lambda \widetilde{M}_t(q, \lambda), \qquad \widetilde{M}_t(q, \lambda) := \prod_{i=1}^{t} \big(1 + \lambda(W_i \mathbf{1}\{X_i = r\} - q)\big),$$

and set $\widetilde{L}_t(r) := \sup(\{0\} \cup \{q \in (0, \widehat{\mu}_t(r)] : \widetilde{M}_t(q) \ge 3/\varepsilon\})$.

**Component (iii): Unseen Upper Bound.** The unseen bound $U_t$ is unchanged since $\mu_a \le p_a$.

**Stopping Time.** W-CITE stops when the weighted pairwise check passes for every observed competitor and the weighted LCB exceeds the unseen bound, $\widetilde{L}_t(r) > U_t$; if no non-target competitor has been observed, the pairwise part is treated as vacuous. This all-competitors check raises the per-round cost from $O(1)$ to $O(|\mathcal{A}_t|)$, since runner-up monotonicity (Lemma B.3) fails under heterogeneous weights. We write $\tau^{(w)}$ for the resulting stopping time, and give the validity analysis in Appendix D.

### 5.2 Theoretical Guarantee for Weighted CITE

We develop theoretical analysis for W-CITE, including the Type-I error and the stopping time.

**Theorem 5.2** (Type-I error, weighted case)**.** *Under Assumption 5.1,* $\sup_{P \in H_{0,r}^{(w)}} \mathbb{P}_P(\tau^{(w)} < \infty) \le \varepsilon$.

**Condition 5.3** (Weighted grid coverage)**.** $\Lambda_{\mathrm{pw}}$ *contains some* $\lambda_{\mathrm{pw}} \in [\Delta_w/8, \Delta_w/4]$ *and* $\Lambda_r$ *contains some* $\lambda_r \in [1/32, 1/16]$.

**Theorem 5.4** (Weighted stopping-time bound)**.** *Suppose Assumption 5.1 holds with* $\Delta_w > 0$ *and Condition 5.3, and the grid weights* $w_{\lambda_{\mathrm{pw}}}, v_{\lambda_r}$ *are bounded below by a constant independent of* $\Delta_w, \mu_r, \varepsilon$*. Then we have* $\mathbb{P}(\tau^{(w)} < \infty) = 1$ *and*

$$\mathbb{E}[\tau^{(w)}] = O\left(\frac{\log(1/\varepsilon) + \log(1/\Delta_w)}{\Delta_w^2} + \frac{\log(1/\varepsilon) + \log(1/\mu_r)}{\mu_r}\right). \tag{10}$$

The weighted guarantee mirrors the unweighted theory (7) with $(p_r, \delta)$ replaced by $(\mu_r, \Delta_w)$. The bound suggests that weighting can improve stopping time when it increases the effective gap $\Delta_w$ without reducing the target weighted mass $\mu_r$ too much.

## 6 Empirical Validation

We evaluate CITE on simulated multinomial distributions and on LLM self-consistency benchmarks. The experiments are designed to evaluate the following quantities: empirical false certification (Type-I error) under null targets, empirical power measured by certification probability under alternatives, and stopping time or sample efficiency, both on synthetic multinomial models and on LLM answer distributions. Baselines are **Bonferroni** (fixed-sample union bound), **KR** [Kim and Ramdas, 2025] (fixed-target sample-split test), and **MMC** [Cordero-Encinar and Duncan, 2025] (sequential leader-tracking), all at level $\varepsilon = 0.05$. We do not include adaptive stopping methods without theoretical guarantees of Type-I error control. Detailed protocols, per-setting tables, and ablations on component scaling and the confidence-weighted W-CITE are deferred to Appendices G and H.

**Simulations.** We evaluate CITE on five multinomial settings (Appendix G.1). First, we check false certification under the fixed-target null. In Case B, across all five settings and all sample sizes, CITE has empirical false-certification rate zero, consistent with Theorem 4.1; the same is true for W-CITE and fixed-sample Bonferroni. By contrast, KR shows small nonzero false-certification rates on Settings 3–4, up to $0.002$, and MMC shows nonzero false certifications in the same regimes, reaching $0.006$ on Setting 3 and $0.014$ on Setting 4 (Table 3).

Second, we evaluate certification power and stopping time under the alternative. In the easy concentrated Setting 2, all methods saturate quickly: at $N = 64$, CITE certifies with rate $0.890$ and mean stopping time $\bar{\tau} \approx 36$, while Bonferroni and KR achieve rates $0.928$ and $0.946$, respectively. In the diffuse Setting 1, CITE is more efficient than Bonferroni before saturation, with certification rates $0.188$ vs. $0.100$ at $N = 64$ and $0.892$ vs. $0.804$ at $N = 128$, reaching full certification by $N = 256$ with $\bar{\tau} \approx 90$. In the moderate LLM-like Setting 5, CITE reaches rate $0.544$ at $N = 256$ and $0.932$ at $N = 512$ with $\bar{\tau} \approx 234$, whereas MMC remains at rate at most $0.002$ through $N = 2048$. KR often has higher finite-sample power, but it is a fixed-sample baseline rather than an anytime-valid stopping rule. In the near-tie Setting 3 and very diffuse Setting 4, all methods remain low-power within the evaluated budget range.

Finally, the component-level simulations are consistent with the two terms in Theorem 4.3. When $p_r$ is varied at fixed gap, the LCB–unseen component follows the predicted $p_r^{-1}$ dependence; in

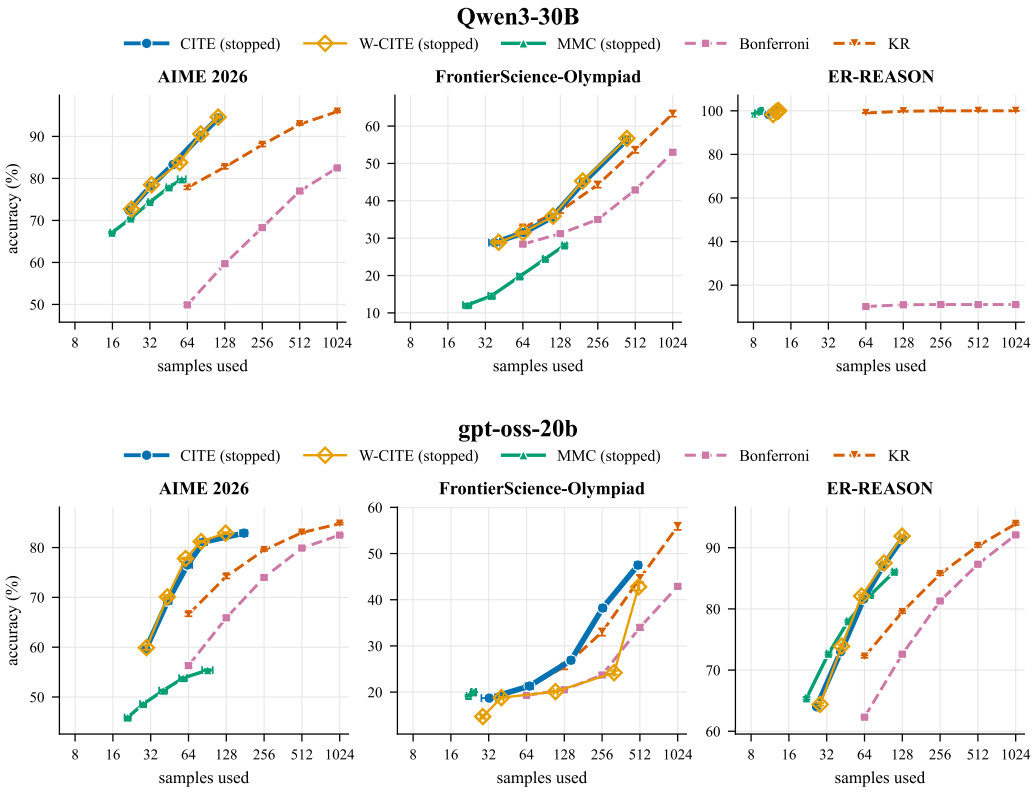



Figure 2: Mean accuracy vs. sample budget on Qwen3-30B (top) and gpt-oss-20b (bottom), across AIME 2026, FrontierScience-Olympiad, and ER-REASON. For the sequential methods (CITE, W-CITE, MMC) the horizontal axis is the mean stopping time; for the fixed-sample baselines (Bonferroni, KR) it is the prefix sample size $N$.

the reported sweep, $\bar{\tau}_{\mathrm{lu}} p_r$ stays between roughly 14 and 20. When $\delta$ is varied at fixed $p_r = 0.24$, the LCB–unseen component stays approximately constant, around $\bar{\tau}_{\mathrm{lu}} \approx 70$, while the pairwise component is the part most affected by the modal gap (Appendix G.3).

**LLM self-consistency.** We evaluate the performance of Qwen3-30B (`Qwen3-30B-A3B-Instruct-2507`) [Qwen Team, 2025] and gpt-oss-20b (`openai/gpt-oss-20b`) [OpenAI, 2025] on AIME 2026 [Dekoninck et al., 2026], FrontierScience-Olympiad [Wang et al., 2026], and the clinical benchmark ER-REASON [Mehandru et al., 2025]. The LLM experiments are conditional bootstrap evaluations with targets fixed from empirical answer pools; they do not establish end-to-end guarantees for the original LLM distribution (Appendix H). Figure 2 plots mean accuracy against the sample budget; the horizontal axis is the mean stopping time for the sequential methods (CITE, W-CITE, MMC) and the prefix sample size $N$ for the fixed-sample baselines (Bonferroni, KR). The LLM experiments are conditional bootstrap evaluations with targets fixed from empirical answer pools; they do not establish end-to-end guarantees for the original LLM distribution (Appendix H).

When the target is set to the empirical runner-up rather than the mode, the maximum false-certification rate over all evaluated methods, problems, and budgets is $0.052$ on gpt-oss-20b/ER-REASON and at most $0.012$ on every other tested (model, dataset) pair; both lie within Monte Carlo error of the nominal level $\varepsilon = 0.05$ given the multiplicity of comparisons. On Qwen3-30B/ER-REASON, fixed-sample Bonferroni saturates at certification rate $0.111$ for $N \geq 256$, whereas CITE attains certification rate $\geq 0.98$ at mean stopping time $\bar{\tau} \approx 12$. This gives a direct stopping-time diagnostic: in this concentrated-answer regime, CITE certifies after roughly a dozen samples on average, while the fixed-sample baseline remains weak even at much larger budgets. At the largest evaluated budget $N = 1024$, the CITE certification rate exceeds that of MMC by a factor of approximately 2 on FrontierScience-Olympiad (where the empirical answer set contains about 250 distinct values per

problem on gpt-oss-20b), while on ER-REASON the two methods differ by at most $0.054$ across all budgets, attained on gpt-oss-20b. Relative to the KR baseline, the CITE certification rate is lower by at most $0.087$, attained on Qwen3-30B/FrontierScience at $N \in \{256, 512\}$, while retaining anytime validity, whereas KR, by construction, is not valid under adaptive stopping. Per-problem and per-budget tables and the corresponding results for the W-CITE are deferred to Appendix H.

## 7 Conclusion

We introduced CITE for sequentially certifying the unique mode of a discrete distribution on a countable, possibly infinite category set. Our theory guarantees that anytime-valid Type-I error control, the stopping time has the category-set-size free order, and the order matches the information-theoretic lower bound up to logarithmic factors. We also extend CITE to confidence-weighted voting and develop its weighted version W-CITE. In our experiments, both numerical simulations and real LLM environments demonstrated that CITE achieves high accuracy while controlling Type-I errors.

**Limitations and future work.** CITE provides statistical guarantee that a prespecified label is the unique mode of the model's response distribution. This is logically distinct from, and complementary to, *answer correctness*: the two compose, and neither subsumes the other, so a downstream system can combine our anytime-valid mode certificate with any separate correctness signal. Within this scope, a natural extension is to certify the empirically observed top category $\widehat{r}_t \in \arg\max_a N_t(a)$, i.e., a data-adaptive target; this requires a Bonferroni-type correction over the random sequence of empirical leaders, paralleling the construction of Cordero-Encinar and Duncan [2025], and whether the category-size-free rate of Theorem 4.3 survives this correction is left open.


## Acknowledgments and Disclosure of Funding

H.O. was supported by JSPS KAKENHI Grant Number 25K23133. M.I. was supported by JSPS KAKENHI (24K02904, 25H01124), JST CREST (JPMJCR21D2), FOREST (JPMJFR216I), ASPIRE (JPMJAP2329), ARiSE (JPMJAR2633) and BOOST (JPMJBY24A9). J. K. was supported by the MBZUAI Start-up Fund [BF0121]. The authors declare no competing interests.


## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Pranjal Aggarwal, Aman Madaan, Yiming Yang, et al. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12375–12396, 2023.

Ehsan Aghazadeh, Ahmad Ghasemi, Hedyeh Beyhaghi, and Hossein Pishro-Nik. Cges: Confidence-guided early stopping for efficient and accurate self-consistency. *arXiv preprint arXiv:2511.02603*, 2025.

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013. doi: 10.1214/ECP.v18-2359.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Paula Cordero-Encinar and Andrew B Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in llms. *arXiv preprint arXiv:2510.17472*, 2025.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvalddson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs. 2026. URL `https://arxiv.org/abs/2605.00674`.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953. doi: 10.1093/biomet/40.3-4.237.

I. J. Good and G. H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956. doi: 10.1093/biomet/43.1-2.45.

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

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17, 2020.

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. doi: 10.1214/20-AOS1991.

Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Ilmun Kim and Aaditya Ramdas. Locally minimax optimal confidence sets for the best model. *arXiv preprint arXiv:2503.21639*, 2025.

Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. Best-of-infinity: Asymptotic performance of test-time llm ensembling. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael Lindon and Alan Malek. Anytime-valid inference for multinomial count data. *Advances in Neural Information Processing Systems*, 35:2817–2831, 2022.

David McAllester and Luis E. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911, 2003.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F. Molina, and Ahmed Alaa. ER-REASON: A Benchmark Dataset for LLM-Based Clinical Reasoning in the Emergency Room. *arXiv preprint arXiv:2505.22919*, 2025. URL `https://arxiv.org/abs/2505.22919`.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*, 2025. URL `https://arxiv.org/abs/2508.10925`.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016. doi: 10.1073/pnas.1607774113.

Amichai Painsky. Confidence intervals for parameters of unobserved events. *Journal of the American Statistical Association*, 120(549):226–236, 2025.

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

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390, 2025.

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

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2): 407–431, 2021. doi: 10.1111/rssa.12647.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20090–20111, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris, 1939.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. Dynscaling: Efficient verifier-free inference scaling via dynamic and integrated sampling. *arXiv preprint arXiv:2506.16043*, 2025a.

Miles Wang, Robi Lin, Kat Hu, Joy Jiao, Neil Chowdhury, Ethan Chang, and Tejal Patwardhan. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks. *arXiv preprint arXiv:2601.21165*, 2026. URL https://arxiv.org/abs/2601.21165.

Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6904–6917, 2025b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

# A Related Works

**Self-consistency and test-time compute for LLM reasoning.** Self-consistency and related inference-time scaling methods improve LLM reasoning by sampling multiple reasoning trajectories and aggregating their terminal answers, typically by majority voting [Wei et al., 2022, Zhou et al., 2023, Wang et al., 2023, Yao et al., 2023, Brown et al., 2024, Snell et al., 2025]. These methods motivate our statistical formulation: after normalization, repeated LLM outputs are viewed as i.i.d. draws from an unknown categorical distribution over terminal response labels. Our work is not a new prompting, search, or reranking method. Instead, it studies when one may stop sampling and certify, with a prescribed error level, that a prespecified response label is the unique mode of this distribution. This is a guarantee about the model-induced response distribution, and is distinct from semantic correctness of the answer.

**Adaptive self-consistency and early stopping.** A growing line of work reduces the cost of self-consistency by adapting the number of sampled trajectories to the apparent difficulty of each query. Adaptive-Consistency uses posterior agreement among sampled answers as a lightweight stopping signal [Aggarwal et al., 2023]; Early-Stopping Self-Consistency stops when a low-entropy window indicates sufficient agreement [Li et al., 2024]; difficulty-adaptive and dynamic variants allocate inference budgets using difficulty estimates, answer-distribution uncertainty, or reasoning-path quality [Wang et al., 2025b,a]; and confidence-aware methods use response-level confidence or verifier-like scores to improve aggregation and stopping [Taubenfeld et al., 2025, Aghazadeh et al., 2025, Fu et al., 2025]. These methods are primarily designed to improve the accuracy–cost tradeoff. By contrast, CITE is designed for statistical certification: it controls the probability of falsely certifying the target as the unique mode under arbitrary data-driven stopping, at a user-specified level $\varepsilon$.

**Certified self-consistency.** The closest LLM-specific statistical work is the Martingale Majority Certificate (MMC) of Cordero-Encinar and Duncan [2025], which develops finite-sample and anytime-valid certificates for self-consistency and uses martingale arguments to determine when enough samples have been drawn. MMC provides an important first step toward certified self-consistency, but it certifies a stronger majority/residual-tail condition than the fixed-target unique-mode property considered here. Our target differs in two respects. First, CITE tests a fixed-target unique-mode alternative, $p_r > \sup_{a \neq r} p_a$, whereas MMC-style certificates are naturally tied to a stronger majority or residual-tail domination condition. These two notions coincide only in non-diffuse regimes. In LLM self-consistency, however, the modal answer may have substantially larger mass than every individual competitor while still being smaller than the aggregate mass of the long tail. Proposition E.1 shows that on a provable least-favourable family in this regime, the MMC certifies with probability at most the nominal level $\varepsilon$. Second, CITE explicitly handles the possibility that a high-mass competitor has not yet been observed. This unseen-category component is essential when the set of normalized free-form answers is unknown and grows during sampling.

**Anytime-valid inference, e-values, and confidence sequences.** Our construction builds on the classical martingale view of sequential testing [Ville, 1939] and the modern theory of e-values, e-processes, testing by betting, and confidence sequences [Howard et al., 2020, 2021, Shafer, 2021, Vovk and Wang, 2021, Lindon and Malek, 2022, Grünwald et al., 2024, Ramdas and Wang, 2025]. These tools provide tests and confidence statements that remain valid under continuous monitoring and data-dependent stopping. CITE instantiates this machinery for a composite categorical unique-mode problem: it combines pairwise e-processes against observed competitors with a time-uniform lower confidence bound for the target mass and an upper bound for unseen competitors. The resulting intersection-union test yields a certification rule whose Type-I error is valid uniformly over all stopping times.

**Best-arm identification and discrete argmin inference.** The fixed-target unique-mode problem is related to fixed-confidence best-arm identification and pure-exploration bandits [Garivier and Kaufmann, 2016, Kaufmann et al., 2016, Kaufmann and Koolen, 2021]. Those works characterize the sample complexity of identifying the best arm under sequential sampling and often rely on a known finite set of arms. Our setting is different: each LLM query returns a single categorical draw from the entire response distribution, the category set is not known in advance, and a previously unseen response can become a relevant competitor. The closest fixed-target finite-dimensional inference problem is the dimension-agnostic discrete argmin test of Kim and Ramdas [2025], which we include

as a baseline. CITE differs by providing an anytime-valid stopping rule for a countable and unknown category set, rather than a fixed-sample or sample-split test over a given finite vector.

**Unseen species, missing mass, and support-agnostic inference.** The need to reason about unobserved response labels connects our work to the classical unseen-species and missing-mass literature, including Good–Turing and Good–Toulmin estimation [Good, 1953, Good and Toulmin, 1956], concentration bounds for the missing mass [McAllester and Ortiz, 2003, Berend and Kontorovich, 2013], and minimax prediction of unseen species [Orlitsky et al., 2016]. These works estimate or bound quantities such as the aggregate probability of unseen outcomes or the number of new species. CITE uses a different object: for certification of a unique mode, it is enough to rule out the existence of a single unseen competitor whose mass could exceed the target. The unseen component of CITE is therefore tailored to sequential modal certification rather than to estimating the entire missing mass or support size.

**Confidence-weighted aggregation.** Recent work shows that response-level confidence can improve self-consistency by weighting or prioritizing high-confidence reasoning paths [Taubenfeld et al., 2025, Aghazadeh et al., 2025, Fu et al., 2025]. Our weighted extension formalizes a complementary question: when the observations are pairs $(X_i, W_i)$ with bounded confidence weights, can one certify that a target label is the unique weighted mode? W-CITE extends the same anytime-valid and category-agnostic guarantee to this weighted setting, with the weighted gap replacing the unweighted modal gap in the stopping-time analysis.

## B Proofs for Section 3

Let $\alpha_{\mathrm{pw}}, \alpha_r, \alpha_u > 0$ be error levels of each component satisfying $\alpha_{\mathrm{pw}} + \alpha_r + \alpha_u \le \varepsilon$. The main text uses the equal split

$$\alpha_{\mathrm{pw}} = \alpha_r = \alpha_u = \varepsilon/3.$$

For readability, the Appendix proofs keep the symbols $\alpha_{\mathrm{pw}}, \alpha_r, \alpha_u$ and substitute the equal split only at the end.

We prepare some supportive results as follow.

For an $(\mathcal{F}_t)$-adapted nonnegative process $(M_t)_{t\ge 0}$ with $\mathbb{E}_P[M_t \mid \mathcal{F}_{t-1}] \le M_{t-1}$ and $M_0 \le 1$ (an NSM), Ville's inequality [Ville, 1939] gives the time-uniform tail bound

$$P\Big(\sup_{t\ge 0} M_t \ge 1/\alpha\Big) \le \alpha, \qquad \alpha \in (0,1], \tag{11}$$

which we invoke repeatedly in what follows.

**Lemma B.1** (Attainment of a positive supremum)**.** *Let $(x_a)_{a\in\mathcal{B}}$ be a nonnegative summable sequence indexed by a countable set $\mathcal{B}$. If $\sup_{a\in\mathcal{B}} x_a > 0$, then the supremum is attained.*

*Proof.* Let $s := \sup_{a\in\mathcal{B}} x_a > 0$. If no index attained $s$, then for each $n$ one could choose a distinct $a_n \in \mathcal{B}$ with $x_{a_n} > s/2$. This would contradict $\sum_{a\in\mathcal{B}} x_a < \infty$. □

**Lemma B.2** (Pairwise NSM)**.** *Fix $a \ne r$ and $\lambda \in (0,1)$. Under any $P$ with $p_r \le p_a$, the process $(E_t^{(r,a)}(\lambda))_{t\ge 0}$ is an NSM with $E_0^{(r,a)}(\lambda) = 1$.*

**Lemma B.3** (Runner-up monotonicity)**.** *For any $a, b \in \mathcal{A}_t \setminus \{r\}$ with $N_t(a) \ge N_t(b)$ and any $\lambda \in (0,1)$, it holds that*

$$E_t^{(r,a)}(\lambda) \le E_t^{(r,b)}(\lambda).$$

*Consequently, we have*

$$E_t^{(r,a)} \le E_t^{(r,b)}.$$

### B.1 Proof of Lemma B.2

*Proof.* We have $E_0 = 1$ (empty product) and $E_t \ge 0$ since each factor $1 + \lambda Z_i \ge 1 - \lambda > 0$. For the supermartingale property:

$$\mathbb{E}[E_t^{(r,a)}(\lambda) \mid \mathcal{F}_{t-1}] = E_{t-1}^{(r,a)}(\lambda) \cdot (1 + \lambda \mathbb{E}[Z_t^{(r,a)}])$$

$$= E_{t-1}^{(r,a)}(\lambda) \cdot (1 + \lambda(p_r - p_a)),$$

using the i.i.d. property. Under $p_r \le p_a$, the factor $1 + \lambda(p_r - p_a) \le 1$. □

### B.2 Proof of Corollary 3.2

*Proof.* For each $\lambda \in \Lambda_{\mathrm{pw}}$, $(E_t^{(r,a)}(\lambda))_{t\ge 0}$ is an NSM under $p_r \le p_a$ by Lemma B.2. Since

$$E_t^{(r,a)} = \sum_{\lambda \in \Lambda_{\mathrm{pw}}} w_\lambda E_t^{(r,a)}(\lambda)$$

is a nonnegative weighted sum with total weight at most one, it is also an NSM and satisfies $E_0^{(r,a)} \le 1$. Ville's inequality (11) then gives

$$\mathbb{P}_P\left(\sup_{t\ge 0} E_t^{(r,a)} \ge \alpha^{-1}\right) \le \alpha.$$

□

### B.3 Proof of Lemma B.3

*Proof.* By (3), $E_t^{(r,a)}(\lambda) = (1+\lambda)^{N_t(r)}(1-\lambda)^{N_t(a)}$. The factor $(1+\lambda)^{N_t(r)}$ depends only on $r$ and is common to both $E_t^{(r,a)}(\lambda)$ and $E_t^{(r,b)}(\lambda)$. Since $0 < 1-\lambda < 1$ and $N_t(a) \ge N_t(b)$,

$$(1-\lambda)^{N_t(a)} \le (1-\lambda)^{N_t(b)},$$

whence $E_t^{(r,a)}(\lambda) \le E_t^{(r,b)}(\lambda)$. Multiplying by the nonnegative weights $w_\lambda$ and summing over $\lambda \in \Lambda_{\mathrm{pw}}$ gives

$$E_t^{(r,a)} \le E_t^{(r,b)}.$$

□

### B.4 Proof of Proposition 3.5

*Proof.* If $p_r = 0$, then $N_t(r) = 0$ a.s. for all $t$, so $\widehat{p}_t(r) = 0$ and $L_t(r) = 0 = p_r$ a.s.; the event $\{\exists t : L_t(r) > p_r\}$ has probability 0. Assume $p_r > 0$ in the remainder.

**Step 1: Monotonicity in $q$.** For any $0 < q_1 < q_2 \le 1$, $\Lambda_r(q_2) \subseteq \Lambda_r(q_1)$ and each factor $1 + \lambda(\mathbf{1}\{X_i = r\} - q)$ is nonincreasing in $q$; hence $q \mapsto M_t(q)$ is nonincreasing.

**Step 2: NSM at $q = p_r$.** Set $M_t(q,\lambda) := \prod_{i=1}^t (1 + \lambda(\mathbf{1}\{X_i = r\} - q))$. For $\lambda \in \Lambda_r(p_r)$, every factor $1 + \lambda(\mathbf{1}\{X_i = r\} - p_r) \ge 1 - \lambda p_r > 0$, and $\mathbb{E}[1 + \lambda(\mathbf{1}\{X_i = r\} - p_r)] = 1$, so $(M_t(p_r,\lambda))_{t\ge 0}$ is a nonnegative martingale with initial value 1. The mixture

$$M_t(p_r) := \sum_{\lambda \in \Lambda_r(p_r)} v_\lambda M_t(p_r,\lambda)$$

is therefore an NSM with $M_0(p_r) = \sum_{\lambda\in\Lambda_r(p_r)} v_\lambda \le 1$ (Corollary-style argument analogous to Corollary 3.2). Ville's inequality (11) gives

$$\mathbb{P}\Big(\sup_{t\ge 0} M_t(p_r) \ge 1/\alpha_r\Big) \le \alpha_r. \tag{12}$$

**Step 3: Event inclusion.** Suppose $L_t(r) > p_r$ for some $t \ge 1$. By the definition of $L_t(r)$, there exists $q \in (p_r, \widehat{p}_t(r)]$ with $M_t(q) \ge 1/\alpha_r$. Step 1 then implies $M_t(p_r) \ge M_t(q) \ge 1/\alpha_r$. Hence

$$\{\exists t \ge 1 : L_t(r) > p_r\} \subseteq \{\sup_{t\ge 0} M_t(p_r) \ge 1/\alpha_r\},$$

and the desired bound follows from (12). □

### B.5 Proof of Proposition 3.7

*Proof.* For a fixed integer $t \geq 1$, define $f_t(u) := u^{-1}(1-u)^t$ for $u \in (0,1]$. Since $(\log f_t)'(u) = -1/u - t/(1-u) < 0$, $f_t$ is strictly decreasing on $(0,1]$.

Now fix $a \in \mathcal{A}$ with $p_a > 0$. Since $U_t$ is nonincreasing in $t$, define $t_*(a) := \inf\{t \geq 1 : p_a > U_t\}$ (possibly $+\infty$). If $t_*(a) = \infty$, then $p_a \leq U_t$ for all $t$, so the event $\{\exists t : N_t(a) = 0, p_a > U_t\}$ is empty. Otherwise, $t_*(a) < \infty$ and $p_a > U_{t_*(a)}$. For any $t \geq t_*(a)$ with $p_a > U_t$, we have $N_t(a) = 0$ only if $N_{t_*(a)}(a) = 0$ (since counts are nondecreasing), so the event $\{\exists t : N_t(a) = 0, p_a > U_t\}$ is contained in $\{N_{t_*(a)}(a) = 0\}$. By definition of $U_{t_*(a)}$, we have $f_{t_*(a)}(U_{t_*(a)}) \leq \alpha_u$. Since $p_a > U_{t_*(a)}$ and $f_{t_*(a)}$ is strictly decreasing, $f_{t_*(a)}(p_a) = p_a^{-1}(1-p_a)^{t_*(a)} < f_{t_*(a)}(U_{t_*(a)}) \leq \alpha_u$, hence

$$\mathbb{P}\big(N_{t_*(a)}(a) = 0\big) = (1-p_a)^{t_*(a)} < \alpha_u p_a.$$

A probability-weighted union bound over the countable category set gives

$$\mathbb{P}\Big(\exists a \in \mathcal{A}, \exists t \geq 1 : N_t(a) = 0 \text{ and } p_a > U_t\Big) \leq \sum_{a \in \mathcal{A}} \alpha_u p_a = \alpha_u. \qquad \square$$

### B.6 Proof of Theorem 4.1

*Proof.* Fix any $P \in H_{0,r}$. We choose a deterministic $a^\star \neq r$ such that

$$p_{a^\star} \geq p_r.$$

We decompose the Type-I event according to whether $a^\star$ has been observed by time $\tau$:

$$\{\tau < \infty\} \subseteq \underbrace{\{\tau < \infty,\ a^\star \in \mathcal{A}_\tau\}}_{E_1} \cup \underbrace{\{\tau < \infty,\ a^\star \notin \mathcal{A}_\tau\}}_{E_2}. \tag{13}$$

**Step 1 (bounding $\mathbb{P}(E_1)$).** On $E_1$, the witness $a^\star$ is observed at time $\tau$, so $\mathcal{A}_\tau \setminus \{r\} \neq \emptyset$. Since CITE stops at $\tau$, the pairwise part of the stopping rule gives

$$E_\tau^{(r,\widehat{a}_\tau)} = E_\tau \geq \alpha_{\mathrm{pw}}^{-1}.$$

By the definition of the empirical runner-up, $N_\tau(\widehat{a}_\tau) \geq N_\tau(a^\star)$. Lemma B.3 then gives

$$E_\tau^{(r,\widehat{a}_\tau)} \leq E_\tau^{(r,a^\star)}.$$

Thus

$$E_\tau^{(r,a^\star)} \geq \alpha_{\mathrm{pw}}^{-1}.$$

Since $a^\star$ is deterministic under $P$ and $p_r \leq p_{a^\star}$, Corollary 3.2 and Ville's inequality yield

$$\mathbb{P}_P(E_1) \leq \mathbb{P}_P\left(\sup_{t \geq 0} E_t^{(r,a^\star)} \geq \alpha_{\mathrm{pw}}^{-1}\right) \leq \alpha_{\mathrm{pw}}. \tag{14}$$

**Step 2 (bounding $\mathbb{P}(E_2)$).** On $E_2$, CITE stops at $\tau$, so $L_\tau(r) > U_\tau$, and $a^\star$ is unseen at $\tau$. Define the two failure events

$$F_r := \{\exists t \geq 1 : L_t(r) > p_r\},$$
$$F_u := \{\exists a \in \mathcal{A},\ \exists t \geq 1 : N_t(a) = 0 \text{ and } p_a > U_t\}.$$

Propositions 3.5 and 3.7 give $\mathbb{P}_P(F_r) \leq \alpha_r$ and $\mathbb{P}_P(F_u) \leq \alpha_u$. On $E_2 \setminus (F_r \cup F_u)$, we have $L_\tau(r) \leq p_r$ and, because $a^\star$ is unseen at $\tau$, $p_{a^\star} \leq U_\tau$. Combining these inequalities with $L_\tau(r) > U_\tau$ yields

$$p_{a^\star} \leq U_\tau < L_\tau(r) \leq p_r,$$

which contradicts $p_{a^\star} \geq p_r$. Hence, $E_2 \subseteq F_r \cup F_u$, and

$$\mathbb{P}_P(E_2) \leq \mathbb{P}_P(F_r) + \mathbb{P}_P(F_u) \leq \alpha_r + \alpha_u. \tag{15}$$

**Step 3 (union bound).** Combining (14)–(15),

$$\mathbb{P}_P(\tau < \infty) \leq \mathbb{P}_P(E_1) + \mathbb{P}_P(E_2) \leq \alpha_{\mathrm{pw}} + \alpha_r + \alpha_u \leq \varepsilon. \qquad \square$$

## C Power Analysis for CITE

Throughout this section, $X_1, X_2, \ldots$ are i.i.d. on the countable category set $\mathcal{A}$, and the fixed target $r$ is the unique mode. Write

$$p_{(2)} := \sup_{a \neq r} p_a, \qquad \delta := p_r - p_{(2)} > 0.$$

For the analysis, define the conservative competitor count

$$R_t := \max_{a \neq r} N_t(a),$$

where zero-count competitors are included in the maximum, and define

$$E_t^\circ := \sum_{\lambda \in \Lambda_{\mathrm{pw}}} w_\lambda (1+\lambda)^{N_t(r)} (1-\lambda)^{R_t}.$$

The corresponding conservative stopping time is

$$\tau^\circ := \inf \left\{ t \geq 1 : E_t^\circ \geq \alpha_{\mathrm{pw}}^{-1} \text{ and } L_t(r) > U_t \right\}.$$

This stopping time is no earlier than the implemented CITE stopping time: replacing the observed runner-up count by $R_t = \max_{a \neq r} N_t(a)$ can only make the pairwise condition harder, while the LCB–unseen condition is unchanged. Hence

$$\tau \leq \tau^\circ \quad \text{pathwise},$$

and it suffices to upper-bound $\mathbb{E}[\tau^\circ]$.

### C.1 Auxiliary lemmas for the clean stopping-time bound

**Lemma C.1** (A deterministic logarithmic inequality)**.** *Let $A > 0$, $C > 0$, and $t > 0$. If*

$$t \geq 2A \log\big((AC) \vee e\big),$$

*then*

$$t \geq A \log(Ct).$$

*Proof.* Set $u = t/A$. The assumption gives $u \geq 2\log((AC) \vee e) \geq 2$, so $\log(CA) \leq u/2$ and $\log u \leq u/2$. Therefore

$$\log(Ct) = \log(CA) + \log u \leq u = t/A.$$

□

**Lemma C.2** (Expected maximum competitor count)**.** *For every integer $t \geq 1$,*

$$\mathbb{E}[R_t] \leq t p_{(2)} + 2\sqrt{2t \log(t+1)}.$$

*Proof.* For $a \neq r$, set $f_a(x) := \mathbf{1}\{x = a\}$. Since $p_a \leq p_{(2)}$,

$$\mathbb{E}[R_t] - t p_{(2)} \leq \mathbb{E} \sup_{a \neq r} \sum_{i=1}^{t} \{f_a(X_i) - \mathbb{E} f_a(X_i)\}.$$

By symmetrization, the right-hand side is at most

$$2\mathbb{E}\mathbb{E}_\sigma \sup_{a \neq r} \sum_{i=1}^{t} \sigma_i f_a(X_i),$$

where $\sigma_i$ are i.i.d. Rademacher variables independent of the sample. Conditional on the sample, the vectors

$$(f_a(X_1), \ldots, f_a(X_t)), \qquad a \neq r,$$

have disjoint supports and hence at most $t+1$ distinct values, each with Euclidean norm at most $\sqrt{t}$. The standard exponential maximal inequality gives

$$\mathbb{E}_\sigma \sup_{a \neq r} \sum_{i=1}^{t} \sigma_i f_a(X_i) \leq \sqrt{2t \log(t+1)}.$$

Combining the displays proves the claim. □

**Lemma C.3** (LCB monotonicity and inversion)**.** *Fix $t \geq 1$. For*

$$M_t(q, \lambda) := \prod_{i=1}^{t} \big\{1 + \lambda(\mathbf{1}\{X_i = r\} - q)\big\},$$

*the following hold.*

(a) *The map $q \mapsto M_t(q)$ is nonincreasing on $(0, 1]$.*
(b) *If $M_t(q) > 1$, then $\widehat{p}_t(r) > q$.*
(c) *If $M_t(q) \geq \alpha_r^{-1}$, then $L_t(r) \geq q$.*

*Proof.* For (a), if $q_1 < q_2$, then $\Lambda_r(q_2) \subseteq \Lambda_r(q_1)$, and each remaining factor $1 + \lambda(\mathbf{1}\{X_i = r\} - q)$ is nonincreasing in $q$. Hence $M_t(q_1) \geq M_t(q_2)$.

For (b), suppose $\widehat{p}_t(r) \leq q$. For fixed $\lambda \in \Lambda_r(q)$, write

$$\phi(\lambda) := \widehat{p}_t(r) \log(1 + \lambda(1 - q)) + (1 - \widehat{p}_t(r)) \log(1 - \lambda q).$$

Then $\log M_t(q, \lambda) = t\phi(\lambda)$, $\phi(0) = 0$, $\phi'(0) = \widehat{p}_t(r) - q \leq 0$, and $\phi$ is concave on $[0, 1/q)$. Therefore $M_t(q, \lambda) \leq 1$ for every admissible $\lambda$, so

$$M_t(q) \leq \sum_{\lambda \in \Lambda_r(q)} v_\lambda \leq 1.$$

This proves the contrapositive.

For (c), if $M_t(q) \geq \alpha_r^{-1} > 1$, then (b) gives $\widehat{p}_t(r) > q$, so $q$ belongs to the set defining $L_t(r)$. Hence $L_t(r) \geq q$. □

**Lemma C.4** (Quadratic lower bound for $\log(1 + x)$)**.** *For every $x \in [-1/2, 1/2]$,*

$$\log(1 + x) \geq x - x^2.$$

*Proof.* Let $h(x) := \log(1 + x) - x + x^2$. Then $h(0) = 0$ and $h'(x) = x(1 + 2x)/(1 + x)$, which is nonpositive on $[-1/2, 0]$ and nonnegative on $[0, 1/2]$. Thus $h(x) \geq 0$ on $[-1/2, 1/2]$. □

**Proposition C.5** (Small-bet pairwise drift)**.** *Assume $\delta > 0$ and*

$$0 < \lambda \leq \frac{\delta}{2(p_r + p_{(2)})}.$$

*Then, for every $a \neq r$ with $p_a \leq p_{(2)}$,*

$$\mathbb{E}\left[\log\{1 + \lambda(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\})\}\right] \geq \frac{\lambda\delta}{2}.$$

*Proof.* Since $\delta \leq p_r + p_{(2)}$, the assumed bound gives $\lambda \leq 1/2$. Let

$$\xi := \lambda(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\}) \in [-1/2, 1/2].$$

By Lemma C.4,

$$\mathbb{E}[\log(1 + \xi)] \geq \lambda(p_r - p_a) - \lambda^2(p_r + p_a).$$

Using $p_a \leq p_{(2)}$ gives

$$\mathbb{E}[\log(1 + \xi)] \geq \lambda\delta - \lambda^2(p_r + p_{(2)}) \geq \lambda\delta/2.$$

□

### C.2 Proof of Theorem 4.3

*Proof.* We prove the bound for $\tau^\circ$; since $\tau \le \tau^\circ$ pathwise, the same expectation bound holds for the implemented CITE stopping time.

*Step 1: pairwise condition.* Let $\lambda := \lambda_{\mathrm{pw}} \in [\delta/8, \delta/4]$ be the grid point from Condition 4.2, with weight $w_\lambda$, and set

$$B_{\mathrm{pw}} := \log\{1/(\alpha_{\mathrm{pw}} w_\lambda)\}.$$

For this fixed $\lambda$, define

$$S_t := N_t(r)\log(1+\lambda) + R_t \log(1-\lambda).$$

Since $E_t^\circ \ge w_\lambda \exp(S_t)$, the pairwise condition $E_t^\circ \ge \alpha_{\mathrm{pw}}^{-1}$ holds whenever $S_t \ge B_{\mathrm{pw}}$.

Using Lemma C.2 and $\log(1-\lambda) < 0$,

$$\mathbb{E}[S_t] \ge t\{p_r \log(1+\lambda) + p_{(2)} \log(1-\lambda)\} - C\lambda\sqrt{t\log(t+1)}.$$

Because $p_r + p_{(2)} \le 1$, $\lambda \le \delta/4$, and $\log(1+x) \ge x - x^2$ on $[-1/2, 1/2]$,

$$p_r \log(1+\lambda) + p_{(2)} \log(1-\lambda) \ge \lambda\delta - \lambda^2(p_r + p_{(2)}) \ge c\delta^2$$

for a universal constant $c > 0$. Hence

$$\mathbb{E}[S_t] \ge c_0\delta^2 t - C_0\delta\sqrt{t\log(t+1)}.$$

Changing one observation changes both $N_t(r)$ and $R_t$ by at most one, so $S_t$ is $C_1\delta$-Lipschitz in each sample. McDiarmid's inequality gives

$$\mathbb{P}\{S_t < \mathbb{E}[S_t] - x\} \le \exp\left\{-\frac{c_1 x^2}{t\delta^2}\right\}.$$

Taking $x = C_2\delta\sqrt{t\log(et)}$ and using Lemma C.1, we obtain, for a universal constant $C$,

$$\mathbb{P}\left(E_t^\circ < \alpha_{\mathrm{pw}}^{-1}\right) \le (et)^{-2}, \qquad t \ge T_{\mathrm{pw}} := C\frac{B_{\mathrm{pw}} + \log(1/\delta)}{\delta^2}. \tag{16}$$

*Step 2: LCB–unseen condition.* Set $q_0 := p_r/2$. Condition 4.2 gives $\lambda_r \in [1/32, 1/16] \subset \Lambda_r(q_0)$ with weight $v_{\lambda_r}$. The unseen bound satisfies $U_t \le p_r/4$ once

$$t \ge Cp_r^{-1}\log\{1/(\alpha_u p_r)\}.$$

Let

$$A_t := \left\{\frac{3}{4}p_r t \le N_t(r) \le \frac{5}{4}p_r t\right\}.$$

Bernstein's inequality gives

$$\mathbb{P}(A_t^c) \le 2\exp(-c_2 p_r t).$$

On $A_t$, the single LCB factor

$$M_t(q_0, \lambda_r) = \prod_{i=1}^{t}\{1 + \lambda_r(\mathbf{1}\{X_i = r\} - q_0)\}$$

satisfies

$$\log M_t(q_0, \lambda_r) \ge c_3 p_r t$$

for a universal constant $c_3 > 0$. Therefore

$$\widetilde{M}_t(q_0) \ge v_{\lambda_r} M_t(q_0, \lambda_r) \ge \alpha_r^{-1}$$

whenever

$$t \ge Cp_r^{-1}\log\{1/(\alpha_r v_{\lambda_r})\}.$$

By Lemma C.3, this implies $L_t(r) \ge q_0 > U_t$. Combining this with the Bernstein tail and Lemma C.1, we obtain

$$\mathbb{P}\{L_t(r) \le U_t\} \le (et)^{-2} \tag{17}$$

for all

$$t \geq T_{\mathrm{rare}} := C\frac{\log(1/\alpha_r) + \log(1/\alpha_u) + \log(1/v_{\lambda_r}) + \log(1/p_r)}{p_r}.$$

*Step 3: combine.* Let

$$T_{\mathrm{cl}} := \max\{T_{\mathrm{pw}}, T_{\mathrm{rare}}\}.$$

For all $t \geq T_{\mathrm{cl}}$,

$$\mathbb{P}(\tau^\circ > t) \leq 2(et)^{-2}.$$

The tail-sum identity gives

$$\mathbb{E}[\tau^\circ] \leq T_{\mathrm{cl}} + 1.$$

Finally, substituting $\alpha_{\mathrm{pw}} = \alpha_r = \alpha_u = \varepsilon/3$ and using the assumption that $w_\lambda$ and $v_{\lambda_r}$ are bounded below by constants,

$$\mathbb{E}[\tau] \leq \mathbb{E}[\tau^\circ] = O\left(\frac{\log(1/\varepsilon) + \log(1/\delta)}{\delta^2} + \frac{\log(1/\varepsilon) + \log(1/p_r)}{p_r}\right),$$

with constants independent of $\delta, p_r$, and $|\mathcal{A}|$. □

### C.3 Oracle pairwise growth rate

**Proposition C.6** (Closed form for the oracle pairwise growth rate)**.** *For $\lambda \in (0,1)$, define*

$$\rho(\lambda) := p_r \log(1+\lambda) + p_{(2)} \log(1-\lambda).$$

*Then $\rho$ is strictly concave on $(0,1)$. If $p_{(2)} > 0$, its unique maximizer is*

$$\lambda^\star = \frac{p_r - p_{(2)}}{p_r + p_{(2)}} = \frac{\delta}{p_r + p_{(2)}},$$

*and the maximal value is*

$$\rho^\star := \rho(\lambda^\star) = p_r \log\frac{2p_r}{p_r + p_{(2)}} + p_{(2)} \log\frac{2p_{(2)}}{p_r + p_{(2)}}.$$

*If $p_{(2)} = 0$, then $\rho(\lambda) = p_r \log(1+\lambda)$ is increasing on $(0,1)$ and*

$$\sup_{0<\lambda<1} \rho(\lambda) = p_r \log 2.$$

*Proof.* We have

$$\rho'(\lambda) = \frac{p_r}{1+\lambda} - \frac{p_{(2)}}{1-\lambda}, \qquad \rho''(\lambda) = -\frac{p_r}{(1+\lambda)^2} - \frac{p_{(2)}}{(1-\lambda)^2} < 0.$$

Thus $\rho$ is strictly concave. When $p_{(2)} > 0$, solving $\rho'(\lambda) = 0$ gives $\lambda^\star = (p_r - p_{(2)})/(p_r + p_{(2)})$, and substituting this value gives the stated formula for $\rho^\star$. The case $p_{(2)} = 0$ is immediate. □

**Lemma C.7** (Stopped KL identity for the swapped alternative)**.** *Assume $p_{(2)} > 0$. Fix*

$$a^\star \in \arg\max_{a \neq r} p_a,$$

*and let $Q$ be obtained from $P$ by swapping the masses of $r$ and $a^\star$:*

$$q_r = p_{(2)}, \qquad q_{a^\star} = p_r, \qquad q_a = p_a \quad \text{for } a \notin \{r, a^\star\}.$$

*Let $P^\tau$ and $Q^\tau$ denote the laws of the stopped experiment $(\tau, X_1, \ldots, X_\tau)$ under $P$ and $Q$. If $\mathbb{E}_P[\tau] < \infty$, then*

$$D(P^\tau \| Q^\tau) = \mathbb{E}_P[\tau]\delta \log\frac{p_r}{p_{(2)}}.$$

*Proof.* Let $\ell(x) := \log(P(x)/Q(x))$. Since $\ell$ is bounded (both $P$ and $Q$ have positive masses on the same support), the stopping factor cancels in the likelihood ratio $dP^\tau/dQ^\tau = \prod_{i\le\tau}(P(X_i)/Q(X_i))$, and Wald's identity for i.i.d. bounded summands yields

$$D(P^\tau\|Q^\tau) = \mathbb{E}_P\left[\sum_{i=1}^{\tau}\ell(X_i)\right] = \mathbb{E}_P[\tau]\mathbb{E}_P[\ell(X_1)].$$

A direct calculation gives $\mathbb{E}_P[\ell(X_1)] = p_r\log(p_r/p_{(2)}) + p_{(2)}\log(p_{(2)}/p_r) = \delta\log(p_r/p_{(2)})$, completing the proof. □

**Corollary C.8** (Information-theoretic lower bound)**.** *Assume $p_{(2)} > 0$. Let $\varepsilon$ be the Type-I error level of a sequential test for $H_{0,r}$, and assume that under the alternative $P$ one has*

$$\mathbb{P}_P(\tau < \infty) = 1 \qquad \textit{and} \qquad \mathbb{E}_P[\tau] < \infty.$$

*Then*

$$\mathbb{E}_P[\tau] \ge \frac{\log(1/\varepsilon)}{\delta\log(p_r/p_{(2)})}.$$

*In particular, if $\delta/p_r \to 0$, then*

$$\mathbb{E}_P[\tau] \ge \frac{p_r\log(1/\varepsilon)}{\delta^2}\big(1+o(1)\big).$$

*Proof.* Let $Q$ be the swap from Lemma C.7, so $Q \in H_{0,r}$ and Type-I validity gives $Q(\tau<\infty)\le\varepsilon$. Set $A := \{\tau<\infty\}$, so $P(A) = 1$. By data processing followed by Lemma C.7,

$$\log\frac{1}{\varepsilon} \le \mathrm{kl}\big(P(A),Q(A)\big) \le D(P^\tau\|Q^\tau) = \mathbb{E}_P[\tau]\delta\log(p_r/p_{(2)}),$$

which is the first claim. For the small-gap expansion, $\delta\log(p_r/p_{(2)}) = \delta\log(1+\delta/p_{(2)}) = \delta^2/p_r + o(\delta^2/p_r)$ as $\delta/p_r\to 0$. □

**Lemma C.9** (Forward I-projection of $P$ onto $H_{0,r}$)**.** *Assume $p_{(2)} > 0$ and pick any $a^\star \in \arg\max_{a\ne r} p_a$ (this set may have ties). Every forward I-projection*

$$P^\star \in \arg\min_{Q\in H_{0,r}} D(P\|Q)$$

*arises by midpointing $r$ with one of the strongest competitors; for the choice $a^\star$ above, the corresponding projection is the* midpoint distribution

$$p^\star_r = p^\star_{a^\star} = \frac{p_r+p_{(2)}}{2}, \qquad p^\star_a = p_a \quad \textit{for } a\notin\{r,a^\star\},$$

*and the per-sample KL divergence is*

$$D(P\|P^\star) = p_r\log\frac{2p_r}{p_r+p_{(2)}} + p_{(2)}\log\frac{2p_{(2)}}{p_r+p_{(2)}}.$$

*In particular, as $\delta/p_r\to 0$,*

$$D(P\|P^\star) = \frac{\delta^2}{4p_r}\big(1+o(1)\big).$$

*Proof.* $H_{0,r} = \bigcup_{a\ne r}\{q_r\le q_a\}$ is a (possibly countable) union of half-spaces, hence not itself convex. We minimise $D(P\|\cdot)$ over each face $\{q_r\le q_a\}$ separately (on each face $D(P\|\cdot)$ is strictly convex and strictly positive since $P\notin H_{0,r}$), and then take the smallest of the per-face minima. The optimum on each face lies on its boundary $\{q_r = q_a\}$. We organise the proof in three steps.

*Step 1 (per-pair midpoint projection).* For each $a\ne r$ with $p_a < p_r$, the I-projection of $P$ onto the affine subspace $\{q_r = q_a\}$ is given in closed form by

$$q_r = q_a = \frac{p_r+p_a}{2}, \qquad q_b = p_b \quad \text{for } b\notin\{r,a\}. \tag{18}$$

The Lagrangian on the simplex with multiplier $\mu$ has first-order conditions $q_b = p_b/\mu$ and $q_r = q_a = (p_r + p_a)/(2\mu)$; the normalisation $\sum_b q_b = 1/\mu = 1$ forces $\mu = 1$, yielding (18), and the positive-definite Hessian $\mathrm{diag}(p/q^2)$ ensures uniqueness. The KL value of this projection is

$$\Phi(p_a) := p_r \log \frac{2p_r}{p_r + p_a} + p_a \log \frac{2p_a}{p_r + p_a}. \tag{19}$$

*Step 2 (the optimal competitor is $a^\star$).* Differentiating (19) gives $\Phi'(p_a) = \log \frac{2p_a}{p_r+p_a} \le 0$ on $(0, p_r)$, so $\Phi$ is non-increasing in $p_a$ and is minimised over $\{p_a : a \neq r\}$ at $p_a = p_{(2)} = p_{a^\star}$. After midpointing on $a^\star$, the remaining masses satisfy $q_b = p_b \le p_{(2)} < (p_r + p_{(2)})/2 = q_r = q_{a^\star}$ for $b \notin \{r, a^\star\}$, confirming that the candidate lies in $H_{0,r}$. Hence midpointing $r$ with the chosen $a^\star$ yields a global forward I-projection $P^\star$ with $D(P\|P^\star) = \Phi(p_{(2)})$ (in case of ties at $p_{(2)}$ each choice of strongest competitor gives a distinct minimiser with the same KL value), proving the first two displays.

*Step 3 (small-gap expansion).* Write $q^\star := (p_r + p_{(2)})/2 = p_r - \delta/2$ and $x := \delta/(2q^\star) \in (0, 1)$, so $p_r/q^\star = 1 + x$ and $p_{(2)}/q^\star = 1 - x$. The closed form factorises as $\Phi(p_{(2)}) = q^\star f(x)$, where

$$f(x) := (1 + x)\log(1 + x) + (1 - x)\log(1 - x).$$

$f$ is even with $f''(x) = 2/(1 - x^2) = 2 + 2x^2 + O(x^4)$, hence $f(x) = x^2 + x^4/6 + O(x^6)$ about 0. Substituting,

$$D(P\|P^\star) = q^\star x^2 + O(q^\star x^4) = \frac{\delta^2}{4q^\star} + O\left(\frac{\delta^4}{q^{\star 3}}\right),$$

and $q^\star = p_r(1 - \delta/(2p_r)) = p_r(1 + o(1))$ as $\delta/p_r \to 0$ yields the third display. □

**Lemma C.10** (Stopped KL identity for the I-projection)**.** *Under the assumptions of Lemma C.9, if $\mathbb{E}_P[\tau] < \infty$, then*

$$D(P^\tau\|(P^\star)^\tau) = \mathbb{E}_P[\tau] D(P\|P^\star).$$

*Proof.* $P$ and $P^\star$ are mutually absolutely continuous on $\mathcal{A}$ (since $q^\star \in (0, 1)$ when $p_r \in (0, 1)$, and other masses agree), and the log-likelihood $\ell(x) := \log(P(x)/P^\star(x))$ is bounded. The stopping-event factor cancels in the likelihood ratio, so Fubini and independence give $D(P^\tau\|(P^\star)^\tau) = \mathbb{E}_P[\tau]\mathbb{E}_P[\ell(X_1)] = \mathbb{E}_P[\tau] D(P\|P^\star)$, exactly as in Lemma C.7. □

**Corollary C.11** (Sharp information-theoretic lower bound)**.** *Under the hypotheses of Corollary C.8,*

$$\mathbb{E}_P[\tau] \ge \frac{\log(1/\varepsilon)}{D(P\|P^\star)} = \frac{4p_r \log(1/\varepsilon)}{\delta^2}\big(1 + o(1)\big) \qquad (\delta/p_r \to 0).$$

*Proof.* Apply the same data-processing argument as in Corollary C.8 with $Q = P^\star$ in place of the swap distribution. Since $P^\star \in H_{0,r}$ (Lemma C.9), Type-I validity gives $P^\star(\tau < \infty) \le \varepsilon$, and the data-processing inequality combined with Lemma C.10 yields $\log(1/\varepsilon) \le \mathbb{E}_P[\tau] D(P\|P^\star)$. The asymptotic constant follows from Lemma C.9. □

### C.4 Information-Theoretic Lower Bound for the Unseen-Category Component

**Lemma C.12** (Stopped KL identity for a hidden-competitor alternative)**.** *Fix a distribution $P$ on a countable category set $\mathcal{A}$ with $p_r \in (0, 1)$, and let $h \in \mathcal{A}$ satisfy $h \neq r$ and $p_h = 0$. Define a second distribution $Q$ by*

$$q_r = \frac{p_r}{2}, \qquad q_h = \frac{p_r}{2}, \qquad q_a = p_a \quad \text{for } a \notin \{r, h\}.$$

*Let $P^\tau$ and $Q^\tau$ denote the laws of the stopped experiment $(\tau, X_1, \ldots, X_\tau)$ under $P$ and $Q$. If $\mathbb{E}_P[\tau] < \infty$, then*

$$D(P^\tau\|Q^\tau) = \mathbb{E}_P[\tau] p_r \log 2.$$

*Proof.* By the same Wald argument as Lemma C.7, $D(P^\tau\|Q^\tau) = \mathbb{E}_P[\tau] \cdot \mathbb{E}_P[\ell(X_1)]$ with $\ell(x) := \log(P(x)/Q(x))$, where we set $\ell(h) := 0$ (an arbitrary finite value on the $P$-null event $\{X = h\}$ that does not affect any $P$-expectation). By construction $\ell(r) = \log 2$ and $\ell(x) = 0$ for $x \neq r$, so $\mathbb{E}_P[\ell(X_1)] = p_r \log 2$. □

**Corollary C.13** (Information-theoretic lower bound for the unseen-category component)**.** *Fix $p \in (0,1)$ and $\delta \in (0,p)$, and assume the category set $\mathcal{A}$ contains at least $\lceil (1-p)/(p-\delta) \rceil + 2$ distinct non-target categories (so that an unused "hidden" label is available; this holds automatically on countably infinite category sets). Define*

$$\mathcal{P}(p,\delta) := \left\{ P : p_r = p, p_{(2)} \le p - \delta \right\}.$$

*Let $\tau$ be any stopping time corresponding to a sequential test such that*

$$\sup_{Q \in H_{0,r}} \mathbb{P}_Q(\tau < \infty) \le \varepsilon,$$

*and assume that, for every $P \in \mathcal{P}(p,\delta)$,*

$$\mathbb{P}_P(\tau < \infty) = 1 \qquad \textit{and} \qquad \mathbb{E}_P[\tau] < \infty.$$

*Then*

$$\sup_{P \in \mathcal{P}(p,\delta)} \mathbb{E}_P[\tau] \ge \frac{\log(1/\varepsilon)}{p \log 2}.$$

*Proof.* Pick $m \ge (1-p)/(p-\delta)$ distinct categories $b_1, \dots, b_m, h \in \mathcal{A} \setminus \{r\}$ and set $P(r) = p$, $P(h) = 0$, $P(b_j) = (1-p)/m$, so $P \in \mathcal{P}(p,\delta)$. The hidden-competitor distribution $Q(r) = Q(h) = p/2$, $Q(b_j) = (1-p)/m$ lies in $H_{0,r}$, so $Q(\tau < \infty) \le \varepsilon$. Data processing combined with Lemma C.12 then gives $\log(1/\varepsilon) \le D(P^\tau \| Q^\tau) = \mathbb{E}_P[\tau] p \log 2$. □

*Proof of Theorem 4.4.* Fix $p \in (0, 1/2]$, $\delta \in (0,p)$, and a level-$\varepsilon$ sequential certifier $\tau'$ satisfying the assumptions of the corollary. For notational simplicity, write $\tau := \tau'$ within this proof, and let

$$\mathcal{P}(p,\delta) := \left\{ P : p_r = p, \quad p_{(2)} \le p - \delta \right\}.$$

We combine two distinct least favorable subfamilies of $\mathcal{P}(p,\delta)$.

**Step 1: Discrimination-hard instance.** Because $p \le 1/2$, the quantity $1 - 2p + \delta$ is nonnegative. Choose an integer

$$m \ge \frac{1 - 2p + \delta}{p - \delta},$$

and distinct categories

$$a^\star, b_1, \dots, b_m \in \mathcal{A} \setminus \{r\}.$$

Define $P_{\mathrm{pw}}$ by

$$P_{\mathrm{pw}}(r) = p, \qquad P_{\mathrm{pw}}(a^\star) = p - \delta, \qquad P_{\mathrm{pw}}(b_j) = \frac{1 - 2p + \delta}{m} \quad (j = 1, \dots, m),$$

and $P_{\mathrm{pw}}(a) = 0$ otherwise. Then $P_{\mathrm{pw}} \in \mathcal{P}(p,\delta)$ and its runner-up satisfies $p_{(2)} = p - \delta > 0$. Corollary C.8 therefore gives

$$\mathbb{E}_{P_{\mathrm{pw}}}[\tau] \ge \frac{\log(1/\varepsilon)}{\delta \log\big(p/(p-\delta)\big)}.$$

If $\delta \le p/2$, then

$$\log \frac{p}{p-\delta} = -\log\left(1 - \frac{\delta}{p}\right) \le \frac{\delta/p}{1 - \delta/p} \le \frac{2\delta}{p},$$

whence

$$\mathbb{E}_{P_{\mathrm{pw}}}[\tau] \ge \frac{p \log(1/\varepsilon)}{2\delta^2}.$$

**Step 2: Exploration-hard instance.** Corollary C.13 provides another distribution $P_{\mathrm{un}} \in \mathcal{P}(p,\delta)$ such that

$$\mathbb{E}_{P_{\mathrm{un}}}[\tau] \ge \frac{\log(1/\varepsilon)}{p \log 2}.$$

**Step 3: Combine the two bounds.** If $\delta \le p/2$, then $\max\{x,y\} \ge (x+y)/2$ applied to the two lower bounds gives the result directly. If $\delta > p/2$, then $p/\delta^2 \le 4/p$, so $p/\delta^2 + 1/p \le 5/p$, and Corollary C.13 alone gives $\sup_P \mathbb{E}_P[\tau] \ge \log(1/\varepsilon)/(p \log 2) \ge (5 \log 2)^{-1} \log(1/\varepsilon)(p/\delta^2 + 1/p)$. Since $(5 \log 2)^{-1} > 1/4$, the same constant works in both cases. □

## D Detailed Power Analysis for the W-CITE

This appendix contains the proofs for the confidence-weighted extension in Section 5.2. We keep the notation parallel to the unweighted analysis, replacing $(p_r, p_{(2)}, \delta)$ by $(\mu_r, \mu_\star, \Delta_w)$ where appropriate.

**Lemma D.1** (Expected maximum weighted competitor penalty)**.** *Fix $\lambda \in (0,1)$ and define $g_\lambda(w) := -\log(1-\lambda w)$ for $w \in [0,1]$. For each competitor $a \neq r$, let $H_t^{(\lambda)}(a) := \sum_{i=1}^t g_\lambda(W_i)\mathbf{1}\{X_i = a\}$, $v_\lambda(a) := \mathbb{E}[g_\lambda(W)\mathbf{1}\{X = a\}]$, and set $R_t^{(\lambda)} := \max_{a\neq r} H_t^{(\lambda)}(a)$, $v_\lambda^\star := \sup_{a\neq r} v_\lambda(a)$. Then, for every integer $t \geq 1$,*

$$\mathbb{E}[R_t^{(\lambda)}] \leq t v_\lambda^\star + 2\sqrt{2}\big(-\log(1-\lambda)\big)\sqrt{t\log(t+1)}.$$

**Lemma D.2** (Weighted LCB inversion at a fixed time)**.** *Fix an integer $t \geq 1$ and a number $q \in (0,1]$.* (a) *If $\widetilde{M}_t(q) > 1$, then $\widehat{\mu}_t(r) > q$.* (b) *If $\widetilde{M}_t(q) \geq 1/\alpha_r$, then $\widetilde{L}_t(r) \geq q$.*

### D.1 Weighted pairwise NSM property (Section 5, item (i))

*Proof.* Fix $a \neq r$ and $\lambda \in (0,1)$. Since $W_i \in [0,1]$, each factor

$$1 + \lambda W_i\big(\mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\}\big)$$

is nonnegative. Under $\mu_r \leq \mu_a$,

$$\mathbb{E}\left[1 + \lambda W_i\big(\mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\}\big) \mid \mathcal{F}_{i-1}\right] = 1 + \lambda(\mu_r - \mu_a) \leq 1,$$

using independence of $(X_i, W_i)$ from $\mathcal{F}_{i-1}$. Therefore $(\widetilde{E}_t^{(r,a)}(\lambda))_{t\geq 0}$ is an NSM. □

### D.2 Weighted mixture NSM (Section 5, item (i))

*Proof.* For each fixed $a \neq r$ and each $\lambda \in \Lambda_{\mathrm{pw}}$, $(\widetilde{E}_t^{(r,a)}(\lambda))_{t\geq 0}$ is an NSM under $\mu_r \leq \mu_a$ by the preceding proof. Since the weights $(w_\lambda)$ are nonnegative and sum to at most one,

$$\widetilde{E}_t^{(r,a)} = \sum_{\lambda\in\Lambda_{\mathrm{pw}}} w_\lambda \widetilde{E}_t^{(r,a)}(\lambda)$$

is also an NSM with initial value at most one. Ville's inequality gives

$$\mathbb{P}_P\left(\sup_{t\geq 0} \widetilde{E}_t^{(r,a)} \geq \alpha^{-1}\right) \leq \alpha.$$

□

### D.3 Weighted LCB monotonicity (Section 5, item (ii))

*Proof.* The proof is identical to the monotonicity part of Lemma C.3: $\Lambda_r(q_2) \subseteq \Lambda_r(q_1)$ and each factor $1 + \lambda(W_i\mathbf{1}\{X_i = r\} - q)$ is pointwise larger when $q$ is smaller, so $\widetilde{M}_t(q_1) \geq \widetilde{M}_t(q_2)$. □

### D.4 Weighted LCB validity (Section 5, item (ii))

*Proof.* If $\mu_r = 0$, then $W_i\mathbf{1}\{X_i = r\} = 0$ almost surely for all $i$, so $\widehat{\mu}_t(r) = 0$ and $\widetilde{L}_t(r) = 0$ almost surely. Hence $\{\exists t : \widetilde{L}_t(r) > \mu_r\}$ is empty. Assume $\mu_r > 0$ in the remainder.

The proof parallels that of Proposition 3.5 (Appendix B.4), with the following substitutions: $\mathbf{1}\{X_i = r\}$ becomes $W_i\mathbf{1}\{X_i = r\}$, $p_r$ becomes $\mu_r$, $\widehat{p}_t(r)$ becomes $\widehat{\mu}_t(r)$, and the mixture weights $(w_\lambda)$ become $(v_\lambda)$. At the true value $q = \mu_r$, one has $\mathbb{E}[W_i\mathbf{1}\{X_i = r\} - \mu_r] = 0$, so each $\widetilde{M}_t(\mu_r, \lambda)$ is a nonnegative martingale with unit initial value, and Ville's inequality gives $\mathbb{P}(\sup_t \widetilde{M}_t(\mu_r) \geq 1/\alpha_r) \leq \alpha_r$. The monotonicity from the weighted LCB monotonicity property then implies $\{\exists t : \widetilde{L}_t(r) > \mu_r\} \subseteq \{\sup_t \widetilde{M}_t(\mu_r) \geq 1/\alpha_r\}$. □

### D.5 Weighted unseen bound validity (Section 5, item (iii))

*Proof.* Since $\mu_a \le p_a$ for every $a \in \mathcal{A}$,

$$\Big\{\exists a, \exists t : N_t(a) = 0 \text{ and } \mu_a > U_t\Big\} \subseteq \Big\{\exists a, \exists t : N_t(a) = 0 \text{ and } p_a > U_t\Big\}.$$

The probability of the latter event is at most $\alpha_u$ by Proposition 3.7. □

### D.6 Proof of Theorem 5.2

*Proof.* The proof parallels that of Theorem 4.1 (Appendix B.6), with $p_a$ replaced by $\mu_a$ throughout and Lemma B.3 replaced by the all-competitors check (runner-up monotonicity fails for weighted observations).

Fix $P \in H_{0,r}^{(w)}$. If $\mathcal{A} \setminus \{r\} = \emptyset$, then the weighted null is vacuous. Otherwise, since

$$\sum_{a \in \mathcal{A}} \mu_a = \mathbb{E}[W] \le 1,$$

Lemma B.1 implies that any positive value of $\mu_\star = \sup_{a \ne r} \mu_a$ is attained. If $\mu_\star > 0$, choose $a^\star \ne r$ with $\mu_{a^\star} = \mu_\star \ge \mu_r$. If $\mu_\star = 0$, then $P \in H_{0,r}^{(w)}$ implies $\mu_r = 0$, and any $a^\star \ne r$ is a valid witness with $\mu_{a^\star} \ge \mu_r$. Decompose

$$\{\tau^{(w)} < \infty\} \subseteq \underbrace{\{\tau^{(w)} < \infty,\ a^\star \in \mathcal{A}_{\tau^{(w)}}\}}_{E_1} \cup \underbrace{\{\tau^{(w)} < \infty,\ a^\star \notin \mathcal{A}_{\tau^{(w)}}\}}_{E_2}.$$

**Step 1 (bounding $\mathbb{P}(E_1)$).** On $E_1$, the weighted pairwise stopping condition at $\tau^{(w)}$ applied to *every* observed competitor gives $\widetilde{E}_{\tau^{(w)}}^{(r,a^\star)} \ge 1/\alpha_{\mathrm{pw}}$. The process $(\widetilde{E}_t^{(r,a^\star)})_{t \ge 0}$ is an NSM under $\mu_r \le \mu_{a^\star}$ by the weighted mixture NSM property (Section 5, item (i)), with initial value $\le 1$. Ville's inequality (11) yields $\mathbb{P}(E_1) \le \alpha_{\mathrm{pw}}$.

**Step 2 (bounding $\mathbb{P}(E_2)$).** Define

$$F_r := \{\exists t \ge 1 : \widetilde{L}_t(r) > \mu_r\}, \quad F_u := \{\exists a \in \mathcal{A},\ \exists t \ge 1 : N_t(a) = 0,\ \mu_a > U_t\}.$$

Weighted LCB validity gives $\mathbb{P}(F_r) \le \alpha_r$, and weighted unseen-bound validity gives $\mathbb{P}(F_u) \le \alpha_u$. On $E_2 \setminus (F_r \cup F_u)$, the LCB–unseen part of the weighted stopping rule gives $\widetilde{L}_{\tau^{(w)}}(r) > U_{\tau^{(w)}}$; combined with $F_r^c$ and $F_u^c$ (recall $a^\star$ is unseen at $\tau^{(w)}$),

$$\mu_{a^\star} \le U_{\tau^{(w)}} < \widetilde{L}_{\tau^{(w)}}(r) \le \mu_r,$$

contradicting $\mu_{a^\star} \ge \mu_r$. Hence $E_2 \subseteq F_r \cup F_u$ and $\mathbb{P}(E_2) \le \alpha_r + \alpha_u$.

**Step 3 (union bound).** $\mathbb{P}(\tau^{(w)} < \infty) \le \mathbb{P}(E_1) + \mathbb{P}(E_2) \le \alpha_{\mathrm{pw}} + \alpha_r + \alpha_u \le \varepsilon$. □

### D.7 Proof of Theorem 5.4: almost-sure stopping part

*Proof.* Lemma D.3 (a) gives $\gamma_{\mathrm{pw}} := \inf_{a \ne r} \mathbb{E}[\log(1 + \lambda_{\mathrm{pw}} W(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\}))] \ge \lambda_{\mathrm{pw}} \Delta_w / 2 > 0$, and Lemma D.3 (c) with $q_0 = \mu_r/2$ gives $\nu_r := \mathbb{E}[\log(1 + \lambda_r(W\mathbf{1}\{X = r\} - q_0))] \ge \lambda_r \mu_r / 4 > 0$. We now show that these positive drifts force $\mathbb{P}(\tau^{(w)} < \infty) = 1$.

Fix $\lambda := \lambda_{\mathrm{pw}}$, and let $w_{\mathrm{pw}}$ denote its mixture weight in Section 5, item (i).

**Step 1: Pairwise condition.** For each $i$, set $A_i := \log(1 + \lambda W_i)\mathbf{1}\{X_i = r\}$ and $B_i(a) := -\log(1 - \lambda W_i)\mathbf{1}\{X_i = a\}$, so that

$$\log \widetilde{E}_t^{(r,a)}(\lambda) = \textstyle\sum_{i=1}^t A_i - \sum_{i=1}^t B_i(a). \tag{20}$$

Write $u := \mathbb{E}[A_1]$, $v_a := \mathbb{E}[B_1(a)]$, and $v^\star := \sup_{a \ne r} v_a$. The definition of $\gamma_{\mathrm{pw}}$ gives $u - v^\star \ge \gamma_{\mathrm{pw}} > 0$.

We control $\max_{a\neq r} t^{-1}\sum_i B_i(a)$ on the countable category set by a tail-truncation argument. Since $\sum_{a\neq r} v_a \le -\log(1-\lambda) < \infty$, for each $m \ge 1$ there is a finite $F_m \subset \mathcal{A}\setminus\{r\}$ with $\sum_{a\notin F_m} v_a < 1/m$. Define the tail process $R_t^{(m)} := \sum_i -\log(1-\lambda W_i)\mathbf{1}\{X_i \notin F_m \cup \{r\}\}$. By the Strong Law of Large Numbers, $t^{-1}\sum_i A_i \to u$, $t^{-1}\sum_i B_i(a) \to v_a$ for each fixed $a$, and $t^{-1}R_t^{(m)} \to \sum_{a\notin F_m} v_a < 1/m$, all almost surely. On the probability-one intersection of these countably many events, $\limsup_t \max_{a\neq r} t^{-1}\sum_i B_i(a) \le \max\{v^\star, 1/m\}$ for every $m$; letting $m\to\infty$ gives the limsup $\le v^\star$. Combining with (20),

$$\liminf_t \min_{a\neq r} t^{-1}\log \widetilde{E}_t^{(r,a)}(\lambda) \ge u - v^\star \ge \gamma_{\mathrm{pw}} > 0.$$

Since $\mathcal{A}_t\setminus\{r\} \subseteq \mathcal{A}\setminus\{r\}$ and $\widetilde{E}_t^{(r,a)} \ge w_{\mathrm{pw}}\widetilde{E}_t^{(r,a)}(\lambda)$, the positive liminf of the normalized log e-process implies there exists a.s. a finite $T'_{\mathrm{pw}}$ after which $\min_{a\in\mathcal{A}_t\setminus\{r\}} \widetilde{E}_t^{(r,a)} \ge 1/\alpha_{\mathrm{pw}}$.

**Step 2: Weighted LCB condition.** Fix $\lambda := \lambda_r$ with mixture weight $v_{\lambda_r} > 0$ and set $C_i := \log(1 + \lambda(W_i\mathbf{1}\{X_i = r\} - q_0))$; each $C_i$ is bounded since $\lambda < 1/q_0$ (Lemma D.3(c)) and $W_i\mathbf{1}\{X_i = r\} \in [0,1]$. Part (c) of Lemma D.3 gives $\mathbb{E}[C_i] = \nu_r \ge \lambda\mu_r/4 > 0$, so by the SLLN $t^{-1}\log \widetilde{M}_t(q_0,\lambda) = t^{-1}\sum_{i=1}^t C_i \to \nu_r$ a.s., hence $\widetilde{M}_t(q_0,\lambda)\to\infty$ and $\widetilde{M}_t(q_0) \ge v_{\lambda_r}\widetilde{M}_t(q_0,\lambda) \ge 1/\alpha_r$ eventually a.s. An independent SLLN on $W_i\mathbf{1}\{X_i = r\}$ gives $\widehat{\mu}_t(r) \to \mu_r > q_0$, so $\widehat{\mu}_t(r) \ge q_0$ eventually, and the weighted LCB construction then yields $\widetilde{L}_t(r) \ge q_0$ for all large $t$ almost surely. Let $T'_r$ denote the resulting (random, a.s. finite) eventual time after which both $\widetilde{M}_t(q_0) \ge 1/\alpha_r$ and $\widetilde{L}_t(r) \ge q_0$ hold.

**Step 3: Unseen upper bound.** The sequence $(U_t)_{t\ge1}$ is deterministic and $U_t \downarrow 0$ as $t\to\infty$. Since $q_0 > 0$, there exists a deterministic index $T_u$ such that

$$U_t < q_0 \qquad \forall t \ge T_u.$$

**Step 4: Combine.** Choose $t$ large enough that $t \ge T'_{\mathrm{pw}} \vee T'_r \vee T_u$. Then the pairwise threshold is met, $\widetilde{L}_t(r) \ge q_0 > U_t$, and both stopping conditions hold, giving $\mathbb{P}(\tau^{(w)} < \infty) = 1$. □

**Lemma D.3** (Weighted small-bet drift bounds)**.** *Assume $\Delta_w > 0$.*

(a) ***Pairwise:*** *if $\lambda \in \Lambda_{\mathrm{pw}}$ satisfies $\lambda \le \Delta_w/\big(2(\mu_r+\mu_\star)\big)$, then*

$$\inf_{a\neq r} \mathbb{E}\big[\log\big(1 + \lambda W(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\})\big)\big] \ge \lambda\Delta_w/2 > 0.$$

(b) ***LCB:*** *for any $q_0 \in (0,\mu_r)$ and any $\lambda \in \Lambda_r$ with $\lambda \le (\mu_r - q_0)/(2\mu_r)$, one has $\lambda < 1/q_0$ and*

$$\mathbb{E}\big[\log\big(1 + \lambda(W\mathbf{1}\{X = r\} - q_0)\big)\big] \ge \lambda(\mu_r - q_0)/2 > 0.$$

(c) ***Universal LCB grid point:*** *with $q_0 := \mu_r/2$ in part (b), the condition reduces to $\lambda_r \in (0, 1/4]$, and*

$$\mathbb{E}\big[\log\big(1 + \lambda_r(W\mathbf{1}\{X = r\} - \mu_r/2)\big)\big] \ge \lambda_r\mu_r/4 > 0.$$

*Proof.* All three parts follow from the quadratic lower bound $\log(1+\xi) \ge \xi - \xi^2$ on $[-1/2, 1/2]$ (Lemma C.4), exactly as in Proposition C.5, with two additional ingredients: $W \in [0,1]$ implies $W^2 \le W$, and the squared indicators satisfy $(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\})^2 = \mathbf{1}\{X = r\} + \mathbf{1}\{X = a\}$.

**(a)** Set $\xi_a := \lambda W(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\})$, so $|\xi_a| \le \lambda \le 1/2$. Taking expectations of $\log(1+\xi_a) \ge \xi_a - \xi_a^2$ and using $W^2 \le W$,

$$\mathbb{E}[\log(1+\xi_a)] \ge \lambda(\mu_r - \mu_a) - \lambda^2(\mu_r + \mu_a) \ge \lambda\Delta_w - \lambda^2(\mu_r + \mu_\star).$$

The hypothesis gives $\lambda^2(\mu_r + \mu_\star) \le \lambda\Delta_w/2$, hence $\mathbb{E}[\log(1+\xi_a)] \ge \lambda\Delta_w/2$ uniformly in $a$.

**(b)** Since $\mu_r \le 1$, $(\mu_r - q_0)/(2\mu_r) < 1/2 < 1/q_0$, so the hypothesis implies $\lambda < 1/2$ and $\lambda < 1/q_0$. Setting $Y := W\mathbf{1}\{X = r\}$ and $\xi := \lambda(Y - q_0)$, the same quadratic bound and $Y^2 \le Y$ yield

$$\mathbb{E}[\log(1+\xi)] \ge \lambda(\mu_r - q_0) - \lambda^2\mu_r \ge \lambda(\mu_r - q_0)/2.$$

**(c)** Specializing (b) to $q_0 = \mu_r/2$ gives $(\mu_r - q_0)/(2\mu_r) = 1/4$, so $\lambda_r \le 1/4$ verifies the hypothesis and the conclusion follows. □

### D.8 Proof of Lemma D.1

*Proof.* The proof follows the same symmetrization argument as Lemma C.2, applied to the bounded function class $\{f_a(x, w) := g_\lambda(w)\mathbf{1}\{x = a\} : a \neq r\}$ in place of $\{\mathbf{1}\{x = a\}\}$. The key modifications are:

- The centering term $tp_{(2)}$ becomes $tv^\star_\lambda$, where $v^\star_\lambda := \sup_{a\neq r} \mathbb{E}[g_\lambda(W)\mathbf{1}\{X = a\}]$.
- The per-coordinate bound $\|v_a\|_\infty \leq 1$ becomes $\|v_a\|_\infty \leq g_\lambda(1) = -\log(1 - \lambda)$, so the sub-Gaussian parameter scales by $-\log(1 - \lambda)$.
- The Rademacher complexity bound remains $O(\sqrt{t \log(t + 1)})$ because there are still at most $t + 1$ distinct vectors in the function class evaluated on $t$ samples.

Optimizing the Donsker–Varadhan parameter $s$ as in Lemma C.2 yields the stated bound. □

### D.9 Proof of Lemma D.2

*Proof.* **(a)** Assume $\widehat{\mu}_t(r) \leq q$. Set $Y_i := W_i\mathbf{1}\{X_i = r\}$. By Jensen's inequality (concavity of $\log$), $t^{-1} \log \widetilde{M}_t(q, \lambda) \leq \log(1 + \lambda(\widehat{\mu}_t(r) - q)) \leq 0$ for every $\lambda \in \Lambda_r(q)$. Summing gives $\widetilde{M}_t(q) \leq \sum v_\lambda \leq 1$, contradicting $\widetilde{M}_t(q) > 1$.

**(b)** If $\widetilde{M}_t(q) \geq 1/\alpha_r > 1$, part (a) gives $\widehat{\mu}_t(r) > q$, so $q \in (0, \widehat{\mu}_t(r)]$ and $q$ belongs to the set defining $\widetilde{L}_t(r)$; hence $\widetilde{L}_t(r) \geq q$. □

### D.10 Proof of Theorem 5.4: expected stopping-time bound

*Proof.* We use the same structure as the proof of Theorem 4.3, with $(p_r, p_{(2)}, \delta)$ replaced by $(\mu_r, \mu_\star, \Delta_w)$.

*Step 1: pairwise condition.* Let $\lambda := \lambda_{\mathrm{pw}} \in [\Delta_w/8, \Delta_w/4]$ and let $w_\lambda$ be its mixture weight. Set

$$B^{(w)}_{\mathrm{pw}} := \log\{1/(\alpha_{\mathrm{pw}} w_\lambda)\}.$$

For each competitor $a \neq r$, write

$$A_i := \log(1 + \lambda W_i)\mathbf{1}\{X_i = r\}, \qquad B_i(a) := -\log(1 - \lambda W_i)\mathbf{1}\{X_i = a\}.$$

The relevant fixed-$\lambda$ log pairwise statistic is

$$S^{(w)}_t := \sum_{i=1}^t A_i - \max_{a\neq r} \sum_{i=1}^t B_i(a).$$

If $S^{(w)}_t \geq B^{(w)}_{\mathrm{pw}}$, then the weighted all-competitors pairwise condition holds at time $t$.

Lemma D.3 gives the drift lower bound

$$\inf_{a\neq r} \mathbb{E} \log\{1 + \lambda W(\mathbf{1}\{X = r\} - \mathbf{1}\{X = a\})\} \geq \lambda\Delta_w/2 \geq c\Delta^2_w.$$

Combining this with Lemma D.1 yields

$$\mathbb{E}[S^{(w)}_t] \geq c_0\Delta^2_w t - C_0\Delta_w\sqrt{t \log(t + 1)}.$$

Changing one observation changes $S^{(w)}_t$ by at most $C_1\Delta_w$, so McDiarmid's inequality and Lemma C.1 imply that, for a universal constant $C$,

$$\mathbb{P}\{\text{weighted pairwise condition fails at time } t\} \leq (et)^{-2}$$

whenever

$$t \geq T^{(w)}_{\mathrm{pw}} := C\frac{\log\{1/(\alpha_{\mathrm{pw}} w_\lambda)\} + \log(1/\Delta_w)}{\Delta^2_w}.$$

*Step 2: LCB–unseen condition.* Set $q_0 := \mu_r/2$. The grid condition gives $\lambda_r \in [1/32, 1/16] \subset \Lambda_r(q_0)$ with weight $v_{\lambda_r}$. Let

$$Y_i := W_i\mathbf{1}\{X_i = r\}, \qquad A_t := \left\{\frac{3}{4}\mu_r t \leq \sum_{i=1}^t Y_i \leq \frac{5}{4}\mu_r t\right\}.$$

Since $Y_i \in [0,1]$ and $\mathbb{E}Y_i = \mu_r$, Bernstein's inequality gives

$$\mathbb{P}(A_t^c) \le 2\exp(-c\mu_r t).$$

On $A_t$, the single weighted LCB factor satisfies

$$\log \widetilde{M}_t(q_0, \lambda_r) = \sum_{i=1}^{t} \log\{1 + \lambda_r(Y_i - q_0)\} \ge c'\mu_r t$$

for a universal constant $c' > 0$. Thus

$$\widetilde{M}_t(q_0) \ge v_{\lambda_r} \widetilde{M}_t(q_0, \lambda_r) \ge \alpha_r^{-1}$$

once

$$t \ge C\mu_r^{-1} \log\{1/(\alpha_r v_{\lambda_r})\}.$$

Moreover, the unchanged unseen bound satisfies $U_t \le \mu_r/4 < q_0$ once

$$t \ge C\mu_r^{-1} \log\{1/(\alpha_u \mu_r)\}.$$

Combining these bounds with Lemma C.1, we obtain

$$\mathbb{P}\{\widetilde{L}_t(r) \le U_t\} \le (et)^{-2}$$

for all

$$t \ge T_{\text{rare}}^{(w)} := C\frac{\log(1/\alpha_r) + \log(1/\alpha_u) + \log(1/v_{\lambda_r}) + \log(1/\mu_r)}{\mu_r}.$$

*Step 3: combine.* Let

$$T_{\text{cl}}^{(w)} := \max\{T_{\text{pw}}^{(w)}, T_{\text{rare}}^{(w)}\}.$$

For all $t \ge T_{\text{cl}}^{(w)}$,

$$\mathbb{P}(\tau^{(w)} > t) \le 2(et)^{-2}.$$

The tail-sum identity gives

$$\mathbb{E}[\tau^{(w)}] \le T_{\text{cl}}^{(w)} + 1.$$

Substituting $\alpha_{\text{pw}} = \alpha_r = \alpha_u = \varepsilon/3$ and using the selected-weight lower bounds yields

$$\mathbb{E}[\tau^{(w)}] = O\left(\frac{\log(1/\varepsilon) + \log(1/\Delta_w)}{\Delta_w^2} + \frac{\log(1/\varepsilon) + \log(1/\mu_r)}{\mu_r}\right).$$

□

**Corollary D.4** (Independent confidence scores)**.** *Suppose $W$ is independent of $X$ and let $\bar{w} := \mathbb{E}[W] > 0$. If $r$ is the unweighted unique mode with gap $\delta := p_r - p_{(2)} > 0$, then*

$$\mu_r = \bar{w}p_r, \qquad \mu_\star = \bar{w}p_{(2)}, \qquad \Delta_w = \bar{w}\delta.$$

*Thus Theorem 5.4 applies whenever the weighted grid condition holds, equivalently when*

$$\lambda_{\text{pw}} \in \left[\frac{\bar{w}\delta}{8}, \frac{\bar{w}\delta}{4}\right] \quad \textit{and} \quad \lambda_r \in [1/32, 1/16]$$

*are available on the grids.*

### D.11 Proof of Corollary D.4

*Proof.* Independence gives

$$\mu_a = \mathbb{E}[W\mathbf{1}\{X = a\}] = \bar{w}p_a \qquad \forall a \in \mathcal{A}.$$

Therefore

$$\mu_r = \bar{w}p_r, \qquad \mu_\star = \bar{w}p_{(2)}, \qquad \Delta_w = \bar{w}\delta.$$

Substituting these identities into the grid condition and the rate in Theorem 5.4 proves the claim. □

**Proposition D.5** (Recovery of the unweighted CITE)**.** *Suppose $W_i \equiv 1$ almost surely. Then $\mu_a = p_a$ for every $a \in \mathcal{A}$; $\widetilde{E}_t^{(r,a)}(\lambda) = (1+\lambda)^{N_t(r)}(1-\lambda)^{N_t(a)}$; $\widehat{\mu}_t(r) = \widehat{p}_t(r)$; and the weighted stopping rule agrees with the all-competitors version of the unweighted CITE, with runner-up monotonicity reducing the pairwise check to the empirical runner-up.*

### D.12 Proof of Proposition D.5

*Proof.* **Claim (i): Mass equality.** If $W_i \equiv 1$ almost surely, then

$$\mu_a = \mathbb{E}[W\mathbf{1}\{X = a\}] = \mathbb{E}[\mathbf{1}\{X = a\}] = p_a \quad \forall a \in \mathcal{A}.$$

**Claim (ii): Pairwise e-process recovery.** Under $W_i \equiv 1$, the weighted e-process simplifies:

$$\widetilde{Z}_i^{(r,a)} = W_i(\mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\}) = \mathbf{1}\{X_i = r\} - \mathbf{1}\{X_i = a\}.$$

Thus

$$\widetilde{E}_t^{(r,a)}(\lambda) = \prod_{i=1}^{t}\big(1+\lambda\widetilde{Z}_i^{(r,a)}\big) = \prod_{i=1}^{t}\big(1+\lambda(\mathbf{1}\{X_i = r\}-\mathbf{1}\{X_i = a\})\big) = (1+\lambda)^{N_t(r)}(1-\lambda)^{N_t(a)}.$$

**Claim (iii): Empirical mean equality.**

$$\widehat{\mu}_t(r) = \frac{1}{t}\sum_{i=1}^{t} W_i\mathbf{1}\{X_i = r\} = \frac{1}{t}\sum_{i=1}^{t}\mathbf{1}\{X_i = r\} = \widehat{p}_t(r).$$

Substituting $W_i \equiv 1$ into the weighted LCB e-process yields

$$\widetilde{M}_t(q,\lambda) = \prod_{i=1}^{t}(1 + \lambda(\mathbf{1}\{X_i = r\} - q)),$$

which is identical to the unweighted LCB e-process after dropping the tilde.

**Claim (iv): Stopping rule recovery.** From (ii), in the unweighted specialization the weighted pairwise condition

$$\widetilde{E}_t^{(r,a)} = \sum_{\lambda} w_\lambda \widetilde{E}_t^{(r,a)}(\lambda) \geq 1/\alpha_{\mathrm{pw}} \quad \text{for all } a \in \mathcal{A}_t \setminus \{r\}$$

checks the same pairwise e-values as the unweighted all-competitors CITE. By Lemma B.3, those e-values are monotone in the competitor count, so among observed competitors the minimum is attained at the empirical runner-up. Thus the all-competitors pairwise check reduces to a single empirical runner-up, recovering the original unweighted fixed-target CITE. □

## E Comparison with Martingale Majority Certificates

### E.1 The MMC Hypothesis

The top-$m$ MMC [Cordero-Encinar and Duncan, 2025] simultaneously tests the leader against $m-1$ pairwise competitors and the aggregated residual mass. The alternative certified upon stopping is

$$\mathcal{H}_{\mathrm{MMC},m} := \{P : p_{(1)} > p_{(i)} \forall i = 2,\dots,m\} \cap \{P : p_{(1)} > s_m\}, \qquad s_m := \sum_{j>m} p_{(j)}. \tag{21}$$

**Proposition E.1** (Strict inclusion)**.** *Suppose the target $r$ coincides with the (unique) modal label $p_r = p_{(1)}$, equivalently $r$ is the unique maximiser of $a \mapsto p_a$. Then for every $m \geq 2$, $\mathcal{H}_{\mathrm{MMC},m} \subsetneq H_{1,r}$; without the unique-mode-target identification, this strict inclusion must be read with $H_{1,r}$ replaced by $\bigcup_{r' \in \mathcal{A}} H_{1,r'}$.*

**Remark E.2** (The diffuse-tail regime)**.** *Distributions in $H_{1,r} \setminus \mathcal{H}_{\mathrm{MMC},m}$ have a unique mode that does not dominate the aggregate tail: $p_{(1)} > p_{(2)}$ but $p_{(1)} \leq s_m$. Several of our LLM self-consistency settings (Section 6) fall into this regime, with the mode-answer carrying 20–30% of the mass and thousands of alternative answers sharing the remainder. For each fixed $m$, the MMC controls its certification rate at level $\alpha$ on any such $P$: its residual-mass channel at that $m$ tests the null $\{p_{(1)} \leq s_m\}$, which is true under $P$, so the corresponding e-process cannot grow past the rejection threshold with probability $> \alpha$. An MMC variant that tries multiple $m$ in parallel pays a Bonferroni-style cost over the active tuples (Appendix E.3).*

### E.2 The Switching-Null Challenge

At each round $n$, the MMC selects the empirical leader $A_{n-1}$ and runner-up $B_{n-1}$ (both $\mathcal{F}_{n-1}$-measurable), maintaining e-processes: $e_n^{\mathrm{run}}$ (leader vs. runner-up) and $e_n^{\mathrm{oth}}$ (leader vs. residual mass). The per-round nulls are:

$$R_n := \{p_{A_{n-1}} \le p_{B_{n-1}}\}, \quad O_n := \Big\{p_{A_{n-1}} \le \sum\nolimits_{j \notin \{A_{n-1}, B_{n-1}\}} p_j\Big\}. \tag{22}$$

Theorem 3.1 of Cordero-Encinar and Duncan [2025] establishes that $(e_n^{\mathrm{run}})$ is an NSM under the all-rounds null $H_0^{\mathrm{run}} := \bigcap_{n \ge 1} R_n$, and similarly for the residual-mass channel. Corollary 3.2 proves

$$\sup_{P \in H_0} \mathbb{P}_P(N < \infty) \le \varepsilon, \qquad H_0 := H_0^{\mathrm{run}} \cup H_0^{\mathrm{oth}}. \tag{23}$$

The intersection-union structure (stop requires *both* channels, null is a *union*) gives $\varepsilon$ without Bonferroni.

The challenge lies in bridging from (23) to the adaptive misclassification guarantee $\mathbb{P}(A_{\tau-1} \ne a^\star) \le \varepsilon$ for the certified pre-round leader (cf. Proposition E.3 below). At any fixed time $n$, misclassification of the pre-round leader satisfies

$$\{A_{n-1} \ne a^\star\} \subseteq R_n \cup O_n. \tag{24}$$

However, Corollary 3.2 controls the all-rounds event

$$H_0 = \Big(\bigcap_{t \ge 1} R_t\Big) \cup \Big(\bigcap_{t \ge 1} O_t\Big),$$

which is logically distinct from the existential per-round inclusion:

$$(\forall t R_t) \vee (\forall t O_t) \quad \not\Longleftarrow \quad \exists n (R_n \vee O_n). \tag{25}$$

Bridging this gap requires a *switching-null reduction* showing that misclassification at $\tau$ is contained in the all-rounds null. Since the labels $(A_{n-1}, B_{n-1})$ are data-dependent and may switch while e-processes accumulate evidence without restart, this reduction is nontrivial but feasible. It introduces Bonferroni costs absent from CITE.

### E.3 Tuple-Indexed Construction

To handle the switching-null challenge, *index e-processes by tuple*: let

$$\mathcal{T} := \big\{(m, a, b_1, \ldots, b_{m-1}) : m \ge 2,\ a \in \mathcal{A},\ b_1, \ldots, b_{m-1} \in \mathcal{A} \setminus \{a\} \text{ pairwise distinct}\big\}$$

denote the (countable) set of admissible tuples and assign each $\ell = (m, a, b_1, \ldots, b_{m-1}) \in \mathcal{T}$ its own e-process, with multiplicative factor 1 on rounds where the current tuple differs from $\ell$.

**Proposition E.3** (Tuple-indexed repair)**.** *Assume the data-generating distribution has a unique mode $a^\star \in \arg\max_{a \in \mathcal{A}} p_a$. Let $\{\alpha_\ell\}_{\ell \in \mathcal{T}}$ satisfy $\sum_\ell \alpha_\ell \le \varepsilon$. For each $\ell$, maintain separate tuple-indexed e-processes $E_t^{\ell,i}$ (pairwise, slot $i$) and $E_t^{\ell,\mathrm{oth}}$ (residual mass), active only when $L_{t-1} = \ell$. Define*

$$\tau := \inf\Big\{n : \exists \ell \in \mathcal{T} \text{ s.t. } L_{n-1} = \ell, E_n^{\ell,i} \ge 1/\alpha_\ell \forall i, E_n^{\ell,\mathrm{oth}} \ge 1/\alpha_\ell\Big\}.$$

*Let $A_{\tau-1}$ denote the leader component of the active tuple $L_{\tau-1}$ (the pre-round empirical leader at the certifying time). Then $\mathbb{P}(\tau < \infty, A_{\tau-1} \ne a^\star) \le \varepsilon$.*

**The Bonferroni cost.** Each tuple receives error allocation $\alpha_\ell$, yielding threshold $1/\alpha_\ell$. With $\pi_\ell := \alpha_\ell/\varepsilon$:

$$\log \frac{1}{\alpha_\ell} = \underbrace{\log \frac{1}{\varepsilon}}_{\text{base}} + \underbrace{\log \frac{1}{\pi_\ell}}_{\text{multiplicity penalty}}. \tag{26}$$

Under uniform allocation over $K$ tuples, the multiplicity penalty is $\log K$. With $L_t$ observed categories: $K_{m,t} = L_t \binom{L_t - 1}{m-1} = O(L_t^m)$, yielding additional sample complexity of order $m \log L_t / I_\ell$, where $I_\ell$ denotes the per-observation KL information.

**Contrast with CITE.** CITE avoids this multiplicity cost: with $r$ fixed a priori, Ville's inequality applies to a single e-process with threshold $1/\alpha_{\mathrm{pw}}$ and no multiplicity correction.

### E.4 Structural Comparison Summary

| | **Tuple-Indexed MMC** | **CITE (ours)** |
|---|---|---|
| Hypothesis | $p_{(1)} > p_{(i)}$ for all $i \le m$ and $p_{(1)} > s_m$ | $p_{(1)} > p_{(2)}$ (exact unique mode) |
| Target selection | Adaptive (tuple may switch) | Fixed target $r$ a priori |
| Multiplicity cost | $\log K_{m,t}$ threshold inflation | None |
| Diffuse-tail power | $\le \alpha$ ($p_{(1)} \le s_m \Rightarrow$ null true; Ville bound) | Full (pairwise margin only) |
| Unseen control | Aggregated into residual mass | Explicit: $L_t(r) > U_t$ |

Table 1: Structural comparison of the tuple-indexed MMC and CITE.

### E.5 Proof of Proposition E.1

*Proof. Inclusion* $\mathcal{H}_{\mathrm{MMC},m} \subseteq H_{1,r}$. Any $P \in \mathcal{H}_{\mathrm{MMC},m}$ satisfies $p_{(1)} > p_{(i)}$ for every $i \in \{2,\dots,m\}$, and $p_{(1)} > s_m \ge p_{(j)}$ for every $j > m$; combining, $p_r = p_{(1)}$ strictly dominates every other category, i.e. $P \in H_{1,r}$.

*Strict inclusion.* Fix $m \ge 2$ and $K \ge 2m$, and consider the perturbed-uniform distribution

$$p_{(1)} := \frac{1}{K} + \eta, \qquad p_{(i)} := \frac{1}{K} - \frac{\eta}{K-1} \quad (i \ge 2), \qquad \eta \in (0, \tfrac{1}{K}),$$

so that $\sum_i p_{(i)} = 1$ and $p_{(1)} > p_{(2)}$. Then $P \in H_{1,r}$ (since $r$ is strictly the unique mode), while

$$s_m = \sum_{j>m} p_{(j)} = (K-m)\left(\tfrac{1}{K} - \tfrac{\eta}{K-1}\right) \ge \tfrac{K-m}{K} - \eta \ge \tfrac{1}{2} - \eta > \tfrac{1}{K} + \eta = p_{(1)}$$

for sufficiently small $\eta$ (using $K \ge 2m$ and $K \ge 2$). Hence $P \notin \mathcal{H}_{\mathrm{MMC},m}$, proving strict inclusion. □

### E.6 Proof of Proposition E.3

*Proof.* Fix a tuple $\ell = (a, b_1, \dots, b_{m-1}) \in \mathcal{T}$ with leader $a$ and slots $b_1, \dots, b_{m-1}$. Define its stopping time and misclassification event

$$\tau_\ell := \inf\big\{n \ge 1 : L_{n-1} = \ell,\ E_n^{\ell,i} \ge 1/\alpha_\ell\ \forall i,\ E_n^{\ell,\mathrm{oth}} \ge 1/\alpha_\ell\big\}, \quad G_\ell := \{\tau = \tau_\ell < \infty,\ a \ne a^\star\}.$$

Suppose $G_\ell$ occurs. Since $a \ne a^\star$ and $a^\star$ lies in $\mathcal{A}$, exactly one of the following holds:

**Case (i): $a^\star = b_i$ for some $i \in \{1,\dots,m-1\}$.** Then the pairwise null for slot $i$ is $R^{(\ell,i)} := \{p_a \le p_{b_i}\}$, which is *true* because $p_a < p_{a^\star} = p_{b_i}$ (as $a^\star$ is the unique mode). Under $R^{(\ell,i)}$ the pairwise e-process $(E_n^{\ell,i})_{n\ge 0}$ is an NSM with initial value $\le 1$, so Ville's inequality gives $\mathbb{P}(E_{\tau_\ell}^{\ell,i} \ge 1/\alpha_\ell) \le \alpha_\ell$. Hence $\mathbb{P}(G_\ell \cap \text{Case (i)}) \le \alpha_\ell$.

**Case (ii): $a^\star \notin \{a, b_1, \dots, b_{m-1}\}$.** Then the residual-mass null $O^{(\ell)} := \{p_a \le \sum_{j \notin \{a,b_1,\dots,b_{m-1}\}} p_j\}$ is true, since the sum on the right side includes $p_{a^\star} > p_a$. Under $O^{(\ell)}$, $(E_n^{\ell,\mathrm{oth}})_{n\ge 0}$ is an NSM with initial value $\le 1$, and Ville gives $\mathbb{P}(E_{\tau_\ell}^{\ell,\mathrm{oth}} \ge 1/\alpha_\ell) \le \alpha_\ell$. Hence $\mathbb{P}(G_\ell \cap \text{Case (ii)}) \le \alpha_\ell$.

Combining both cases, $\mathbb{P}(G_\ell) \le \alpha_\ell$.

**Mutual exclusion across tuples.** At the stopping time $\tau$, only the tuple $L_{\tau-1}$ has $L_{\tau-1} = \ell$, so the events $\{\tau = \tau_\ell, L_{\tau_\ell - 1} = \ell\}$ are disjoint across $\ell$. Therefore

$$\mathbb{P}(\tau < \infty,\ A_{\tau-1} \ne a^\star) = \sum_{\ell \in \mathcal{T}} \mathbb{P}(G_\ell) \le \sum_{\ell \in \mathcal{T}} \alpha_\ell \le \varepsilon. \qquad \square$$

# F Additional Theoretical Results

This appendix collects additional results used to interpret and extend the main theory: a sharp oracle pairwise benchmark (F.1), an adaptive geometric grid (F.2), a growth-rate optimality result for the oracle pairwise bet and its finite mixture (F.3), and a top-$k$ extension (F.4).

## F.1 Sharp oracle pairwise benchmark

The main text compares the stopping-time upper bound with the minimax lower bound at the level of rates. Here we isolate a sharper constant-level statement for an oracle pairwise comparison between the target and its true runner-up. This result is not a sharp-constant theorem for the full CITE stopping time; rather, it shows that the pairwise e-process has the correct leading constant when both the true runner-up and the oracle betting parameter are available.

**Proposition F.1** (Identity and small-gap expansion of $\rho^\star$)**.** *The oracle pairwise growth rate $\rho^\star := \max_{\lambda\in(0,1)}\big[p_r\log(1+\lambda)+p_{(2)}\log(1-\lambda)\big]$ (Proposition C.6) coincides exactly with the forward I-projection KL of $P$ onto $H_{0,r}$:*

$$\rho^\star = D(P\|P^\star) = p_r\log\frac{2p_r}{p_r+p_{(2)}} + p_{(2)}\log\frac{2p_{(2)}}{p_r+p_{(2)}}. \tag{27}$$

*In particular, by Lemma C.9,*

$$\rho^\star = \frac{\delta^2}{2(p_r+p_{(2)})} + O\left(\frac{\delta^4}{(p_r+p_{(2)})^3}\right) = \frac{\delta^2}{4p_r}\big(1+o(1)\big).$$

*Proof.* Substituting $\lambda^\star = (p_r - p_{(2)})/(p_r + p_{(2)})$ from Proposition C.6 into the definition of $\rho^\star$ yields the closed form of $D(P\|P^\star)$ in Lemma C.9, since $1 \pm \lambda^\star = 2p_r/(p_r + p_{(2)})$ and $2p_{(2)}/(p_r + p_{(2)})$ respectively. Both displays now follow directly from Lemma C.9 after substituting $q^\star = (p_r + p_{(2)})/2$. □

**Theorem F.2** (Sharp constant for the oracle pairwise benchmark)**.** *Assume $p_{(2)} > 0$, $p_r \in (0, 1/2]$, and $\delta/p_r \to 0$. Suppose*

$$\lambda^\star := \frac{\delta}{p_r+p_{(2)}} \in \Lambda_{\mathrm{pw}}, \qquad w_{\lambda^\star} > 0.$$

*Let $a^\star \in \arg\max_{a\neq r} p_a$, and let $P^\star \in H_{0,r}$ be the forward I-projection of $P$ from Lemma C.9. Define*

$$\rho^\star := \max_{\lambda\in(0,1)}\{p_r\log(1+\lambda)+p_{(2)}\log(1-\lambda)\}.$$

*Then*

$$\rho^\star = D(P\|P^\star) = \frac{\delta^2}{4p_r}\{1+o(1)\}.$$

*For the oracle pairwise hitting time*

$$\sigma^\star_{\mathrm{pw}} := \inf\left\{t\geq 1 : w_{\lambda^\star}E_t^{(r,a^\star)}(\lambda^\star) \geq \alpha_{\mathrm{pw}}^{-1}\right\},$$

*we have*

$$\mathbb{E}_P[\sigma^\star_{\mathrm{pw}}] \leq \frac{\log\{1/(\alpha_{\mathrm{pw}}w_{\lambda^\star})\}}{\rho^\star}\{1+o(1)\}.$$

*Conversely, any level-$\varepsilon$ sequential certifier $\tau'$ for $H_{0,r}$ with $\mathbb{P}_P(\tau'<\infty)=1$ and $\mathbb{E}_P[\tau']<\infty$ obeys*

$$\mathbb{E}_P[\tau'] \geq \frac{\log(1/\varepsilon)}{\rho^\star}.$$

*In particular, if $\alpha_{\mathrm{pw}} = c\varepsilon$ for a constant $c\in(0,1]$ and $\log(1/w_{\lambda^\star}) = o\{\log(1/\varepsilon)\}$, the oracle pairwise upper bound and the information lower bound share the leading constant $4p_r/\delta^2$.*

*Proof.* Proposition F.1 gives

$$\rho^\star = D(P\|P^\star) = \frac{\delta^2}{4p_r}\{1+o(1)\}.$$

*(a) Upper bound.* Define

$$Y_i := \log\{1+\lambda^\star Z_i^{(r,a^\star)}\}.$$

Then the $Y_i$ are i.i.d., bounded, and $\mathbb{E}_P[Y_i] = \rho^\star > 0$. Let

$$B := \log\{1/(\alpha_{\mathrm{pw}} w_{\lambda^\star})\}.$$

At the hitting time $\sigma^\star_{\mathrm{pw}}$,

$$\sum_{i=1}^{\sigma^\star_{\mathrm{pw}}} Y_i \geq B.$$

Moreover, the crossing overshoot is at most the largest positive increment, so

$$\sum_{i=1}^{\sigma^\star_{\mathrm{pw}}} Y_i \leq B + \log(1+\lambda^\star).$$

Since the increments are bounded and the drift is positive, Wald's identity gives

$$\rho^\star \mathbb{E}_P[\sigma^\star_{\mathrm{pw}}] = \mathbb{E}_P\left[\sum_{i=1}^{\sigma^\star_{\mathrm{pw}}} Y_i\right] \leq B + \log(1+\lambda^\star).$$

As $\lambda^\star \to 0$ in the small-gap regime,

$$\mathbb{E}_P[\sigma^\star_{\mathrm{pw}}] \leq \frac{\log\{1/(\alpha_{\mathrm{pw}} w_{\lambda^\star})\} + o(1)}{\rho^\star}.$$

Substituting the expansion of $\rho^\star$ proves the oracle upper bound.

*(b) Lower bound.* Apply the sequential change-of-measure argument to the I-projection $P^\star \in H_{0,r}$. Type-I validity gives $P^\star(\tau' < \infty) \leq \varepsilon$, while $\mathbb{P}_P(\tau' < \infty) = 1$. By data processing and Lemma C.10,

$$\log(1/\varepsilon) \leq \mathbb{E}_P[\tau'] D(P\|P^\star) = \mathbb{E}_P[\tau']\rho^\star.$$

Hence

$$\mathbb{E}_P[\tau'] \geq \frac{\log(1/\varepsilon)}{\rho^\star}.$$

*(c) Ratio.* With $\alpha_{\mathrm{pw}} = c\varepsilon$,

$$\frac{\log(1/(\alpha_{\mathrm{pw}} w_{\lambda^\star}))}{\log(1/\varepsilon)} = 1 + \frac{\log(1/(c w_{\lambda^\star}))}{\log(1/\varepsilon)} \to 1$$

using $\alpha_{\mathrm{pw}} = c\varepsilon$ and $\log(1/w_{\lambda^\star}) = o\{\log(1/\varepsilon)\}$. □

**Remark F.3** (Scope of the oracle result)**.** *Theorem F.2 is an oracle benchmark for comparing $r$ with its true runner-up. It is not a sharp-constant theorem for the full CITE stopping time, which also uses the empirical runner-up, the LCB–unseen check, and a finite grid.*

### F.2 Adaptive geometric grid

Condition 4.2 requires a grid point in $[\delta/8, \delta/4]$, which depends on the unknown gap $\delta$. A geometric grid gives a uniform version of the same rate over all $\delta \geq \delta_0$.

**Condition F.4** (Geometric grid)**.** *For a fixed minimum gap $\delta_0 \in (0,1]$, set*

$$\Lambda^{\mathrm{geo}}_{\mathrm{pw}} := \{2^{-k} : 1 \leq k \leq K\}, \qquad K := \lceil \log_2(8/\delta_0) \rceil,$$

*with uniform weights $w_\lambda = 1/K$.*

**Theorem F.5** (Adaptive grid rate)**.** *Suppose Assumption 2.1 holds and CITE uses $\Lambda_{\mathrm{pw}}^{\mathrm{geo}}$ in the pairwise component. Keep the LCB-grid and selected-weight assumptions of Theorem 4.3. Then, uniformly over all $P$ with modal gap $\delta \geq \delta_0$,*

$$\mathbb{E}[\tau] = O\left(\frac{\log(1/\varepsilon) + \log(\log(8/\delta_0) \vee e) + \log(1/\delta)}{\delta^2} + \frac{\log(1/\varepsilon) + \log(1/p_r)}{p_r}\right),$$

*with constants independent of $\delta, p_r$, and $|\mathcal{A}|$.*

*Proof.* Fix $P$ with $\delta \geq \delta_0$ and set

$$k^\star := \lfloor \log_2(8/\delta) \rfloor, \qquad \lambda := 2^{-k^\star}.$$

Then $\lambda \in [\delta/8, \delta/4]$ and, since $\delta \geq \delta_0$, we have $k^\star \leq K$. Hence $\lambda \in \Lambda_{\mathrm{pw}}^{\mathrm{geo}}$.

The proof of Theorem 4.3 applies with this grid point. The only change is the pairwise mixture weight:

$$\log \frac{1}{\alpha_{\mathrm{pw}} w_\lambda} = \log \frac{1}{\alpha_{\mathrm{pw}}} + \log K = \log \frac{1}{\alpha_{\mathrm{pw}}} + O\{\log(\log(8/\delta_0) \vee e)\}.$$

Substituting this into the pairwise part of the stopping-time bound, and keeping the LCB–unseen part unchanged, proves the display. □

**Remark F.6.** *The geometric grid removes the need to know $\delta$ exactly, but it still requires a lower design scale $\delta_0$. Removing this design parameter would require a separate adaptive-betting analysis, which we do not pursue here.*

### F.3 Growth-rate optimality of the oracle pairwise bet

We record a class-restricted growth-rate optimality property for the oracle pairwise bet, in the GRO/GROW sense of Grünwald et al. [2024]. The comparison is restricted to predictable linear bets in the pairwise statistic.

**Definition F.7** (Class-restricted GRO/GROW)**.** *Fix a pair $(r, a)$ with $r \neq a$, a class $\mathcal{Q}$ of alternatives, and a class $\mathcal{E}$ of e-processes for the pairwise null $H_0^{(a)} = \{P : p_r \leq p_a\}$. An element $E^\star \in \mathcal{E}$ is* GRO/GROW within $\mathcal{E}$ with respect to $\mathcal{Q}$ *if, for every fixed horizon $T$,*

$$E^\star \in \arg\max_{E \in \mathcal{E}} \inf_{Q \in \mathcal{Q}} \mathbb{E}_Q[\log E_T].$$

**Theorem F.8** (GRO bet and mixture near-optimality)**.** *Let $\mathcal{E}_Z$ denote the class of processes $E_t = \prod_{i=1}^t (1 + \lambda_i Z_i^{(r,a)})$ with predictable bets $\lambda_i \in [0, 1)$ measurable with respect to $\mathcal{F}_{i-1} := \sigma(X_1, \ldots, X_{i-1})$. Every $E \in \mathcal{E}_Z$ is an e-process for the composite null $H_0^{(a)} = \{P : p_r \leq p_a\}$ in the sense of Definition F.7.*

*Fix a simple alternative $Q$ with $\mathbb{P}_Q(X = r) = p$, $\mathbb{P}_Q(X = a) = q$, and $p > q > 0$ (the boundary case $q = 0$ is degenerate: $\lambda^\star = 1$ lies on the boundary of the open betting interval $[0, 1)$, so the supremum is approached only in the limit $\lambda \uparrow 1$ and is not attained in $\mathcal{E}_Z$), and define*

$$\lambda^\star := \frac{p - q}{p + q}, \qquad \rho^\star := p \log(1 + \lambda^\star) + q \log(1 - \lambda^\star).$$

(a) ***GRO bet.*** *For every $T \geq 1$ and every $E \in \mathcal{E}_Z$, $\mathbb{E}_Q[\log E_T] \leq T\rho^\star$, with equality if and only if $\lambda_i = \lambda^\star$ holds $Q$-almost surely for every $i \leq T$. In particular, $E_t^{(r,a)}(\lambda^\star)$ is the unique growth-rate maximizer within $\mathcal{E}_Z$ under the simple alternative $Q$.*

(b) ***Mixture near-optimality.*** *For any finite grid $\Lambda_{\mathrm{pw}} \subset (0, 1)$ with $\lambda^\star \in \Lambda_{\mathrm{pw}}$ and positive weights $(w_\lambda)$ summing to at most 1, the mixture $E_t^{(r,a)} := \sum_{\lambda \in \Lambda_{\mathrm{pw}}} w_\lambda E_t^{(r,a)}(\lambda)$ satisfies, for every $T \geq 1$,*

$$T\rho^\star - \log \frac{1}{w_{\lambda^\star}} \leq \mathbb{E}_Q[\log E_T^{(r,a)}] \leq T\rho^\star.$$

*The mixture therefore has the same asymptotic expected log-growth rate as the oracle fixed-$\lambda^\star$ bet, with the additive log-weight term $\log(1/w_{\lambda^\star})$.*

*Proof. Preliminary: e-process property.* For any $P \in H_0^{(a)}$, $X_i$ is independent of $\mathcal{F}_{i-1}$ with $\mathbb{E}_P[Z_i^{(r,a)}] = p_r - p_a \leq 0$, so

$$\mathbb{E}_P[1 + \lambda_i Z_i^{(r,a)} \mid \mathcal{F}_{i-1}] = 1 + \lambda_i(p_r - p_a) \leq 1$$

($\lambda_i \geq 0$). Thus $(E_t)$ is a nonneg. supermartingale with $E_0 = 1$, and optional stopping gives $\mathbb{E}_P[E_\tau] \leq 1$ for any stopping time $\tau$.

*Part (a): GRO bet.* Fix $E \in \mathcal{E}_Z$. For each $i$, $\lambda_i$ is $\mathcal{F}_{i-1}$-measurable and $X_i$ is independent of $\mathcal{F}_{i-1}$ under $Q$, taking values $r$, $a$, and "other" with probabilities $p$, $q$, and $1 - p - q$ (in which last case $Z_i = 0$ and the log term vanishes). Conditioning on $\mathcal{F}_{i-1}$,

$$\mathbb{E}_Q\big[\log(1 + \lambda_i Z_i^{(r,a)}) \mid \mathcal{F}_{i-1}\big] = p\log(1 + \lambda_i) + q\log(1 - \lambda_i) =: \rho(\lambda_i).$$

Summing over $i = 1, \ldots, T$ via the tower property gives

$$\mathbb{E}_Q[\log E_T] = \sum_{i=1}^{T} \mathbb{E}_Q[\rho(\lambda_i)]. \tag{28}$$

By Proposition C.6, $\rho$ is strictly concave on $(0, 1)$ with unique maximiser $\lambda^\star$, so $\rho(\lambda_i) \leq \rho^\star$ pointwise. Substituting into (28) yields $\mathbb{E}_Q[\log E_T] \leq T\rho^\star$, with equality iff $\lambda_i = \lambda^\star$ $Q$-a.s. for each $i$.

*Lower bound for part (b).*

For every $\lambda \in \Lambda_{\mathrm{pw}}$, $E_t^{(r,a)}(\lambda) > 0$, hence $E_T^{(r,a)} \geq w_{\lambda^\star} E_T^{(r,a)}(\lambda^\star)$ $Q$-a.s. Taking logs and $Q$-expectations and applying (a) to the constant bet $\lambda_i \equiv \lambda^\star$:

$$\mathbb{E}_Q[\log E_T^{(r,a)}] \geq \log w_{\lambda^\star} + T\rho^\star = T\rho^\star - \log(1/w_{\lambda^\star}).$$

*Upper bound for part (b), via variational duality.* Let $P^\star$ be the boundary distribution

$$P^\star(X = r) = P^\star(X = a) := \tfrac{p+q}{2}, \qquad P^\star(X = b) := Q(X = b) \text{ for } b \notin \{r, a\}.$$

The Lagrangian argument of Lemma C.9, applied to $H_0^{(a)} = \{P : p_r \leq p_a\}$ in place of $H_{0,r}$, identifies $P^\star$ as the forward I-projection of $Q$ onto $H_0^{(a)}$, with the closed form

$$D(Q\|P^\star) = p\log\frac{2p}{p+q} + q\log\frac{2q}{p+q} = p\log(1 + \lambda^\star) + q\log(1 - \lambda^\star) = \rho^\star, \tag{29}$$

the second equality from $1 \pm \lambda^\star = 2p/(p+q)$ and $2q/(p+q)$.

Writing $W := \sum_{\lambda \in \Lambda_{\mathrm{pw}}} w_\lambda \leq 1$, $E_T^{(r,a)}$ is an e-process for $H_0^{(a)}$ (mixture of e-processes), so $\mathbb{E}_{P^\star}[E_T^{(r,a)}] \leq W \leq 1$. Decomposing the $Q$-expectation via change of measure and applying Jensen to the concave $\log$,

$$\begin{aligned}\mathbb{E}_Q[\log E_T^{(r,a)}] &= \mathbb{E}_Q\left[\log\left(E_T^{(r,a)} \cdot \frac{d(P^\star)^T}{dQ^T}\right)\right] + \mathbb{E}_Q\left[\log\frac{dQ^T}{d(P^\star)^T}\right] \\ &\leq \log \mathbb{E}_{P^\star}[E_T^{(r,a)}] + D(Q^T\|(P^\star)^T) \leq 0 + TD(Q\|P^\star) = T\rho^\star,\end{aligned}$$

where the last line uses $\mathbb{E}_{P^\star}[E_T^{(r,a)}] \leq 1$, tensorisation of KL, and (29). Combining with the lower bound finishes the proof. □

### F.4 Top-$k$ mode certification

Fix a target set $S \subseteq \mathcal{A}$ of size $k \geq 1$ with $\mathcal{A} \setminus S \neq \emptyset$ (so that the outsider competitor set is meaningful). We extend CITE to certify that $S$ contains the top-$k$ categories in the sense

$$H_{1,S}^{(k)} := \big\{P : \min_{s \in S} p_s > \sup_{a \notin S} p_a\big\}, \qquad H_{0,S}^{(k)} := \neg H_{1,S}^{(k)}.$$

**Top-$k$ CITE.** Let

$$\delta_S := \min_{s\in S} p_s - \sup_{a\notin S} p_a, \qquad p_S := \min_{s\in S} p_s.$$

For each $s \in S$ and $a \notin S$, define $E_t^{(s,a)}$ as in Definition 3.1, and let $L_t(s)$ be the LCB for $p_s$.

At time $t$, if $\mathcal{A}_t \setminus S \neq \emptyset$, define

$$\widehat{a}_t^{\,\mathrm{out}} \in \arg\max_{a\in\mathcal{A}_t\setminus S} N_t(a)$$

with deterministic tie-breaking. Define

$$\mathsf{PW}_t^{(k)} := \begin{cases} \text{true}, & \mathcal{A}_t \setminus S = \emptyset, \\ \left\{ \min_{s\in S} \widetilde{E}_t^{(s,\widehat{a}_t^{\,\mathrm{out}})} \geq \alpha_{\mathrm{pw}}^{-1} \right\}, & \mathcal{A}_t \setminus S \neq \emptyset. \end{cases}$$

The top-$k$ CITE stopping time is

$$\tau^{(k)} := \inf\left\{ t \geq 1 : \mathsf{PW}_t^{(k)} \text{ and } \min_{s\in S} L_t(s) > U_t \right\}. \tag{30}$$

**Theorem F.9** (Top-$k$ Type-I and power)**.** *Run the top-$k$ CITE* (30) *under the same component-budget and selected-grid-weight bounded-below assumptions as Theorem 4.3, with the same per-component budget as the single-target CITE, $\alpha_{\mathrm{pw}} + \alpha_r + \alpha_u \leq \varepsilon$.*

(a) ***Type-I:*** *Under Assumption 2.1, $\sup_{P\in H_{0,S}^{(k)}} \mathbb{P}_P(\tau^{(k)} < \infty) \leq \varepsilon$.*

(b) ***Power:*** *Suppose $\delta_S > 0$ and the grids satisfy the analogue of Condition 4.2 with $\delta_S$ in place of $\delta$. Define the threshold*

$$T_{\mathrm{cl}}^{(k)} := C_0 \left( \frac{\log(k/\varepsilon) + \log(1/\delta_S)}{\delta_S^2} + \frac{\log(k/\varepsilon) + \log(1/p_S)}{p_S} \right)$$

*for a universal constant $C_0$. For* fixed *$k$, as $\varepsilon \to 0$,*

$$\mathbb{E}[\tau^{(k)}] \leq T_{\mathrm{cl}}^{(k)}\big(1 + o(1)\big) = O(T_{\mathrm{cl}}^{(k)}).$$

*The upper bound extends to growing $k$ provided $k = O(T_{\mathrm{cl}}^{(k)})$, e.g. when $\delta_S^{-2}\log(k/\varepsilon)$ dominates $k$.*

*Proof. (a) Type-I.* Fix any $P \in H_{0,S}^{(k)}$. Choose $s^\star \in S$ such that

$$p_{s^\star} = \min_{s\in S} p_s.$$

Since $P \in H_{0,S}^{(k)}$,

$$p_{s^\star} \leq \sup_{a\notin S} p_a.$$

If this supremum is positive, Lemma B.1 gives $a^\star \notin S$ with $p_{a^\star} \geq p_{s^\star}$. If $p_{s^\star} = 0$, any outsider $a^\star \notin S$ can be used.

Decompose

$$\{\tau^{(k)} < \infty\} \subseteq \underbrace{\{\tau^{(k)} < \infty, a^\star \in \mathcal{A}_{\tau^{(k)}}\}}_{E_1} \cup \underbrace{\{\tau^{(k)} < \infty, a^\star \notin \mathcal{A}_{\tau^{(k)}}\}}_{E_2}.$$

On $E_1$, the procedure's pairwise condition holds at $\tau^{(k)}$ for *every* $s \in S$, including the witness $s^\star$: $E_{\tau^{(k)}}^{(s^\star,\widehat{a}_{\tau^{(k)}}^{\,\mathrm{out}})} \geq 1/\alpha_{\mathrm{pw}}$. Runner-up monotonicity (Lemma B.3) restricted to outsiders (since $a^\star \in \mathcal{A}_{\tau^{(k)}} \setminus S$ implies $N_{\tau^{(k)}}(a^\star) \leq N_{\tau^{(k)}}(\widehat{a}_{\tau^{(k)}}^{\,\mathrm{out}})$) gives $E_{\tau^{(k)}}^{(s^\star,a^\star)} \geq E_{\tau^{(k)}}^{(s^\star,\widehat{a}_{\tau^{(k)}}^{\,\mathrm{out}})} \geq 1/\alpha_{\mathrm{pw}}$. Since $s^\star, a^\star$ are deterministic and $(E_t^{(s^\star,a^\star)})_{t\geq 0}$ is an NSM with initial value $\leq 1$ under $p_{s^\star} \leq p_{a^\star}$ (Corollary 3.2), Ville's inequality (11) yields $\mathbb{P}(E_1) \leq \alpha_{\mathrm{pw}}$.

On $E_2$, condition (b) at $\tau^{(k)}$ holds for the witness $s^\star$: $L_{\tau^{(k)}}(s^\star) > U_{\tau^{(k)}}$. Define the two failure events

$$F_r := \{\exists t : L_t(s^\star) > p_{s^\star}\}, \qquad F_u := \{\exists a, t : N_t(a) = 0,\ p_a > U_t\}.$$

Proposition 3.5 (applied to the single index $s^\star$) gives $\mathbb{P}(F_r) \le \alpha_r$, and Proposition 3.7 gives $\mathbb{P}(F_u) \le \alpha_u$. On $E_2 \setminus (F_r \cup F_u)$, $L_{\tau^{(k)}}(s^\star) \le p_{s^\star}$ and $p_{a^\star} \le U_{\tau^{(k)}}$ (since $a^\star \notin \mathcal{A}_{\tau^{(k)}}$), yielding the contradiction $p_{a^\star} \le U_{\tau^{(k)}} < L_{\tau^{(k)}}(s^\star) \le p_{s^\star}$, contrary to $p_{a^\star} \ge p_{s^\star}$. Hence $\mathbb{P}(E_2) \le \alpha_r + \alpha_u$.

A union bound gives $\mathbb{P}(\tau^{(k)} < \infty) \le \alpha_{\mathrm{pw}} + \alpha_r + \alpha_u \le \varepsilon$.

*(b) Power.* The single-target CITE of Definition 3.8 applied to a target $s \in S$ is *not* appropriate here, because its pairwise check ranges over all of $\mathcal{A} \setminus \{s\}$ and may include competitors $s' \in S$ with $p_{s'} \ge p_s$, on which the pairwise null is true; the resulting stopping time would be $+\infty$ a.s. Instead, fix $s \in S$ and consider the *outsider-restricted single-target CITE for target $s$*, defined exactly as Definition 3.8 except that the empirical runner-up $\widehat{a}_t$ is replaced by $\widehat{a}_t^{\mathrm{out}} := \arg\max_{a \in \mathcal{A}_t \setminus S} N_t(a)$ and the pairwise threshold is $k/\alpha_{\mathrm{pw}}$, the LCB threshold is $k/\alpha_r$, the unseen budget $\alpha_u$ is unchanged. We adopt the same convention as for the actual top-$k$ procedure: if $\mathcal{A}_t \setminus S = \emptyset$, the per-$s$ pairwise condition is treated as vacuously satisfied at round $t$. The tightened thresholds $k/\alpha_{\mathrm{pw}}$ and $k/\alpha_r$ are *strictly stronger sufficient* requirements than the actual top-$k$ stopping criteria (which use the original thresholds $1/\alpha_{\mathrm{pw}}$ and the LCB inequality $L_t(s) > U_t$); they are introduced solely so that a Bonferroni union bound over $s \in S$ on the per-target failure events yields a uniform $\varepsilon$-level conclusion. Write $\tau_s^{\mathrm{out}}$ for its stopping time.

We use a fixed-time tail bound rather than a max-of-first-hitting-times argument (the latter fails because pairwise e-processes and LCBs are not monotone in $t$). For each fixed $t \ge 1$,

$$\{\tau^{(k)} > t\} \subseteq \bigcup_{s \in S} \left( \left\{ \mathcal{A}_t \setminus S \neq \emptyset,\ E_t^{(s, \widehat{a}_t^{\mathrm{out}})} < \alpha_{\mathrm{pw}}^{-1} \right\} \cup \{L_t(s) \le U_t\} \right). \tag{31}$$

Apply the proof of Theorem 4.3 to the *outsider-restricted single-target CITE for target $s$*, with $\mathcal{A} \setminus \{r\}$ replaced by $\mathcal{A} \setminus S$ in the pairwise/LCB analysis (the unseen bound $U_t$ is unchanged since it is already over all of $\mathcal{A}$): the runner-up monotonicity (Lemma B.3) holds within any subset, the LCB construction is local to target $s$, and the effective modal gap $p_s - \max_{a \notin S} p_a \ge \delta_S$ and target mass $p_s \ge p_S$. With per-$s$ thresholds tightened to $k/\alpha_{\mathrm{pw}}$ for the pairwise check and $k/\alpha_r$ for the LCB (a Bonferroni split over $|S| = k$), the same McDiarmid + Bernstein arguments give, for each fixed $s$ and each $t \ge T_{\mathrm{cl}}^{(k)}$,

$$\mathbb{P}\big(E_t^{(s, \widehat{a}_t^{\mathrm{out}})} < 1/\alpha_{\mathrm{pw}} \text{or} L_t(s) \le U_t\big) \le 2(et)^{-2},$$

where $T_{\mathrm{cl}}^{(k)} = C_0\big((\log(k/\varepsilon) + \log(1/\delta_S))/\delta_S^2 + (\log(k/\varepsilon) + \log(1/p_S))/p_S\big)$ for a universal constant $C_0$. Substituting into (31) and applying a union bound over $s \in S$,

$$\mathbb{P}(\tau^{(k)} > t) \le 2k(et)^{-2} \qquad \text{for all } t \ge T_{\mathrm{cl}}^{(k)}.$$

The tail-sum identity ($\sum_{t \ge T} 1/t^2 \le 2/T$ for $T \ge 2$) yields

$$\mathbb{E}[\tau^{(k)}] \le T_{\mathrm{cl}}^{(k)} + \sum_{t > T_{\mathrm{cl}}^{(k)}} 2k(et)^{-2} \le T_{\mathrm{cl}}^{(k)}\big(1 + o(1)\big) \quad (\varepsilon \to 0,\ k \text{ fixed}),$$

which gives the claimed bound. □

**Remark F.10.** *The top-$k$ extension preserves the runner-up reduction within the "outsider" set, so the per-round cost is $k$ pairwise checks and $k$ LCB updates, plus one shared unseen bound. The case $k = 1$ recovers CITE of Definition 3.8, with Type-I and rate guarantees given by Theorems 4.1 and 4.3 respectively.*

# G Simulation Details

## G.1 Distribution Settings

Table 2 summarizes the five distribution settings used in Section 6. In each setting, the tail probabilities follow a shifted-Zipf distribution capped so that the stated modal gap $\delta = p_{(1)} - p_{(2)}$ is preserved exactly.

Table 2: Distribution settings for the simulation study.

| Setting | Name | $K$ | $p_r$ | $\delta$ | Tail |
|---|---|---|---|---|---|
| 1 | LLM-like Zipf (diffuse) | 5000 | 0.24 | 0.215 | $\text{Zipf}_{1.1}$ on 4987 labels |
| 2 | Concentrated (strong mode) | 100 | 0.60 | 0.450 | Uniform on 94 labels |
| 3 | Near-tie (small gap) | 500 | 0.12 | 0.010 | $\text{Zipf}_{1.3}$ on 497 labels |
| 4 | Very diffuse (huge category set) | 10000 | 0.06 | 0.010 | $\text{Zipf}_{1.0}$ on 9998 labels |
| 5 | Moderate LLM-like | 1000 | 0.35 | 0.150 | $\text{Zipf}_{1.2}$ on 997 labels |

### G.2 Results for All Reported Settings

**Type-I error control.** Across all five settings and all budgets $N \in \{64, 128, 256, 512, 1024, 2048\}$, CITE, W-CITE, and Bonferroni have empirical false-certification rate exactly zero in every Case B experiment. KR also has zero false certifications in Settings 1, 2, and 5, with only small nonzero rates on Settings 3–4. MMC shows nonzero false certifications in the same two regimes, illustrating that leader-tracking certificates are not designed for the fixed-target null considered here. Table 3 reports the full per-budget Case B false-certification rates.

Table 3: Case B false-certification rates by budget. Entries are empirical false-certification rates at level $\varepsilon = 0.05$ under the fixed-target null.

| Setting | $N$ | Bonferroni | CITE | W-CITE | KR | MMC |
|---|---|---|---|---|---|---|
| 1: LLM-like Zipf | 64 | .000 | .000 | .000 | .000 | .000 |
| | 128 | .000 | .000 | .000 | .000 | .000 |
| | 256 | .000 | .000 | .000 | .000 | .000 |
| | 512 | .000 | .000 | .000 | .000 | .000 |
| | 1024 | .000 | .000 | .000 | .000 | .000 |
| | 2048 | .000 | .000 | .000 | .000 | .000 |
| 2: Concentrated | 64 | .000 | .000 | .000 | .000 | .000 |
| | 128 | .000 | .000 | .000 | .000 | .000 |
| | 256 | .000 | .000 | .000 | .000 | .000 |
| | 512 | .000 | .000 | .000 | .000 | .000 |
| | 1024 | .000 | .000 | .000 | .000 | .000 |
| | 2048 | .000 | .000 | .000 | .000 | .000 |
| 3: Near-tie | 64 | .000 | .000 | .000 | .002 | .000 |
| | 128 | .000 | .000 | .000 | .000 | .000 |
| | 256 | .000 | .000 | .000 | .000 | .002 |
| | 512 | .000 | .000 | .000 | .002 | .004 |
| | 1024 | .000 | .000 | .000 | .002 | .006 |
| | 2048 | .000 | .000 | .000 | .002 | .006 |
| 4: Very diffuse | 64 | .000 | .000 | .000 | .000 | .000 |
| | 128 | .000 | .000 | .000 | .000 | .012 |
| | 256 | .000 | .000 | .000 | .000 | .012 |
| | 512 | .000 | .000 | .000 | .000 | .012 |
| | 1024 | .000 | .000 | .000 | .002 | .014 |
| | 2048 | .000 | .000 | .000 | .000 | .014 |
| 5: Moderate LLM-like | 64 | .000 | .000 | .000 | .000 | .000 |
| | 128 | .000 | .000 | .000 | .000 | .000 |
| | 256 | .000 | .000 | .000 | .000 | .000 |
| | 512 | .000 | .000 | .000 | .000 | .000 |
| | 1024 | .000 | .000 | .000 | .000 | .000 |
| | 2048 | .000 | .000 | .000 | .000 | .000 |

**Power (Case A).** Tables 4–8 report certification rates for all five settings. Setting 2 saturates quickly for all methods, whereas Settings 3 and 4 are the most difficult regimes. Setting 5 highlights a diffuse-tail regime in which CITE substantially outperforms MMC at moderate budgets. Bonferroni can be slightly more powerful in the easiest concentrated regime at the smallest budget, but CITE matches or improves on it in the remaining reported regimes and budgets.

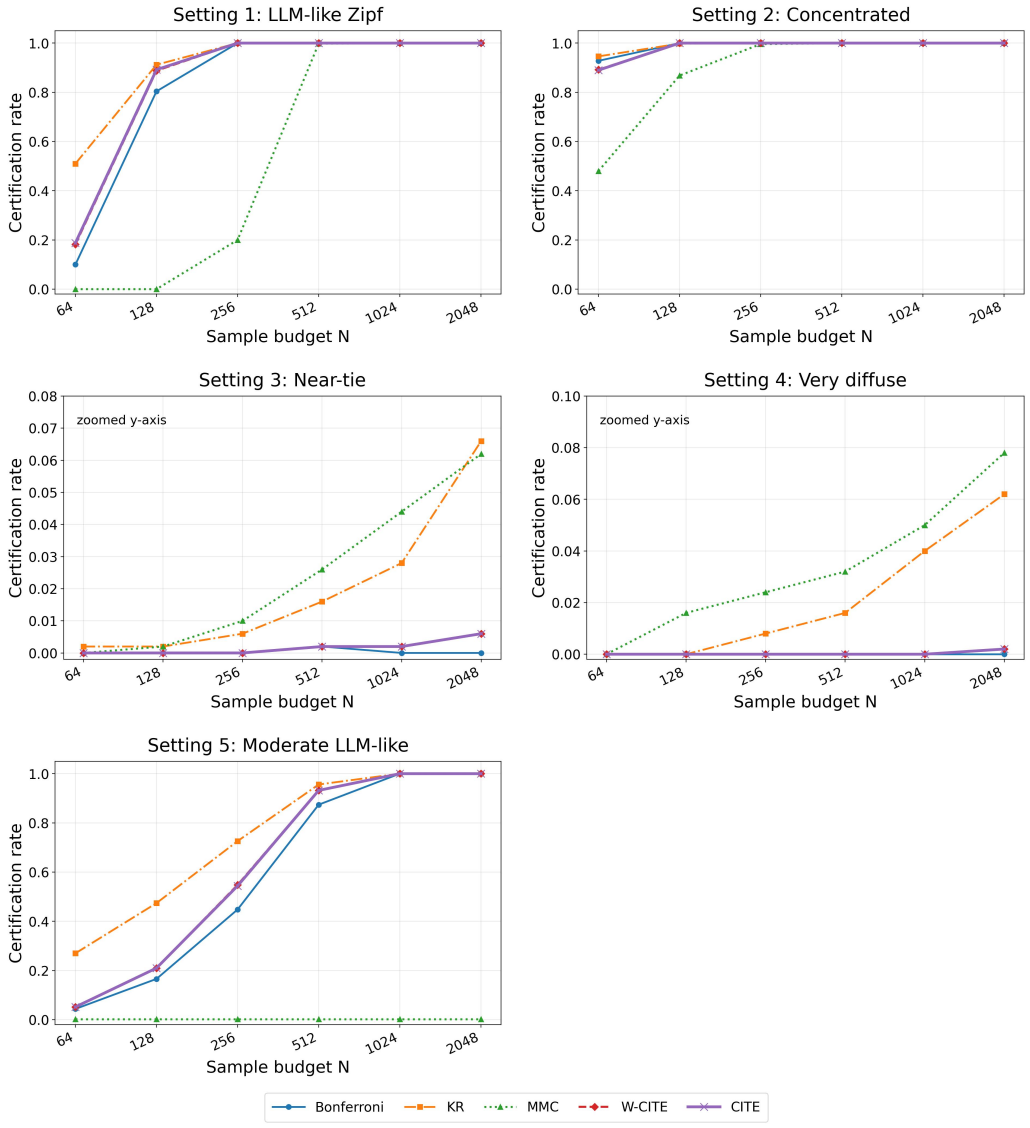



Figure 3: Certification rate versus sample budget $N$ for CITE, W-CITE, fixed-sample Bonferroni, the sample-split fixed-target test KR [Kim and Ramdas, 2025], and the leader-tracking MMC [Cordero-Encinar and Duncan, 2025], across all five reported simulation settings. The panels for Settings 3 and 4 use zoomed vertical axes because all methods remain low-power over the evaluated budget range. CITE matches or exceeds Bonferroni in most regimes, with the easy concentrated Setting 2 saturating quickly for both methods. Settings 3–4 are also the regimes where KR and MMC show nonzero false certifications under Case B (Table 3).

Table 4: Setting 1 (LLM-like Zipf): certification rate, Case A.

| $N$ | Bonferroni | CITE | W-CITE | KR | MMC | $\bar{\tau}_{\text{CITE}}$ |
|---|---|---|---|---|---|---|
| 64 | .100 | .188 | .182 | .510 | .000 | 53 |
| 128 | .804 | .892 | .888 | .912 | .000 | 84 |
| 256 | 1.00 | 1.00 | 1.00 | 1.00 | .200 | 90 |
| 512 | 1.00 | 1.00 | 1.00 | 1.00 | .998 | 90 |
| 1024 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 90 |
| 2048 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 90 |

Table 5: Setting 2 (Concentrated, strong mode): certification rate, Case A.

| $N$ | Bonferroni | CITE | W-CITE | KR | MMC | $\bar{\tau}_{\text{CITE}}$ |
|---|---|---|---|---|---|---|
| 64 | .928 | .890 | .892 | .946 | .480 | 36 |
| 128 | 1.00 | 1.00 | 1.00 | .998 | .868 | 41 |
| 256 | 1.00 | 1.00 | 1.00 | 1.00 | .996 | 41 |
| 512 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 41 |
| 1024 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 41 |
| 2048 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 41 |

Table 6: Setting 3 (Near-tie, $\delta = 0.01$): certification rate, Case A.

| $N$ | Bonferroni | CITE | W-CITE | KR | MMC |
|---|---|---|---|---|---|
| 64 | .000 | .000 | .000 | .002 | .000 |
| 128 | .000 | .000 | .000 | .002 | .002 |
| 256 | .000 | .000 | .000 | .006 | .010 |
| 512 | .002 | .002 | .002 | .016 | .026 |
| 1024 | .000 | .002 | .002 | .028 | .044 |
| 2048 | .000 | .006 | .006 | .066 | .062 |

Table 7: Setting 4 (Very diffuse, huge category set): certification rate, Case A.

| $N$ | Bonferroni | CITE | W-CITE | KR | MMC | $\bar{\tau}_{\text{CITE}}$ |
|---|---|---|---|---|---|---|
| 64 | .000 | .000 | .000 | .000 | .000 | — |
| 128 | .000 | .000 | .000 | .000 | .016 | — |
| 256 | .000 | .000 | .000 | .008 | .024 | — |
| 512 | .000 | .000 | .000 | .016 | .032 | — |
| 1024 | .000 | .000 | .000 | .040 | .050 | — |
| 2048 | .000 | .002 | .002 | .062 | .078 | 1149 |

Table 8: Setting 5 (Moderate LLM-like): certification rate, Case A.

| $N$ | Bonferroni | CITE | W-CITE | KR | MMC | $\bar{\tau}_{\text{CITE}}$ |
|---|---|---|---|---|---|---|
| 64 | .044 | .052 | .052 | .270 | .002 | 49 |
| 128 | .166 | .210 | .210 | .474 | .002 | 84 |
| 256 | .448 | .544 | .548 | .726 | .002 | 151 |
| 512 | .874 | .932 | .932 | .956 | .002 | 234 |
| 1024 | 1.00 | 1.00 | 1.00 | 1.00 | .002 | 259 |
| 2048 | 1.00 | 1.00 | 1.00 | 1.00 | .002 | 259 |

### G.3 Bottleneck Analysis

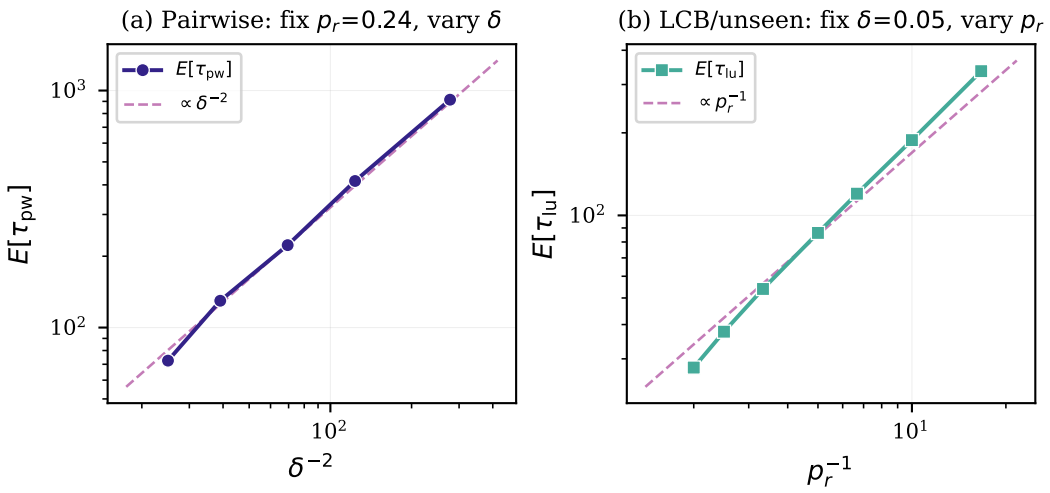



Figure 4: Component-level diagnostics for the two terms in the stopping-time upper bound of Theorem 4.3, on log-log axes. Dashed reference lines show the theoretical slopes. (a) Fixing $p_r = 0.24$ and varying $\delta$, the pairwise component $\mathbb{E}[\tau_{\text{pw}}]$ (first time the pairwise e-value crosses $1/\alpha_{\text{pw}}$) is consistent with the $\delta^{-2}$ term. (b) Fixing $\delta = 0.05$ and varying $p_r$, the LCB/unseen component $\mathbb{E}[\tau_{\text{lu}}]$ (first time $L_t(r) > U_t$) is consistent with the $p_r^{-1}$ term.

### G.4 Logistic Weight Model Saturation

As discussed in Section 6, the logistic weight model can saturate when $K$ is large. For Settings 1 and 5, sweeping $\beta \in \{0.5, 1, 2, 4, 8, 16\}$ produces certification rates and stopping times that are nearly unchanged relative to unweighted CITE. The effective gap ratio $\Delta_w/\mu_r$ is essentially the same as $\delta/p_r$ for $\beta \geq 2$. This occurs because the top categories are already assigned weights close to one, so the logistic transformation does not substantially separate the mode from its runner-up.

### G.5 Rank-Based Weight Model

The rank-based model assigns $E[W \mid X = a] = 0.95 \cdot e^{-\gamma \cdot \text{rank}(a)}$ for the top $K_0 = 10$ categories and $w_{\text{low}} = 0.1$ otherwise, with additive $\text{Uniform}(-0.05, 0.05)$ noise clipped to $[0.01, 1]$.

Table 9 reports detailed results for Setting 5 at $N = 512$. We omit a separate figure here to avoid duplicating Figure 3; the table directly shows how the effective weighted gap and stopping time change with the rank-decay parameter.

Table 9: Setting 5: rank-based W-CITE at $N = 512$ in the supplementary weight sweep.

| $\gamma$ | W-CITE rate | $\bar{\tau}_{\text{W-CITE}}$ | $\bar{\tau}_{\text{CITE}}$ | $\Delta_w/\mu_r$ | $\delta/p_r$ |
|---|---|---|---|---|---|
| 0.0 | .920 | 237 | 235 | 0.429 | 0.429 |
| 0.1 | .990 | 185 | 235 | 0.483 | 0.429 |
| 0.2 | 1.00 | 142 | 235 | 0.532 | 0.429 |
| 0.5 | 1.00 | 71 | 235 | 0.653 | 0.429 |
| 1.0 | 1.00 | 48 | 235 | 0.790 | 0.429 |
| 2.0 | 1.00 | 45 | 235 | 0.923 | 0.429 |
| 3.0 | 1.00 | 45 | 235 | 0.972 | 0.429 |

In this synthetic rank-based model, performance improves as the effective gap ratio $\Delta_w/\mu_r$ increases. A modest $\gamma = 0.2$ already reaches full certification in this sweep and reduces $\bar{\tau}$ by about 40%; at $\gamma = 1.0$, the reduction is about 80%. For Setting 1, the rank-based model with $\gamma \geq 0.1$ reaches 99.6% certification at $N = 128$ (versus 88.6% unweighted) with $\bar{\tau} \approx 72$ (versus 92).

Table 10: Certification rate on LLM self-consistency outputs (Case A, target = mode), averaged across problems. $\bar{K}_N$: Monte Carlo mean number of observed answer categories at budget $N$. $\bar{\tau}$ columns report mean stopping time among certified replicates.

| Model | Dataset | $N$ | Bonferroni | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\text{CITE}}$ | $\bar{\tau}_{\text{W-CITE}}$ | $\bar{\tau}_{\text{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-30B | AIME 2026 | 64 | 0.499±0.002 | 0.724±0.001 | 0.727±0.001 | 0.778±0.002 | 0.671±0.001 | 7.1±0.0 | 21.8±0.3 | 22.8±0.2 | 15.9±0.3 |
| | | 128 | 0.597±0.002 | 0.781±0.001 | 0.785±0.001 | 0.827±0.002 | 0.705±0.001 | 10.2±0.0 | 32.4±0.5 | 33.0±0.5 | 22.5±0.5 |
| | | 256 | 0.683±0.002 | 0.833±0.002 | 0.838±0.002 | 0.880±0.002 | 0.744±0.001 | 14.5±0.0 | 49.1±0.6 | 55.5±0.8 | 32.1±0.6 |
| | | 512 | 0.770±0.002 | 0.902±0.001 | 0.906±0.001 | 0.929±0.001 | 0.779±0.001 | 19.9±0.0 | 81.8±1.3 | 82.0±1.1 | 45.6±1.1 |
| | | 1024 | 0.825±0.001 | 0.944±0.001 | 0.946±0.001 | 0.959±0.001 | 0.798±0.000 | 25.7±0.0 | 113.7±1.8 | 113.2±1.4 | 57.8±2.2 |
| | FrontierScience-Olympiad | 64 | 0.284±0.002 | 0.288±0.002 | 0.289±0.002 | 0.328±0.003 | 0.121±0.002 | 28.0±0.0 | 37.9±1.9 | 40.9±1.6 | 22.9±0.9 |
| | | 128 | 0.312±0.002 | 0.315±0.002 | 0.315±0.002 | 0.373±0.003 | 0.146±0.003 | 46.0±0.0 | 65.5±2.5 | 64.3±2.9 | 35.9±1.1 |
| | | 256 | 0.350±0.003 | 0.356±0.003 | 0.359±0.003 | 0.443±0.004 | 0.198±0.003 | 73.7±0.1 | 110.0±2.7 | 112.0±2.5 | 60.6±1.5 |
| | | 512 | 0.429±0.003 | 0.449±0.004 | 0.453±0.004 | 0.536±0.004 | 0.245±0.002 | 112.9±0.1 | 195.5±3.3 | 193.5±3.2 | 97.0±2.2 |
| | | 1024 | 0.530±0.003 | 0.563±0.004 | 0.567±0.004 | 0.633±0.004 | 0.281±0.002 | 160.0±0.1 | 436.1±15.0 | 439.2±12.3 | 138.5±2.9 |
| | ER-REASON | 64 | 0.102±0.000 | 0.983±0.000 | 0.983±0.000 | 0.990±0.000 | 0.988±0.000 | 1.1±0.0 | 10.6±0.0 | 11.5±0.0 | 8.2±0.0 |
| | | 128 | 0.110±0.000 | 0.996±0.000 | 0.996±0.000 | 0.998±0.000 | 0.996±0.000 | 1.1±0.0 | 11.3±0.0 | 12.2±0.0 | 8.8±0.0 |
| | | 256 | 0.111±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 11.8±0.0 | 12.6±0.0 | 9.3±0.0 |
| | | 512 | 0.111±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 11.8±0.0 | 12.7±0.0 | 9.3±0.1 |
| | | 1024 | 0.111±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 11.8±0.0 | 12.7±0.0 | 9.3±0.1 |
| gpt-oss-20b | AIME 2026 | 64 | 0.563±0.002 | 0.596±0.002 | 0.599±0.002 | 0.666±0.002 | 0.459±0.001 | 12.6±0.0 | 29.3±0.2 | 29.6±0.2 | 21.0±0.6 |
| | | 128 | 0.659±0.002 | 0.693±0.001 | 0.701±0.002 | 0.742±0.002 | 0.486±0.001 | 18.3±0.0 | 43.8±1.6 | 43.5±1.6 | 27.9±0.8 |
| | | 256 | 0.740±0.002 | 0.765±0.001 | 0.778±0.001 | 0.795±0.001 | 0.513±0.001 | 25.7±0.0 | 62.8±2.9 | 60.7±3.2 | 40.3±1.5 |
| | | 512 | 0.799±0.001 | 0.811±0.001 | 0.812±0.001 | 0.830±0.001 | 0.538±0.001 | 34.8±0.0 | 84.2±2.9 | 80.7±3.1 | 58.0±2.1 |
| | | 1024 | 0.825±0.001 | 0.829±0.001 | 0.829±0.001 | 0.848±0.001 | 0.554±0.001 | 44.0±0.0 | 176.7±6.0 | 126.6±3.7 | 91.0±4.5 |
| | FrontierScience-Olympiad | 64 | 0.193±0.001 | 0.187±0.002 | 0.147±0.002 | 0.209±0.002 | 0.192±0.001 | 37.8±0.0 | 32.3±2.2 | 28.7±0.5 | 22.1±0.4 |
| | | 128 | 0.205±0.001 | 0.213±0.002 | 0.188±0.001 | 0.255±0.003 | 0.199±0.000 | 65.1±0.1 | 67.6±2.4 | 40.4±0.7 | 24.0±0.5 |
| | | 256 | 0.237±0.002 | 0.269±0.003 | 0.201±0.001 | 0.330±0.004 | 0.200±0.000 | 108.5±0.1 | 144.6±2.9 | 108.9±1.8 | 24.3±0.5 |
| | | 512 | 0.340±0.003 | 0.382±0.003 | 0.242±0.002 | 0.446±0.004 | 0.200±0.000 | 172.0±0.1 | 258.7±4.5 | 319.0±2.3 | 24.3±0.5 |
| | | 1024 | 0.429±0.002 | 0.475±0.003 | 0.428±0.004 | 0.559±0.004 | 0.200±0.000 | 250.3±0.1 | 494.7±13.2 | 499.9±3.7 | 24.3±0.5 |
| | ER-REASON | 64 | 0.623±0.001 | 0.640±0.001 | 0.644±0.001 | 0.722±0.001 | 0.654±0.001 | 3.2±0.0 | 26.6±0.2 | 28.4±0.2 | 22.1±0.2 |
| | | 128 | 0.726±0.001 | 0.730±0.001 | 0.739±0.001 | 0.795±0.001 | 0.727±0.001 | 3.6±0.0 | 41.4±0.5 | 42.1±0.5 | 33.1±0.6 |
| | | 256 | 0.813±0.001 | 0.816±0.001 | 0.821±0.001 | 0.857±0.001 | 0.780±0.001 | 3.8±0.0 | 63.0±0.7 | 60.7±0.6 | 47.5±0.7 |
| | | 512 | 0.873±0.001 | 0.870±0.001 | 0.875±0.001 | 0.903±0.001 | 0.823±0.001 | 4.0±0.0 | 89.7±1.0 | 91.5±1.6 | 71.4±1.4 |
| | | 1024 | 0.921±0.001 | 0.915±0.001 | 0.919±0.001 | 0.939±0.001 | 0.861±0.001 | 4.2±0.0 | 128.3±2.3 | 127.5±2.2 | 110.8±2.1 |

Table 11: Certification rate on LLM self-consistency outputs (Case B, target = runner-up), averaged across problems. $\bar{K}_N$: Monte Carlo mean number of observed answer categories at budget $N$. $\bar{\tau}$ columns report mean stopping time among certified replicates.

| Model | Dataset | $N$ | Bonferroni | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\text{CITE}}$ | $\bar{\tau}_{\text{W-CITE}}$ | $\bar{\tau}_{\text{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen3-30B | AIME 2026 | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 7.1±0.0 | — | — | 17.2±0.7 |
| | | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 10.2±0.0 | — | — | 17.2±0.7 |
| | | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 14.5±0.0 | — | — | 17.2±0.7 |
| | | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 19.9±0.0 | — | — | 17.2±0.7 |
| | | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 25.7±0.0 | — | — | 17.2±0.7 |
| | FrontierScience-Olympiad | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 28.0±0.0 | — | — | 26.0±0.0 |
| | | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 46.0±0.0 | — | — | 26.0±0.0 |
| | | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 73.7±0.1 | — | — | 26.0±0.0 |
| | | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 112.9±0.1 | — | — | 26.0±0.0 |
| | | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 160.0±0.1 | — | — | 26.0±0.0 |
| | ER-REASON | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 1.1±0.0 | — | — | 7.0±0.0 |
| | | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 1.1±0.0 | — | — | 7.0±0.0 |
| | | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 1.1±0.0 | — | — | 7.0±0.0 |
| | | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 1.1±0.0 | — | — | 7.0±0.0 |
| | | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 1.1±0.0 | — | — | 7.0±0.0 |
| gpt-oss-20b | AIME 2026 | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.000 | 0.000±0.000 | 12.6±0.0 | 9.0±0.0 | 9.0±0.0 | 26.0±7.8 |
| | | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.000 | 0.000±0.000 | 18.3±0.0 | 40.0±31.0 | 40.0±31.0 | 26.0±7.8 |
| | | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.000 | 0.000±0.000 | 25.7±0.0 | 40.0±31.0 | 40.0±31.0 | 26.0±7.8 |
| | | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 34.8±0.0 | 40.0±31.0 | 40.0±31.0 | 26.0±7.8 |
| | | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 44.0±0.0 | 354.5±15.5 | 485.5±13.4 | 26.0±7.8 |
| | FrontierScience-Olympiad | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 37.8±0.0 | — | — | — |
| | | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 65.1±0.1 | — | — | — |
| | | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 108.5±0.1 | — | — | — |
| | | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.000±0.000 | 172.0±0.1 | — | — | — |
| | | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 250.3±0.1 | — | — | — |
| | ER-REASON | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.000 | 0.001±0.000 | 3.2±0.0 | 24.6±5.3 | 30.0±9.3 | 16.0±1.6 |
| | | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.000 | 0.001±0.000 | 3.6±0.0 | 28.1±6.3 | 36.8±11.5 | 16.0±1.6 |
| | | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.001±0.000 | 3.8±0.0 | 28.1±6.3 | 36.8±11.5 | 21.1±3.2 |
| | | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.001±0.000 | 4.0±0.0 | 28.1±6.3 | 36.8±11.5 | 21.1±3.2 |
| | | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.001±0.000 | 0.001±0.000 | 4.2±0.0 | 123.4±44.4 | 36.8±11.5 | 27.1±6.6 |

# H LLM Self-Consistency Details

Table 10 reports the certification rate averaged across problems for each (model, dataset, budget) under Case A (target = mode); Table 11 reports the corresponding Case B (target = runner-up) certification rates as a sanity check on Type-I error. For each (model, problem) pair we draw 1,000 i.i.d. samples, then bootstrap 500 replicates of each budget $N \in \{64, 128, 256, 512, 1024\}$ from the empirical answer pool to compute certification rates.

## H.1 Model settings

**Sampling parameters.** Qwen3-30B-A3B-Instruct-2507: `temperature=0.7`, `top_p=0.8`, `top_k=20`. gpt-oss-20b: `temperature=0.4`, `top_p=0.9`, `top_k=20`, `reasoning_effort=low`.

**Response probability weight for W-CITE.** For each sample $i$, let $(y_1, \ldots, y_{T_i})$ denote the tokens of the extracted answer span (the final `\boxed{...}` contents for AIME and FrontierScience, the `Triage:` ⟨`acuity`⟩ | `Disposition:` ⟨`disposition`⟩ string for ER-REASON). The W-CITE weight is the response probability used by the weighted variant of self-consistency [Wang et al., 2023],

$$w_i = \prod_{t=1}^{T_i} p_\theta \left( y_t \mid y_{<t}, \text{prompt} \right),$$

i.e., the model's own likelihood of the extracted answer under token-by-token factorization. Weights are not renormalized across samples.

### H.2 Dataset protocols

**AIME 2026.** All 30 problems of AIME 2026 [Dekoninck et al., 2026] (AIME I and AIME II combined), each with an integer gold answer in $[0, 999]$. Prompt is the problem statement with a per-model suffix: "Solve in under 100 tokens. Put your final answer within `\boxed{}`." (Qwen3-30B-A3B-Instruct-2507) or "Do not show your reasoning. Put only the final answer within `\boxed{}`." (gpt-oss-20b). Answers are extracted from the last `\boxed{...}` and matched by exact integer equality.

**FrontierScience-Olympiad.** FrontierScience-Olympiad [Wang et al., 2026] contains 100 olympiad-level physics, chemistry, and biology problems with numeric or symbolic answers. We use a 10-problem subset obtained by taking every 10th problem (problems 10, 20, …, 100). Prompt suffix is the same as for AIME 2026. Each extracted answer is compared against the gold by an LLM self-judge (the generating model decides whether its own output is equivalent to the gold).

**ER-REASON.** ER-REASON [Mehandru et al., 2025] contains 3,984 emergency-room encounters. We use the 72-encounter subset with expert-authored clinical rationales, which the paper designates as the primary evaluation set. Both models use a two-turn chat prompt in place of the `\boxed{}` suffix. Turn 1 asks the model to assign an ESI acuity level (1–5) from the chief complaint, demographics, and triage vital signs. Turn 2, which echoes the Turn 1 acuity, asks for the disposition (one of 12 labels) given the full ED provider note and all historical notes. The final answer is parsed as the fixed string "Triage: ⟨acuity⟩ | Disposition: ⟨disposition⟩".

### H.3 Per-problem certification rates: Qwen3-30B-A3B-Instruct-2507

**AIME 2026.**

Table 12: Per-problem certification rate on AIME 2026 (30 problems) (Case A, target = mode). **Bold**: CITE > Bonferroni. $\bar{K}_N$ is the Monte Carlo mean number of observed answer categories at budget $N$.

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\text{W-CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P6 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| gold=441, mode=441 ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P20 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| gold=190, mode=190 ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P30 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.0±0.0 | 7.0±0.0 |
| gold=393, mode=243 ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.0±0.0 | 7.0±0.0 |
| P1 ($\delta$=0.998) | 64 | 0.066±0.011 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 9.1±0.0 | 10.0±0.0 | 7.0±0.0 |
| gold=277, mode=277 ✓ | 128 | 0.130±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 9.1±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.224±0.019 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.2±0.0 | 9.1±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.406±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.4±0.0 | 9.1±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.638±0.021 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.6±0.0 | 9.1±0.0 | 10.0±0.0 | 7.0±0.0 |
| P7 ($\delta$=0.995) | 64 | 0.186±0.017 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.2±0.0 | 9.1±0.0 | 10.1±0.0 | 7.1±0.0 |
| gold=396, mode=396 ✓ | 128 | 0.326±0.021 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.4±0.0 | 9.1±0.0 | 10.1±0.0 | 7.1±0.0 |
| | 256 | 0.530±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.6±0.0 | 9.1±0.0 | 10.1±0.0 | 7.1±0.0 |
| | 512 | 0.794±0.018 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.1±0.0 | 10.1±0.0 | 7.1±0.0 |
| | 1024 | 0.960±0.009 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.5±0.0 | 9.1±0.0 | 10.1±0.0 | 7.1±0.0 |
| P24 ($\delta$=0.994) | 64 | 0.234±0.019 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.3±0.0 | 9.2±0.0 | 10.2±0.0 | 7.1±0.0 |
| gold=669, mode=669 ✓ | 128 | 0.406±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.5±0.0 | 9.2±0.0 | 10.2±0.0 | 7.1±0.0 |
| | 256 | 0.652±0.021 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 9.2±0.0 | 10.2±0.0 | 7.1±0.0 |
| | 512 | 0.886±0.014 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.5±0.0 | 9.2±0.0 | 10.2±0.0 | 7.1±0.0 |
| | 1024 | 0.982±0.006 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.2±0.0 | 9.2±0.0 | 10.2±0.0 | 7.1±0.0 |

Table 12 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P21 ($\delta$=0.987) | 64 | 0.430±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.5±0.0 | 9.4±0.1 | 11.0±0.1 | 7.3±0.1 |
| gold=50, mode=50 ✓ | 128 | 0.706±0.020 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.4±0.1 | 11.0±0.1 | 7.3±0.1 |
| | 256 | 0.912±0.013 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 9.4±0.1 | 11.0±0.1 | 7.3±0.1 |
| | 512 | 0.986±0.005 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 9.4±0.1 | 11.0±0.1 | 7.3±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 9.4±0.1 | 11.0±0.1 | 7.3±0.1 |
| P5 ($\delta$=0.976) | 64 | 0.508±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.5±0.0 | 9.6±0.1 | 10.5±0.1 | 7.5±0.1 |
| gold=65, mode=65 ✓ | 128 | 0.792±0.018 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.8±0.0 | 9.6±0.1 | 10.5±0.1 | 7.5±0.1 |
| | 256 | 0.950±0.010 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 9.6±0.1 | 10.5±0.1 | 7.5±0.1 |
| | 512 | 0.994±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.6±0.1 | 10.5±0.1 | 7.5±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.6±0.1 | 10.5±0.1 | 7.5±0.1 |
| P25 ($\delta$=0.975) | 64 | 0.714±0.020 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 9.9±0.1 | 10.8±0.1 | 7.8±0.1 |
| gold=850, mode=850 ✓ | 128 | 0.924±0.012 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.3±0.1 | 9.9±0.1 | 10.8±0.1 | 7.8±0.1 |
| | 256 | 0.992±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.9±0.1 | 9.9±0.1 | 10.8±0.1 | 7.8±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.0±0.1 | 9.9±0.1 | 10.8±0.1 | 7.8±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 8.9±0.1 | 9.9±0.1 | 10.8±0.1 | 7.8±0.1 |
| P15 ($\delta$=0.970) | 64 | 0.752±0.019 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 10.1±0.1 | 12.8±0.1 | 7.9±0.1 |
| gold=83, mode=1 ✗ | 128 | 0.920±0.012 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 10.1±0.1 | 12.8±0.1 | 7.9±0.1 |
| | 256 | 0.996±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.1±0.0 | 10.1±0.1 | 12.8±0.1 | 7.9±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.2±0.0 | 10.1±0.1 | 12.8±0.1 | 7.9±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 10.1±0.1 | 12.8±0.1 | 7.9±0.1 |
| P12 ($\delta$=0.956) | 64 | 0.878±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.9±0.1 | 10.5±0.1 | 11.3±0.1 | 8.3±0.1 |
| gold=161, mode=161 ✓ | 128 | 0.988±0.005 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.5±0.1 | 10.5±0.1 | 11.3±0.1 | 8.3±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.9±0.1 | 10.5±0.1 | 11.3±0.1 | 8.3±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 10.0±0.1 | 10.5±0.1 | 11.3±0.1 | 8.3±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 12.8±0.1 | 10.5±0.1 | 11.3±0.1 | 8.3±0.1 |
| P8 ($\delta$=0.910) | 64 | 0.968±0.008 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 11.1±0.1 | 12.0±0.1 | 8.6±0.1 |
| gold=244, mode=244 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.6±0.0 | 11.1±0.1 | 12.0±0.1 | 8.6±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 11.1±0.1 | 12.0±0.1 | 8.6±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.5±0.0 | 11.1±0.1 | 12.0±0.1 | 8.6±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.8±0.0 | 11.1±0.1 | 12.0±0.1 | 8.6±0.1 |
| P22 ($\delta$=0.891) | 64 | 0.992±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.6±0.1 | 12.2±0.2 | 12.9±0.1 | 10.4±0.2 |
| gold=754, mode=754 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.9±0.1 | 12.2±0.2 | 12.9±0.1 | 10.4±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 10.3±0.1 | 12.2±0.2 | 12.9±0.1 | 10.4±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 14.8±0.1 | 12.2±0.2 | 12.9±0.1 | 10.4±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 20.3±0.1 | 12.2±0.2 | 12.9±0.1 | 10.4±0.2 |
| P3 ($\delta$=0.859) | 64 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 12.8±0.2 | 13.7±0.2 | 10.1±0.2 |
| gold=79, mode=79 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.1±0.0 | 12.8±0.2 | 13.7±0.2 | 10.1±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 12.8±0.2 | 13.7±0.2 | 10.1±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 12.8±0.2 | 13.7±0.2 | 10.1±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 12.8±0.2 | 13.7±0.2 | 10.1±0.2 |
| P2 ($\delta$=0.850) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.7±0.1 | 13.7±0.2 | 14.3±0.2 | 13.6±0.3 |
| gold=62, mode=62 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 9.9±0.1 | 13.7±0.2 | 14.3±0.2 | 13.6±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 13.6±0.1 | 13.7±0.2 | 14.3±0.2 | 13.6±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 17.4±0.1 | 13.7±0.2 | 14.3±0.2 | 13.6±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 21.0±0.1 | 13.7±0.2 | 14.3±0.2 | 13.6±0.3 |
| P26 ($\delta$=0.776) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.994±0.003 | 12.0±0.1 | 16.5±0.2 | 17.7±0.2 | 23.5±0.6 |
| gold=132, mode=132 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 19.6±0.1 | 16.5±0.2 | 17.7±0.2 | 23.8±0.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 30.2±0.2 | 16.5±0.2 | 17.7±0.2 | 23.8±0.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 43.1±0.2 | 16.5±0.2 | 17.7±0.2 | 23.8±0.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 57.4±0.1 | 16.5±0.2 | 17.7±0.2 | 23.8±0.6 |
| P27 ($\delta$=0.728) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.994±0.003 | 9.8±0.1 | 17.4±0.2 | 17.9±0.2 | 21.4±0.6 |
| gold=223, mode=223 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 15.7±0.1 | 17.4±0.2 | 17.9±0.2 | 21.8±0.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 24.1±0.1 | 17.4±0.2 | 17.9±0.2 | 21.8±0.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 34.7±0.1 | 17.4±0.2 | 17.9±0.2 | 21.8±0.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 46.4±0.1 | 17.4±0.2 | 17.9±0.2 | 21.8±0.6 |
| P28 ($\delta$=0.695) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.944±0.010 | 10.5±0.1 | 19.1±0.2 | 20.9±0.2 | 28.1±0.7 |
| gold=107, mode=12 ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 15.2±0.1 | 19.1±0.2 | 20.9±0.2 | 30.9±0.8 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 20.5±0.1 | 19.1±0.2 | 20.9±0.2 | 30.9±0.8 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 26.7±0.1 | 19.1±0.2 | 20.9±0.2 | 30.9±0.8 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 32.6±0.1 | 19.1±0.2 | 20.9±0.2 | 30.9±0.8 |
| P11 ($\delta$=0.690) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.990±0.004 | 8.6±0.1 | 19.1±0.3 | 20.5±0.3 | 23.3±0.6 |
| gold=896, mode=896 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 12.8±0.1 | 19.1±0.3 | 20.5±0.3 | 23.8±0.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 19.1±0.1 | 19.1±0.3 | 20.5±0.3 | 23.8±0.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 27.2±0.1 | 19.1±0.3 | 20.5±0.3 | 23.8±0.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 35.7±0.1 | 19.1±0.3 | 20.5±0.3 | 23.8±0.6 |
| P14 ($\delta$=0.429) | 64 | 0.976±0.007 | **0.990±0.004** | 0.990±0.004 | 0.996±0.003 | 0.014±0.005 | 24.5±0.1 | 33.8±0.5 | 35.2±0.4 | 17.9±5.4 |
| gold=681, mode=681 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.018±0.006 | 39.4±0.2 | 34.2±0.5 | 35.7±0.5 | 33.1±11.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.018±0.006 | 60.8±0.2 | 34.2±0.5 | 35.7±0.5 | 33.1±11.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.020±0.006 | 87.7±0.2 | 34.2±0.5 | 35.7±0.5 | 56.5±25.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.022±0.007 | 116.8±0.2 | 34.2±0.5 | 35.7±0.5 | 105.1±53.8 |
| P16 ($\delta$=0.392) | 64 | 0.808±0.018 | 0.550±0.022 | 0.564±0.022 | 0.792±0.018 | 0.682±0.021 | 2.0±0.0 | 35.3±1.0 | 36.2±0.9 | 29.0±1.0 |
| gold=178, mode=32 ✗ | 128 | 0.988±0.005 | 0.922±0.012 | 0.928±0.012 | 0.948±0.010 | 0.946±0.010 | 2.0±0.0 | 57.2±1.5 | 57.2±1.4 | 46.0±1.5 |
| | 256 | 1.000±0.000 | 0.998±0.002 | 1.000±0.000 | 0.996±0.003 | 0.998±0.002 | 2.0±0.0 | 65.0±1.9 | 64.8±1.9 | 52.6±1.9 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 65.4±1.9 | 64.8±1.9 | 53.1±2.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 65.4±1.9 | 64.8±1.9 | 53.1±2.0 |
| P13 ($\delta$=0.319) | 64 | 0.574±0.022 | 0.394±0.022 | 0.466±0.022 | 0.648±0.021 | 0.176±0.017 | 5.9±0.0 | 45.1±0.9 | 44.8±0.8 | 38.6±1.8 |
| gold=39, mode=39 ✓ | 128 | 0.930±0.011 | 0.872±0.015 | 0.898±0.014 | 0.932±0.011 | 0.518±0.022 | 6.7±0.0 | 70.2±1.3 | 66.2±1.3 | 77.9±2.1 |
| | 256 | 1.000±0.000 | 0.996±0.003 | 0.998±0.002 | 0.998±0.002 | 0.962±0.009 | 7.3±0.0 | 81.4±1.8 | 75.2±1.7 | 123.5±2.7 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.7±0.0 | 82.1±1.9 | 75.6±1.7 | 129.7±3.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 8.0±0.0 | 82.1±1.9 | 75.6±1.7 | 129.7±3.0 |
| P23 ($\delta$=0.312) | 64 | 0.570±0.022 | 0.544±0.022 | 0.548±0.022 | 0.768±0.019 | 0.006±0.003 | 19.3±0.1 | 43.7±0.8 | 44.3±0.7 | 8.7±1.7 |
| gold=245, mode=245 ✓ | 128 | 0.942±0.010 | **0.946±0.010** | 0.954±0.009 | 0.964±0.008 | 0.006±0.003 | 29.7±0.2 | 62.5±1.2 | 62.8±1.2 | 8.7±1.7 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 0.006±0.003 | 44.8±0.2 | 67.6±1.5 | 67.2±1.5 | 8.7±1.7 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 65.0±0.2 | 67.6±1.5 | 67.2±1.5 | 8.7±1.7 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 89.1±0.2 | 67.6±1.5 | 67.2±1.5 | 8.7±1.7 |
| P4 ($\delta$=0.199) | 64 | 0.152±0.016 | 0.096±0.013 | 0.100±0.013 | 0.318±0.021 | 0.182±0.017 | 4.8±0.0 | 38.8±2.5 | 41.8±2.2 | 35.8±1.8 |
| gold=70, mode=70 ✓ | 128 | 0.352±0.021 | 0.268±0.020 | 0.274±0.020 | 0.526±0.022 | 0.338±0.021 | 6.1±0.1 | 76.4±2.9 | 76.8±2.7 | 64.3±2.7 |
| | 256 | 0.706±0.020 | 0.594±0.022 | 0.618±0.022 | 0.788±0.018 | 0.606±0.022 | 7.9±0.1 | 141.2±4.0 | 142.0±3.8 | 121.5±4.2 |
| | 512 | 0.978±0.007 | 0.944±0.010 | 0.950±0.010 | 0.962±0.009 | 0.924±0.012 | 10.2±0.1 | 219.8±5.7 | 213.5±5.4 | 201.4±6.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 12.7±0.1 | 241.6±6.8 | 232.6±6.4 | 232.9±7.6 |
| P19 ($\delta$=0.158) | 64 | 0.098±0.013 | 0.050±0.010 | 0.048±0.010 | 0.262±0.020 | 0.106±0.014 | 2.9±0.0 | 28.3±2.9 | 29.9±2.7 | 21.3±2.1 |
| gold=279, mode=279 ✓ | 128 | 0.210±0.018 | 0.106±0.014 | 0.096±0.013 | 0.374±0.022 | 0.166±0.017 | 3.6±0.0 | 68.5±5.7 | 66.5±5.8 | 52.0±4.8 |
| | 256 | 0.482±0.022 | 0.300±0.020 | 0.288±0.020 | 0.590±0.022 | 0.326±0.021 | 4.8±0.0 | 147.8±5.7 | 149.9±5.8 | 122.4±6.5 |
| | 512 | 0.814±0.017 | 0.688±0.021 | 0.664±0.021 | 0.814±0.017 | 0.660±0.021 | 6.0±0.0 | 276.0±7.2 | 278.7±7.4 | 253.0±8.3 |
| | 1024 | 0.988±0.005 | 0.960±0.009 | 0.950±0.010 | 0.974±0.007 | 0.940±0.011 | 7.1±0.0 | 399.0±10.9 | 409.8±11.1 | 393.1±12.0 |
| P10 ($\delta$=0.154) | 64 | 0.050±0.010 | **0.058±0.010** | 0.078±0.012 | 0.270±0.020 | 0.054±0.010 | 10.7±0.1 | 39.2±2.4 | 39.4±2.1 | 42.7±3.3 |

Table 12 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| gold=156, mode=156 ✓ | 128 | 0.130±0.015 | **0.152±0.016** | 0.222±0.019 | 0.412±0.022 | 0.154±0.016 | 16.1±0.1 | 76.9±3.9 | 76.0±3.0 | 78.2±3.6 |
| | 256 | 0.358±0.021 | **0.374±0.022** | 0.516±0.022 | 0.624±0.022 | 0.400±0.022 | 24.4±0.1 | 145.6±5.0 | 141.0±4.2 | 148.4±4.6 |
| | 512 | 0.760±0.019 | **0.776±0.019** | 0.902±0.013 | 0.860±0.016 | 0.770±0.019 | 35.4±0.1 | 263.8±6.8 | 236.6±6.1 | 258.5±6.9 |
| | 1024 | 0.996±0.003 | **0.998±0.002** | 1.000±0.000 | 0.984±0.006 | 0.980±0.006 | 47.9±0.1 | 360.1±10.0 | 274.6±7.7 | 355.0±10.3 |
| P29 ($\delta$=0.131) | 64 | 0.008±0.004 | 0.008±0.004 | 0.008±0.004 | 0.112±0.014 | 0.000±0.000 | 14.9±0.1 | 59.2±3.1 | 59.2±3.1 | — |
| gold=157, mode=12 ✗ | 128 | 0.102±0.014 | 0.086±0.013 | 0.090±0.013 | 0.302±0.021 | 0.000±0.000 | 17.7±0.1 | 100.0±3.5 | 100.3±3.4 | — |
| | 256 | 0.500±0.022 | 0.450±0.022 | 0.452±0.022 | 0.684±0.021 | 0.000±0.000 | 20.4±0.1 | 176.6±3.4 | 173.9±3.4 | — |
| | 512 | 0.954±0.009 | 0.954±0.009 | 0.952±0.010 | 0.962±0.009 | 0.000±0.000 | 23.5±0.1 | 268.6±4.8 | 266.9±4.9 | — |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 0.000±0.000 | 26.4±0.1 | 283.8±5.6 | 282.9±5.7 | — |
| P17 ($\delta$=0.095) | 64 | 0.008±0.004 | **0.016±0.006** | 0.016±0.006 | 0.134±0.015 | 0.000±0.000 | 19.4±0.1 | 54.2±4.8 | 55.4±3.8 | — |
| gold=243, mode=243 ✓ | 128 | 0.046±0.009 | **0.058±0.010** | 0.058±0.010 | 0.246±0.019 | 0.000±0.000 | 29.0±0.2 | 87.7±4.7 | 88.5±4.6 | — |
| | 256 | 0.140±0.016 | **0.214±0.018** | 0.196±0.018 | 0.464±0.022 | 0.000±0.000 | 43.1±0.2 | 169.3±5.8 | 166.1±6.2 | — |
| | 512 | 0.404±0.022 | **0.538±0.022** | 0.504±0.022 | 0.744±0.020 | 0.000±0.000 | 62.1±0.2 | 300.4±7.7 | 305.0±8.2 | — |
| | 1024 | 0.812±0.017 | **0.878±0.015** | 0.846±0.016 | 0.930±0.011 | 0.000±0.000 | 82.6±0.2 | 467.4±11.9 | 478.5±12.2 | — |
| P18 ($\delta$=0.065) | 64 | 0.004±0.003 | 0.002±0.002 | 0.002±0.002 | 0.032±0.008 | 0.000±0.000 | 13.2±0.1 | 28.0±0.0 | 27.0±0.0 | — |
| gold=503, mode=505 ✗ | 128 | 0.014±0.005 | 0.014±0.005 | 0.016±0.006 | 0.080±0.012 | 0.000±0.000 | 18.1±0.1 | 79.0±10.8 | 80.4±10.3 | — |
| | 256 | 0.044±0.009 | **0.056±0.010** | 0.068±0.011 | 0.216±0.018 | 0.000±0.000 | 24.8±0.1 | 169.3±12.3 | 174.1±11.1 | — |
| | 512 | 0.116±0.014 | **0.154±0.016** | 0.182±0.017 | 0.408±0.022 | 0.000±0.000 | 33.0±0.1 | 306.3±14.2 | 311.9±13.5 | — |
| | 1024 | 0.364±0.022 | **0.430±0.022** | 0.490±0.022 | 0.694±0.021 | 0.000±0.000 | 40.9±0.1 | 603.9±18.0 | 599.2±16.6 | — |
| P9 ($\delta$=0.024) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.018±0.006 | 0.000±0.000 | 20.4±0.1 | — | — | — |
| gold=29, mode=29 ✓ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.022±0.007 | 0.000±0.000 | 29.1±0.1 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.008±0.004 | 0.050±0.010 | 0.000±0.000 | 41.3±0.2 | — | 227.5±16.8 | — |
| | 512 | 0.004±0.003 | **0.014±0.005** | 0.032±0.008 | 0.120±0.015 | 0.000±0.000 | 56.3±0.2 | 344.6±32.0 | 357.6±26.8 | — |
| | 1024 | 0.016±0.006 | **0.040±0.009** | 0.098±0.013 | 0.196±0.018 | 0.000±0.000 | 72.1±0.1 | 580.6±46.9 | 628.5±33.7 | — |

## FrontierScience-Olympiad.

Table 13: Per-problem certification rate on FrontierScience-Olympiad (10 problems, every-10 subset) (Case A, target = mode). **Bold**: CITE $>$ Bonferroni. $\bar{K}_N$ is the Monte Carlo mean number of observed answer categories at budget $N$.

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P60 ($\delta$=0.630) | 64 | 1.000±0.000 | 0.996±0.003 | 0.998±0.002 | 0.994±0.003 | 1.000±0.000 | 2.7±0.0 | 23.5±0.5 | 23.4±0.5 | 18.2±0.5 |
| gold=6.98, mode=6.49 ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.1±0.0 | 23.7±0.5 | 23.5±0.5 | 18.2±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.4±0.0 | 23.7±0.5 | 23.5±0.5 | 18.2±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.8±0.0 | 23.7±0.5 | 23.5±0.5 | 18.2±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.3±0.0 | 23.7±0.5 | 23.5±0.5 | 18.2±0.5 |
| P50 ($\delta$=0.524) | 64 | 0.994±0.003 | 0.994±0.003 | 0.994±0.003 | 0.992±0.004 | 0.156±0.016 | 14.1±0.1 | 29.4±0.4 | 30.1±0.4 | 34.7±2.1 |
| gold=0.2, mode=0.2 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.354±0.021 | 19.2±0.1 | 29.7±0.4 | 30.4±0.4 | 70.3±2.7 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.756±0.019 | 24.6±0.1 | 29.7±0.4 | 30.4±0.4 | 133.9±3.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.992±0.004 | 30.4±0.1 | 29.7±0.4 | 30.4±0.4 | 181.8±4.9 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 36.2±0.1 | 29.7±0.4 | 30.4±0.4 | 184.8±5.1 |
| P30 ($\delta$=0.347) | 64 | 0.800±0.018 | **0.840±0.016** | 0.858±0.016 | 0.930±0.011 | 0.006±0.003 | 26.0±0.1 | 42.1±0.6 | 43.5±0.5 | 10.7±1.9 |
| ✗ | 128 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 41.1±0.2 | 47.6±0.8 | 47.9±0.7 | 10.7±1.9 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 61.6±0.2 | 47.6±0.8 | 47.9±0.7 | 10.7±1.9 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 86.9±0.2 | 47.6±0.8 | 47.9±0.7 | 10.7±1.9 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 113.2±0.2 | 47.6±0.8 | 47.9±0.7 | 10.7±1.9 |
| P70 ($\delta$=0.116) | 64 | 0.038±0.009 | 0.028±0.007 | 0.030±0.008 | 0.162±0.016 | 0.040±0.009 | 8.3±0.1 | 42.2±4.7 | 47.3±3.7 | 43.7±3.4 |
| ✗ | 128 | 0.080±0.012 | 0.076±0.012 | 0.080±0.012 | 0.306±0.021 | 0.094±0.013 | 11.3±0.1 | 76.6±5.3 | 77.5±4.6 | 73.3±4.5 |
| | 256 | 0.200±0.018 | 0.178±0.017 | 0.196±0.018 | 0.430±0.022 | 0.210±0.018 | 15.3±0.1 | 137.6±6.7 | 138.8±6.3 | 133.2±6.3 |
| | 512 | 0.494±0.022 | 0.476±0.022 | 0.518±0.022 | 0.696±0.021 | 0.446±0.022 | 20.6±0.1 | 290.7±8.9 | 287.4±8.3 | 267.1±9.7 |
| | 1024 | 0.870±0.015 | 0.850±0.016 | 0.882±0.014 | 0.920±0.012 | 0.804±0.018 | 26.6±0.1 | 484.1±12.6 | 464.5±11.9 | 472.0±13.5 |
| P100 ($\delta$=0.112) | 64 | 0.006±0.003 | **0.014±0.005** | 0.010±0.004 | 0.128±0.015 | 0.004±0.003 | 17.5±0.1 | 36.1±8.0 | 47.0±5.7 | 7.0±0.0 |
| ✗ | 128 | 0.042±0.009 | **0.060±0.011** | 0.060±0.011 | 0.244±0.019 | 0.004±0.003 | 25.2±0.1 | 89.7±6.4 | 95.3±5.0 | 7.0±0.0 |
| | 256 | 0.264±0.020 | **0.306±0.021** | 0.310±0.021 | 0.562±0.022 | 0.004±0.003 | 35.4±0.1 | 182.2±4.6 | 181.6±4.3 | 7.0±0.0 |
| | 512 | 0.662±0.021 | **0.726±0.020** | 0.726±0.020 | 0.860±0.016 | 0.004±0.003 | 47.1±0.1 | 298.5±6.3 | 296.3±6.3 | 7.0±0.0 |
| | 1024 | 0.972±0.007 | **0.978±0.007** | 0.978±0.007 | 0.980±0.006 | 0.004±0.003 | 58.2±0.1 | 394.5±9.2 | 391.4±9.2 | 7.0±0.0 |
| P90 ($\delta$=0.057) | 64 | 0.000±0.000 | **0.004±0.003** | 0.004±0.003 | 0.050±0.010 | 0.000±0.000 | 21.7±0.1 | 54.0±7.0 | 54.0±7.0 | — |
| ✗ | 128 | 0.000±0.000 | **0.010±0.004** | 0.006±0.003 | 0.080±0.012 | 0.000±0.000 | 33.2±0.2 | 87.8±14.5 | 72.0±18.4 | — |
| | 256 | 0.024±0.007 | **0.040±0.009** | 0.042±0.009 | 0.206±0.018 | 0.000±0.000 | 51.8±0.2 | 164.6±12.8 | 175.8±12.2 | — |
| | 512 | 0.072±0.012 | **0.148±0.016** | 0.138±0.015 | 0.380±0.022 | 0.000±0.000 | 79.8±0.3 | 331.7±14.3 | 329.7±14.7 | — |
| | 1024 | 0.244±0.019 | **0.388±0.022** | 0.394±0.022 | 0.638±0.021 | 0.000±0.000 | 115.5±0.2 | 591.1±17.7 | 606.5±17.6 | — |
| P80 ($\delta$=0.044) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.028±0.007 | 0.000±0.000 | 30.8±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | **0.004±0.003** | 0.004±0.003 | 0.098±0.013 | 0.000±0.000 | 45.0±0.2 | 103.5±4.5 | 103.5±4.5 | — |
| | 256 | 0.014±0.005 | **0.034±0.008** | 0.040±0.009 | 0.216±0.018 | 0.000±0.000 | 63.0±0.2 | 184.5±10.8 | 186.1±9.9 | — |
| | 512 | 0.058±0.010 | **0.136±0.015** | 0.144±0.016 | 0.378±0.022 | 0.000±0.000 | 85.6±0.2 | 346.4±14.0 | 339.3±13.8 | — |
| | 1024 | 0.216±0.018 | **0.398±0.022** | 0.396±0.022 | 0.616±0.022 | 0.000±0.000 | 110.4±0.2 | 608.2±16.5 | 606.3±17.4 | — |
| P10 ($\delta$=0.009) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 52.6±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 93.0±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.002 | 0.000±0.000 | 156.6±0.3 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.006±0.003 | 0.000±0.000 | 248.5±0.4 | — | — | — |
| | 1024 | 0.000±0.000 | **0.004±0.003** | 0.004±0.003 | 0.054±0.010 | 0.000±0.000 | 360.0±0.4 | 613.0±38.0 | 613.0±38.0 | — |
| P20 ($\delta$=0.007) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 58.6±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 109.8±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.006±0.003 | 0.000±0.000 | 198.6±0.3 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.028±0.007 | 0.000±0.000 | 338.1±0.4 | — | — | — |
| | 1024 | 0.000±0.000 | **0.006±0.003** | 0.010±0.004 | 0.078±0.012 | 0.000±0.000 | 520.0±0.5 | 704.0±127.6 | 733.4±87.0 | — |
| P40 ($\delta$=0.007) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 47.1±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.002 | 0.000±0.000 | 79.1±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.004±0.003 | 0.000±0.000 | 126.6±0.3 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.016±0.006 | 0.000±0.000 | 188.3±0.3 | — | — | — |
| | 1024 | 0.000±0.000 | **0.006±0.003** | 0.006±0.003 | 0.048±0.010 | 0.000±0.000 | 255.4±0.3 | 865.3±62.6 | 875.3±72.5 | — |

## ER-REASON.

Table 14: Per-problem certification rate on ER-REASON (Case A, target = mode). **Bold**: CITE $>$ Bonferroni. $\bar{K}_N$ is the Monte Carlo mean number of observed answer categories at budget $N$.

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P1 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |

Table 14 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P2 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P3 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P4 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P5 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P6 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P7 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P8 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P9 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P10 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P11 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P12 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P13 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P14 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P15 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P16 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P17 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P18 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P19 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P20 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |

Table 14 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P21 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P22 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P23 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P24 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P25 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P26 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P27 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| P29 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P30 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P32 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P34 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P36 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P37 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P38 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P39 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P41 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P43 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P44 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P45 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |

Table 14 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P46 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P47 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P48 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P49 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P50 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P51 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P52 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P53 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P54 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P55 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P56 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P57 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 9.0±0.0 | 7.0±0.0 |
| P59 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P60 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P61 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P62 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P63 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P64 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✓ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P65 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P66 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |

Table 14 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P67 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P69 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P70 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P71 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P72 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 10.0±0.0 | 7.0±0.0 |
| P35 ($\delta$=0.918) | 64 | 0.934±0.011 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 11.1±0.2 | 11.9±0.1 | 8.6±0.1 |
| ✗ | 128 | 0.994±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.1±0.2 | 11.9±0.1 | 8.6±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.1±0.2 | 11.9±0.1 | 8.6±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.1±0.2 | 11.9±0.1 | 8.6±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.1±0.2 | 11.9±0.1 | 8.6±0.1 |
| P58 ($\delta$=0.908) | 64 | 0.956±0.009 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.4±0.2 | 12.2±0.1 | 8.7±0.1 |
| ✗ | 128 | 0.994±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.4±0.2 | 12.2±0.1 | 8.7±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.4±0.2 | 12.2±0.1 | 8.7±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.4±0.2 | 12.2±0.1 | 8.7±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.4±0.2 | 12.2±0.1 | 8.7±0.1 |
| P33 ($\delta$=0.754) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 17.1±0.4 | 16.9±0.3 | 12.7±0.3 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 17.1±0.4 | 16.9±0.3 | 12.7±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 17.1±0.4 | 16.9±0.3 | 12.7±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 17.1±0.4 | 16.9±0.3 | 12.7±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 17.1±0.4 | 16.9±0.3 | 12.7±0.3 |
| P31 ($\delta$=0.718) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 18.2±0.4 | 18.1±0.3 | 13.8±0.3 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 18.2±0.4 | 18.1±0.3 | 13.8±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 18.2±0.4 | 18.1±0.3 | 13.8±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 18.2±0.4 | 18.1±0.3 | 13.8±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 18.2±0.4 | 18.1±0.3 | 13.8±0.3 |
| P42 ($\delta$=0.704) | 64 | 1.000±0.000 | 0.998±0.002 | 1.000±0.000 | 0.998±0.002 | 0.996±0.003 | 2.0±0.0 | 20.4±0.5 | 19.8±0.4 | 15.5±0.4 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 20.5±0.5 | 19.8±0.4 | 15.8±0.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 20.5±0.5 | 19.8±0.4 | 15.8±0.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 20.5±0.5 | 19.8±0.4 | 15.8±0.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 20.5±0.5 | 19.8±0.4 | 15.8±0.4 |
| P28 ($\delta$=0.470) | 64 | 0.950±0.010 | 0.758±0.019 | 0.820±0.017 | 0.870±0.015 | 0.844±0.016 | 2.0±0.0 | 33.9±0.8 | 33.5±0.7 | 26.6±0.8 |
| ✗ | 128 | 0.998±0.002 | 0.988±0.005 | 0.996±0.003 | 0.990±0.004 | 0.990±0.004 | 2.0±0.0 | 45.2±1.1 | 41.9±1.1 | 35.3±1.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 46.2±1.2 | 42.3±1.1 | 36.4±1.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 46.2±1.2 | 42.3±1.1 | 36.4±1.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 46.2±1.2 | 42.3±1.1 | 36.4±1.3 |
| P68 ($\delta$=0.438) | 64 | 0.890±0.014 | 0.686±0.021 | 0.626±0.022 | 0.840±0.016 | 0.788±0.018 | 2.0±0.0 | 36.0±0.9 | 37.6±0.9 | 28.2±0.9 |
| ✓ | 128 | 0.994±0.003 | 0.976±0.007 | 0.964±0.008 | 0.978±0.007 | 0.970±0.008 | 2.0±0.0 | 51.0±1.3 | 54.8±1.3 | 39.8±1.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 53.7±1.5 | 58.6±1.5 | 43.4±1.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 53.7±1.5 | 58.6±1.5 | 43.4±1.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 53.7±1.5 | 58.6±1.5 | 43.4±1.6 |
| P40 ($\delta$=0.326) | 64 | 0.600±0.022 | 0.356±0.021 | 0.354±0.021 | 0.590±0.022 | 0.494±0.022 | 2.0±0.0 | 37.0±1.2 | 37.5±1.2 | 30.5±1.1 |
| ✗ | 128 | 0.908±0.013 | 0.736±0.020 | 0.732±0.020 | 0.864±0.015 | 0.768±0.019 | 2.0±0.0 | 65.9±1.7 | 65.8±1.7 | 53.1±1.8 |
| | 256 | 1.000±0.000 | 0.980±0.006 | 0.980±0.006 | 0.968±0.008 | 0.976±0.007 | 2.0±0.0 | 91.7±2.5 | 92.0±2.5 | 80.0±2.8 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 95.5±2.7 | 95.6±2.7 | 85.0±3.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 95.5±2.7 | 95.6±2.7 | 85.0±3.1 |

## H.4 Per-problem certification rates: gpt-oss-20b

## AIME 2026.

Table 15: Per-problem certification rate on AIME 2026 (30 problems) (Case A, target = mode). **Bold**: CITE $>$ Bonferroni. $\bar{K}_N$ is the Monte Carlo mean number of observed answer categories at budget $N$.

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P6 ($\delta$=0.995) | 64 | 0.222±0.019 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.2±0.0 | 9.2±0.0 | 9.2±0.0 | 7.2±0.0 |
| gold=441, mode=441 ✓ | 128 | 0.348±0.021 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.4±0.0 | 9.2±0.0 | 9.2±0.0 | 7.2±0.0 |
| | 256 | 0.548±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.6±0.0 | 9.2±0.0 | 9.2±0.0 | 7.2±0.0 |
| | 512 | 0.792±0.018 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 9.2±0.0 | 9.2±0.0 | 7.2±0.0 |
| | 1024 | 0.968±0.008 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.6±0.0 | 9.2±0.0 | 9.2±0.0 | 7.2±0.0 |
| P1 ($\delta$=0.970) | 64 | 0.636±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.7±0.0 | 9.6±0.1 | 9.6±0.1 | 7.5±0.1 |
| gold=277, mode=277 ✓ | 128 | 0.852±0.016 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 9.6±0.1 | 9.6±0.1 | 7.5±0.1 |
| | 256 | 0.976±0.007 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 9.6±0.1 | 9.6±0.1 | 7.5±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.8±0.0 | 9.6±0.1 | 9.6±0.1 | 7.5±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.3±0.0 | 9.6±0.1 | 9.6±0.1 | 7.5±0.1 |
| P24 ($\delta$=0.904) | 64 | 0.992±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.4±0.1 | 12.1±0.1 | 12.4±0.1 | 10.7±0.2 |
| gold=669, mode=669 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 8.8±0.1 | 12.1±0.1 | 12.4±0.1 | 10.7±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 13.6±0.1 | 12.1±0.1 | 12.4±0.1 | 10.7±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 19.4±0.1 | 12.1±0.1 | 12.4±0.1 | 10.7±0.2 |

Table 15 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 24.9±0.1 | 12.1±0.1 | 12.4±0.1 | 10.7±0.2 |
| P16 ($\delta$=0.893) | 64 | 0.976±0.007 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 11.8±0.2 | 11.8±0.2 | 9.2±0.1 |
| gold=178, mode=178 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 11.8±0.2 | 11.8±0.2 | 9.2±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 11.8±0.2 | 11.8±0.2 | 9.2±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.2±0.0 | 11.8±0.2 | 11.8±0.2 | 9.2±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 11.8±0.2 | 11.8±0.2 | 9.2±0.1 |
| P15 ($\delta$=0.878) | 64 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.3±0.1 | 12.9±0.2 | 12.9±0.2 | 11.3±0.3 |
| gold=83, mode=1 ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.8±0.1 | 12.9±0.2 | 12.9±0.2 | 11.3±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 11.3±0.1 | 12.9±0.2 | 12.9±0.2 | 11.3±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 15.8±0.1 | 12.9±0.2 | 12.9±0.2 | 11.3±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 20.7±0.1 | 12.9±0.2 | 12.9±0.2 | 11.3±0.3 |
| P22 ($\delta$=0.868) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.8±0.1 | 13.5±0.1 | 13.5±0.1 | 13.5±0.3 |
| gold=754, mode=754 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 13.1±0.1 | 13.5±0.1 | 13.5±0.1 | 13.5±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 20.8±0.1 | 13.5±0.1 | 13.5±0.1 | 13.5±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 30.6±0.1 | 13.5±0.1 | 13.5±0.1 | 13.5±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 41.0±0.1 | 13.5±0.1 | 13.5±0.1 | 13.5±0.3 |
| P2 ($\delta$=0.777) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.4±0.1 | 15.9±0.2 | 15.9±0.2 | 16.6±0.4 |
| gold=62, mode=62 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 11.0±0.1 | 15.9±0.2 | 15.9±0.2 | 16.6±0.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 16.0±0.1 | 15.9±0.2 | 15.9±0.2 | 16.6±0.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 22.1±0.1 | 15.9±0.2 | 15.9±0.2 | 16.6±0.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 28.6±0.1 | 15.9±0.2 | 15.9±0.2 | 16.6±0.4 |
| P18 ($\delta$=0.741) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.996±0.003 | 6.7±0.1 | 17.0±0.2 | 17.0±0.2 | 17.4±0.4 |
| gold=503, mode=1 ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 9.0±0.1 | 17.0±0.2 | 17.0±0.2 | 17.6±0.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 11.3±0.1 | 17.0±0.2 | 17.0±0.2 | 17.6±0.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 13.5±0.1 | 17.0±0.2 | 17.0±0.2 | 17.6±0.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 15.6±0.1 | 17.0±0.2 | 17.0±0.2 | 17.6±0.4 |
| P7 ($\delta$=0.733) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.958±0.009 | 12.4±0.1 | 17.6±0.2 | 17.6±0.2 | 27.0±0.7 |
| gold=396, mode=396 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 19.9±0.1 | 17.6±0.2 | 17.6±0.2 | 29.2±0.8 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 30.2±0.2 | 17.6±0.2 | 17.6±0.2 | 29.2±0.8 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 43.1±0.2 | 17.6±0.2 | 17.6±0.2 | 29.2±0.8 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 56.5±0.1 | 17.6±0.2 | 17.6±0.2 | 29.2±0.8 |
| P8 ($\delta$=0.700) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 20.9±0.4 | 20.9±0.4 | 18.1±0.4 |
| gold=244, mode=244 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 20.9±0.4 | 20.9±0.4 | 18.1±0.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 20.9±0.4 | 20.9±0.4 | 18.1±0.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 20.9±0.4 | 20.9±0.4 | 18.1±0.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 20.9±0.4 | 20.9±0.4 | 18.1±0.4 |
| P3 ($\delta$=0.699) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 4.7±0.0 | 20.5±0.4 | 20.4±0.3 | 20.4±0.5 |
| gold=79, mode=79 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.7±0.0 | 20.5±0.4 | 20.4±0.3 | 20.5±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.6±0.0 | 20.5±0.4 | 20.4±0.3 | 20.5±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.4±0.0 | 20.5±0.4 | 20.4±0.3 | 20.5±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 8.2±0.0 | 20.5±0.4 | 20.4±0.3 | 20.5±0.5 |
| P30 ($\delta$=0.675) | 64 | 1.000±0.000 | 0.998±0.002 | 0.998±0.002 | 0.998±0.002 | 1.000±0.000 | 2.4±0.0 | 21.9±0.5 | 21.9±0.5 | 16.0±0.5 |
| gold=393, mode=243 ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 22.0±0.5 | 22.0±0.5 | 16.0±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.2±0.0 | 22.0±0.5 | 22.0±0.5 | 16.0±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.8±0.0 | 22.0±0.5 | 22.0±0.5 | 16.0±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 22.0±0.5 | 22.0±0.5 | 16.0±0.5 |
| P5 ($\delta$=0.488) | 64 | 0.940±0.011 | 0.804±0.018 | 0.804±0.018 | 0.928±0.012 | 0.876±0.015 | 3.8±0.1 | 32.1±0.8 | 32.1±0.8 | 26.1±0.8 |
| gold=65, mode=65 ✓ | 128 | 0.996±0.003 | 0.988±0.005 | 0.988±0.005 | 0.996±0.003 | 0.994±0.003 | 5.2±0.1 | 41.3±1.1 | 41.3±1.1 | 32.9±1.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.5±0.1 | 42.6±1.2 | 42.6±1.2 | 33.7±1.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 10.6±0.1 | 42.6±1.2 | 42.6±1.2 | 33.7±1.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 14.5±0.1 | 42.6±1.2 | 42.6±1.2 | 33.7±1.2 |
| P19 ($\delta$=0.444) | 64 | 0.972±0.007 | 0.964±0.008 | 0.964±0.008 | 0.960±0.009 | 0.032±0.008 | 16.6±0.1 | 36.3±0.5 | 36.3±0.5 | 24.4±4.8 |
| gold=279, mode=279 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.050±0.010 | 23.4±0.1 | 37.6±0.6 | 37.6±0.6 | 51.1±8.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.086±0.013 | 31.6±0.1 | 37.6±0.6 | 37.6±0.6 | 107.0±11.9 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.220±0.019 | 39.8±0.1 | 37.6±0.6 | 37.6±0.6 | 288.7±15.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.530±0.022 | 47.3±0.1 | 37.6±0.6 | 37.6±0.6 | 548.1±16.4 |
| P20 ($\delta$=0.415) | 64 | 0.840±0.016 | 0.748±0.019 | 0.752±0.019 | 0.870±0.015 | 0.684±0.021 | 7.5±0.1 | 37.1±0.7 | 37.1±0.7 | 39.0±0.9 |
| gold=190, mode=190 ✓ | 128 | 1.000±0.000 | 0.996±0.003 | 0.998±0.002 | 0.988±0.005 | 0.992±0.004 | 10.6±0.1 | 48.9±1.1 | 48.8±1.1 | 53.4±1.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 14.5±0.1 | 49.2±1.1 | 49.0±1.1 | 54.1±1.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 18.7±0.1 | 49.2±1.1 | 49.0±1.1 | 54.1±1.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 22.3±0.1 | 49.2±1.1 | 49.0±1.1 | 54.1±1.2 |
| P28 ($\delta$=0.402) | 64 | 0.878±0.015 | 0.850±0.016 | 0.898±0.014 | 0.922±0.012 | 0.070±0.011 | 13.9±0.1 | 38.0±0.6 | 37.9±0.6 | 32.9±3.2 |
| gold=107, mode=12 ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.996±0.003 | 0.166±0.017 | 18.3±0.1 | 44.0±0.8 | 42.0±0.8 | 75.1±4.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.484±0.022 | 22.5±0.1 | 44.0±0.8 | 42.0±0.8 | 155.0±4.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.880±0.015 | 27.0±0.1 | 44.0±0.8 | 42.0±0.8 | 245.5±5.8 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 31.3±0.1 | 44.0±0.8 | 42.0±0.8 | 290.5±7.6 |
| P12 ($\delta$=0.345) | 64 | 0.630±0.022 | 0.624±0.022 | 0.624±0.022 | 0.794±0.018 | 0.116±0.014 | 16.0±0.1 | 38.5±0.8 | 38.2±0.8 | 38.4±2.4 |
| gold=161, mode=161 ✓ | 128 | 0.964±0.008 | **0.972±0.007** | 0.970±0.008 | 0.978±0.007 | 0.326±0.021 | 25.1±0.1 | 56.8±1.3 | 56.4±1.3 | 75.9±2.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.756±0.019 | 37.9±0.2 | 59.4±1.4 | 59.2±1.5 | 140.5±3.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.984±0.006 | 55.1±0.2 | 59.4±1.4 | 59.2±1.5 | 183.6±4.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 74.0±0.2 | 59.4±1.4 | 59.2±1.5 | 191.0±5.2 |
| P10 ($\delta$=0.315) | 64 | 0.592±0.022 | **0.654±0.021** | 0.656±0.021 | 0.838±0.016 | 0.006±0.003 | 24.2±0.1 | 44.3±0.7 | 44.3±0.7 | 7.0±0.0 |
| gold=156, mode=168 ✗ | 128 | 0.972±0.007 | **0.982±0.006** | 0.982±0.006 | 0.990±0.004 | 0.006±0.003 | 39.5±0.2 | 57.9±1.1 | 57.8±1.1 | 7.0±0.0 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 61.9±0.2 | 59.3±1.1 | 59.3±1.1 | 7.0±0.0 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 88.9±0.2 | 59.3±1.1 | 59.3±1.1 | 7.0±0.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.006±0.003 | 114.6±0.2 | 59.3±1.1 | 59.3±1.1 | 7.0±0.0 |
| P29 ($\delta$=0.314) | 64 | 0.652±0.021 | 0.604±0.022 | 0.602±0.022 | 0.788±0.018 | 0.004±0.003 | 14.7±0.1 | 44.6±0.7 | 45.0±0.7 | 27.0±11.0 |
| gold=157, mode=0 ✗ | 128 | 0.974±0.007 | 0.974±0.007 | 0.974±0.007 | 0.982±0.006 | 0.004±0.003 | 18.7±0.1 | 59.4±1.0 | 59.9±1.0 | 27.0±11.0 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.004±0.003 | 23.1±0.1 | 61.6±1.2 | 62.1±1.2 | 27.0±11.0 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.004±0.003 | 27.6±0.1 | 61.6±1.2 | 62.1±1.2 | 27.0±11.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.004±0.003 | 31.9±0.1 | 61.6±1.2 | 62.1±1.2 | 27.0±11.0 |
| P26 ($\delta$=0.264) | 64 | 0.368±0.022 | **0.422±0.022** | 0.420±0.022 | 0.650±0.021 | 0.000±0.000 | 20.5±0.1 | 44.8±0.9 | 45.3±0.8 | — |
| gold=132, mode=66 ✗ | 128 | 0.824±0.017 | **0.860±0.016** | 0.860±0.016 | 0.926±0.012 | 0.000±0.000 | 30.1±0.1 | 68.6±1.3 | 69.2±1.3 | — |
| | 256 | 0.994±0.003 | **0.998±0.002** | 0.998±0.002 | 0.998±0.002 | 0.000±0.000 | 41.6±0.1 | 81.3±1.9 | 81.9±1.9 | — |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.000±0.000 | 54.9±0.2 | 81.7±1.9 | 82.3±1.9 | — |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.000±0.000 | 68.3±0.1 | 81.7±1.9 | 82.3±1.9 | — |
| P25 ($\delta$=0.182) | 64 | 0.110±0.014 | 0.106±0.014 | 0.126±0.015 | 0.370±0.022 | 0.002±0.002 | 15.1±0.1 | 48.6±1.5 | 48.6±1.4 | 7.0±0.0 |
| gold=850, mode=850 ✓ | 128 | 0.330±0.021 | **0.344±0.021** | 0.366±0.022 | 0.632±0.022 | 0.002±0.002 | 20.7±0.1 | 80.4±2.1 | 80.5±2.1 | 7.0±0.0 |
| | 256 | 0.774±0.019 | **0.782±0.018** | 0.804±0.018 | 0.874±0.015 | 0.002±0.002 | 27.3±0.1 | 139.8±3.1 | 137.3±3.1 | 7.0±0.0 |
| | 512 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 0.988±0.005 | 0.002±0.002 | 34.2±0.1 | 180.7±4.4 | 173.8±4.3 | 7.0±0.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.002±0.002 | 41.1±0.1 | 180.7±4.4 | 173.8±4.3 | 7.0±0.0 |
| P11 ($\delta$=0.142) | 64 | 0.018±0.006 | **0.036±0.008** | 0.036±0.008 | 0.278±0.020 | 0.000±0.000 | 30.6±0.2 | 48.6±2.1 | 48.5±2.1 | — |
| gold=896, mode=896 ✓ | 128 | 0.190±0.018 | **0.278±0.020** | 0.280±0.020 | 0.568±0.022 | 0.000±0.000 | 49.2±0.2 | 90.8±2.0 | 90.3±2.0 | — |
| | 256 | 0.634±0.022 | **0.760±0.019** | 0.758±0.019 | 0.854±0.016 | 0.000±0.000 | 75.7±0.2 | 150.7±2.9 | 149.6±2.9 | — |
| | 512 | 0.960±0.009 | **0.984±0.006** | 0.984±0.006 | 0.988±0.005 | 0.000±0.000 | 109.3±0.3 | 195.1±4.5 | 194.0±4.5 | — |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.000±0.000 | 145.2±0.2 | 201.2±4.9 | 200.1±4.9 | — |

Table 15 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P27 ($\delta$=0.141) | 64 | 0.014±0.005 | **0.018±0.006** | 0.022±0.007 | 0.176±0.017 | 0.000±0.000 | 22.5±0.1 | 47.9±3.9 | 53.4±2.9 | — |
| gold=223, mode=13 ✗ | 128 | 0.134±0.015 | **0.182±0.017** | 0.234±0.019 | 0.438±0.022 | 0.000±0.000 | 33.3±0.2 | 97.9±2.4 | 95.6±2.0 | — |
| | 256 | 0.594±0.022 | **0.668±0.021** | 0.780±0.019 | 0.808±0.018 | 0.000±0.000 | 48.7±0.2 | 168.6±3.0 | 160.6±2.7 | — |
| | 512 | 0.974±0.007 | **0.986±0.005** | 0.996±0.003 | 0.984±0.006 | 0.000±0.000 | 68.2±0.2 | 223.5±4.5 | 195.2±3.8 | — |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.000±0.000 | 89.5±0.2 | 228.3±4.8 | 196.7±3.9 | — |
| P17 ($\delta$=0.141) | 64 | 0.040±0.009 | 0.038±0.009 | 0.066±0.011 | 0.242±0.019 | 0.000±0.000 | 19.4±0.1 | 52.2±1.9 | 52.9±1.3 | — |
| gold=243, mode=32 ✗ | 128 | 0.180±0.017 | **0.204±0.018** | 0.364±0.022 | 0.474±0.022 | 0.000±0.000 | 26.4±0.1 | 93.8±2.6 | 90.2±1.8 | — |
| | 256 | 0.634±0.022 | **0.668±0.021** | 0.900±0.013 | 0.802±0.018 | 0.000±0.000 | 34.7±0.1 | 160.1±3.0 | 139.9±2.4 | — |
| | 512 | 0.962±0.009 | **0.974±0.007** | 1.000±0.000 | 0.968±0.008 | 0.000±0.000 | 43.7±0.1 | 215.4±4.6 | 155.8±3.1 | — |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.000±0.000 | 53.0±0.1 | 226.6±5.5 | 155.8±3.1 | — |
| P23 ($\delta$=0.068) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.020±0.006 | 0.000±0.000 | 21.3±0.1 | — | — | — |
| gold=245, mode=20 ✗ | 128 | 0.006±0.003 | **0.008±0.004** | 0.008±0.004 | 0.104±0.014 | 0.000±0.000 | 28.6±0.1 | 103.8±12.6 | 103.8±12.6 | — |
| | 256 | 0.044±0.009 | **0.068±0.011** | 0.074±0.012 | 0.276±0.020 | 0.000±0.000 | 37.5±0.1 | 196.2±8.0 | 196.3±7.6 | — |
| | 512 | 0.262±0.020 | **0.360±0.021** | 0.368±0.022 | 0.616±0.022 | 0.000±0.000 | 47.7±0.1 | 359.2±7.7 | 356.4±7.8 | — |
| | 1024 | 0.744±0.020 | **0.814±0.017** | 0.808±0.018 | 0.882±0.014 | 0.000±0.000 | 57.1±0.1 | 566.0±11.1 | 555.4±11.0 | — |
| P21 ($\delta$=0.039) | 64 | 0.004±0.003 | 0.002±0.002 | 0.002±0.002 | 0.080±0.012 | 0.024±0.007 | 8.7±0.1 | 48.0±0.0 | 48.0±0.0 | 24.8±5.2 |
| gold=50, mode=9 ✗ | 128 | 0.002±0.002 | **0.004±0.003** | 0.004±0.003 | 0.098±0.013 | 0.032±0.008 | 13.7±0.1 | 60.5±12.5 | 60.5±12.5 | 39.5±7.9 |
| | 256 | 0.000±0.000 | **0.006±0.003** | 0.006±0.003 | 0.120±0.015 | 0.034±0.008 | 21.0±0.1 | 115.7±55.6 | 115.7±55.6 | 47.6±11.0 |
| | 512 | 0.004±0.003 | **0.012±0.005** | 0.012±0.005 | 0.176±0.017 | 0.038±0.009 | 30.5±0.1 | 236.7±62.7 | 236.7±62.7 | 76.2±22.1 |
| | 1024 | 0.020±0.006 | **0.034±0.008** | 0.036±0.008 | 0.264±0.020 | 0.054±0.010 | 40.8±0.1 | 586.9±75.8 | 574.6±69.5 | 274.6±64.2 |
| P4 ($\delta$=0.026) | 64 | 0.000±0.000 | **0.002±0.002** | 0.002±0.002 | 0.066±0.011 | 0.002±0.002 | 9.9±0.1 | 19.0±0.0 | 19.0±0.0 | 60.0±0.0 |
| gold=70, mode=65 ✗ | 128 | 0.002±0.002 | **0.004±0.003** | 0.004±0.003 | 0.088±0.013 | 0.004±0.003 | 12.7±0.1 | 58.5±39.5 | 58.5±39.5 | 67.0±7.0 |
| | 256 | 0.006±0.003 | **0.008±0.004** | 0.006±0.003 | 0.092±0.013 | 0.010±0.004 | 16.0±0.1 | 146.0±53.3 | 121.3±66.8 | 129.8±26.4 |
| | 512 | 0.016±0.006 | 0.016±0.006 | 0.014±0.005 | 0.114±0.014 | 0.014±0.005 | 19.8±0.1 | 245.5±47.1 | 249.1±54.2 | 176.4±35.4 |
| | 1024 | 0.014±0.005 | **0.022±0.007** | 0.022±0.007 | 0.178±0.017 | 0.026±0.007 | 23.6±0.1 | 373.4±75.7 | 410.4±77.3 | 392.5±72.4 |
| P13 ($\delta$=0.022) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.002 | 0.000±0.000 | 8.5±0.1 | — | — | — |
| gold=39, mode=502 ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.004±0.003 | 0.000±0.000 | 10.2±0.1 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.004±0.003 | 0.000±0.000 | 12.2±0.1 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.018±0.006 | 0.000±0.000 | 13.9±0.0 | — | — | — |
| | 1024 | 0.002±0.002 | 0.002±0.002 | 0.002±0.002 | 0.034±0.008 | 0.000±0.000 | 14.8±0.0 | 661.0±0.0 | 661.0±0.0 | — |
| P14 ($\delta$=0.012) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 29.8±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 42.5±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.018±0.006 | 0.000±0.000 | 58.5±0.2 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.034±0.008 | 0.000±0.000 | 77.7±0.2 | — | — | — |
| | 1024 | 0.000±0.000 | **0.004±0.003** | 0.000±0.000 | 0.060±0.011 | 0.000±0.000 | 97.3±0.2 | 861.5±134.5 | — | — |
| P9 ($\delta$=0.002) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 35.0±0.1 | — | — | — |
| gold=29, mode=13 ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 53.1±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.002 | 0.000±0.000 | 77.5±0.2 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.016±0.006 | 0.000±0.000 | 108.4±0.3 | — | — | — |
| | 1024 | 0.000±0.000 | **0.006±0.003** | 0.000±0.000 | 0.028±0.007 | 0.000±0.000 | 142.0±0.2 | 798.0±49.8 | — | — |

## FrontierScience-Olympiad.

Table 16: Per-problem certification rate on FrontierScience-Olympiad (10 problems) (Case A, target = mode). **Bold**: CITE $>$ Bonferroni. $\bar{K}_N$ is the Monte Carlo mean number of observed answer categories at budget $N$.

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P60 ($\delta$=0.763) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.2±0.1 | 16.1±0.3 | 17.0±0.3 | 14.3±0.3 |
| gold=6.98, mode=6.98 ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.3±0.1 | 16.1±0.3 | 17.0±0.3 | 14.3±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 9.8±0.1 | 16.1±0.3 | 17.0±0.3 | 14.3±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 12.3±0.1 | 16.1±0.3 | 17.0±0.3 | 14.3±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 14.4±0.0 | 16.1±0.3 | 17.0±0.3 | 14.3±0.3 |
| P50 ($\delta$=0.474) | 64 | 0.926±0.012 | 0.862±0.015 | 0.474±0.022 | 0.934±0.011 | 0.916±0.012 | 6.8±0.1 | 33.4±0.7 | 40.3±0.9 | 30.0±0.7 |
| gold=0.2, mode=0.2 ✓ | 128 | 0.998±0.002 | 0.994±0.003 | 0.882±0.014 | 0.996±0.003 | 0.994±0.003 | 9.9±0.1 | 39.6±0.9 | 63.7±1.4 | 33.7±0.9 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 1.000±0.000 | 1.000±0.000 | 14.1±0.1 | 40.2±1.0 | 74.9±1.9 | 34.3±0.9 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 18.7±0.1 | 40.2±1.0 | 75.3±1.9 | 34.3±0.9 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 23.0±0.1 | 40.2±1.0 | 75.3±1.9 | 34.3±0.9 |
| P80 ($\delta$=0.107) | 64 | 0.002±0.002 | **0.004±0.003** | 0.000±0.000 | 0.122±0.015 | 0.000±0.000 | 38.1±0.2 | 47.5±6.5 | — | — |
| ✓ | 128 | 0.052±0.010 | **0.124±0.015** | 0.000±0.000 | 0.354±0.021 | 0.000±0.000 | 64.6±0.2 | 103.3±2.4 | — | — |
| | 256 | 0.280±0.020 | **0.468±0.022** | 0.000±0.000 | 0.700±0.020 | 0.000±0.000 | 104.7±0.3 | 168.9±3.3 | — | — |
| | 512 | 0.860±0.016 | **0.954±0.009** | 0.050±0.010 | 0.952±0.010 | 0.000±0.000 | 160.3±0.3 | 261.3±5.0 | 448.2±9.4 | — |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 0.514±0.022 | 1.000±0.000 | 0.000±0.000 | 224.4±0.3 | 277.4±5.8 | 737.7±9.9 | — |
| P30 ($\delta$=0.077) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.022±0.007 | 0.000±0.000 | 36.4±0.1 | — | — | — |
| ✗ | 128 | 0.002±0.002 | **0.012±0.005** | 0.000±0.000 | 0.116±0.014 | 0.000±0.000 | 56.4±0.2 | 111.3±9.3 | — | — |
| | 256 | 0.086±0.013 | **0.192±0.018** | 0.014±0.005 | 0.370±0.022 | 0.000±0.000 | 83.3±0.2 | 199.0±4.2 | 234.7±5.1 | — |
| | 512 | 0.500±0.022 | **0.700±0.020** | 0.338±0.021 | 0.816±0.017 | 0.000±0.000 | 115.9±0.2 | 332.7±5.7 | 398.9±5.9 | — |
| | 1024 | 0.984±0.006 | **0.990±0.004** | 0.878±0.015 | 0.986±0.005 | 0.000±0.000 | 149.9±0.2 | 422.3±7.9 | 586.3±9.0 | — |
| P40 ($\delta$=0.036) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.002 | 0.000±0.000 | 40.8±0.2 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.038±0.009 | 0.000±0.000 | 66.1±0.2 | — | — | — |
| | 256 | 0.000±0.000 | **0.022±0.007** | 0.000±0.000 | 0.108±0.014 | 0.000±0.000 | 102.0±0.3 | 221.8±11.4 | — | — |
| | 512 | 0.028±0.007 | **0.108±0.014** | 0.028±0.007 | 0.286±0.020 | 0.000±0.000 | 149.2±0.3 | 351.3±13.1 | 476.7±8.0 | — |
| | 1024 | 0.204±0.018 | **0.398±0.022** | 0.756±0.019 | 0.624±0.022 | 0.000±0.000 | 203.1±0.3 | 649.5±16.3 | 759.0±7.6 | — |
| P90 ($\delta$=0.029) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.002±0.002 | 0.000±0.000 | 53.5±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.016±0.006 | 0.000±0.000 | 97.3±0.2 | — | — | — |
| | 256 | 0.000±0.000 | **0.006±0.003** | 0.000±0.000 | 0.042±0.009 | 0.000±0.000 | 169.0±0.3 | 221.7±11.6 | — | — |
| | 512 | 0.008±0.004 | **0.038±0.009** | 0.002±0.002 | 0.220±0.019 | 0.000±0.000 | 276.8±0.4 | 391.5±20.7 | 498.0±0.0 | — |
| | 1024 | 0.092±0.013 | **0.292±0.020** | 0.136±0.015 | 0.540±0.022 | 0.000±0.000 | 412.3±0.4 | 732.0±15.8 | 823.7±16.0 | — |
| P10 ($\delta$=0.020) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.004±0.003 | 0.000±0.000 | 45.3±0.2 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.026±0.007 | 0.000±0.000 | 77.7±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.052±0.010 | 0.000±0.000 | 129.6±0.3 | — | — | — |
| | 512 | 0.004±0.003 | **0.020±0.006** | 0.000±0.000 | 0.114±0.014 | 0.000±0.000 | 206.2±0.4 | 376.7±25.1 | — | — |
| | 1024 | 0.004±0.003 | **0.056±0.010** | 0.000±0.000 | 0.258±0.020 | 0.000±0.000 | 300.6±0.4 | 649.1±44.6 | — | — |
| P70 ($\delta$=0.017) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 41.4±0.2 | — | — | — |
| ✓ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 68.1±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.016±0.006 | 0.000±0.000 | 108.3±0.3 | — | — | — |
| | 512 | 0.000±0.000 | **0.002±0.002** | 0.000±0.000 | 0.052±0.010 | 0.000±0.000 | 163.0±0.3 | 300.0±0.0 | — | — |
| | 1024 | 0.002±0.002 | **0.012±0.005** | 0.000±0.000 | 0.150±0.016 | 0.000±0.000 | 227.0±0.3 | 696.8±105.3 | — | — |
| P20 ($\delta$=0.008) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 48.2±0.2 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.004±0.003 | 0.000±0.000 | 84.0±0.2 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.010±0.004 | 0.000±0.000 | 141.2±0.3 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.016±0.006 | 0.000±0.000 | 225.8±0.4 | — | — | — |
| | 1024 | 0.000±0.000 | **0.004±0.003** | 0.000±0.000 | 0.032±0.008 | 0.000±0.000 | 329.8±0.4 | 968.5±22.5 | — | — |

Table 16 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P100 ($\delta$=0.002) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 61.8±0.1 | — | — | — |
| ✗ | 128 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 119.3±0.1 | — | — | — |
| | 256 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 222.9±0.2 | — | — | — |
| | 512 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 392.2±0.4 | — | — | — |
| | 1024 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 618.9±0.4 | — | — | — |

# ER-REASON.

Table 17: Per-problem certification rate on ER-REASON (72 problems) (Case A, target = mode). **Bold**: CITE $>$ Bonferroni. $\bar{K}_N$ is the Monte Carlo mean number of observed answer categories at budget $N$.

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P8 ($\delta$=1.000) | 64 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.1±0.0 | 7.0±0.0 |
| ✗ | 128 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.1±0.0 | 7.0±0.0 |
| | 256 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.1±0.0 | 7.0±0.0 |
| | 512 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.1±0.0 | 7.0±0.0 |
| | 1024 | 0.000±0.000 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.0±0.0 | 9.0±0.0 | 11.1±0.0 | 7.0±0.0 |
| P47 ($\delta$=0.996) | 64 | 0.104±0.014 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| ✗ | 128 | 0.196±0.018 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.2±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| | 256 | 0.372±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.4±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| | 512 | 0.616±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.6±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| | 1024 | 0.870±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| P54 ($\delta$=0.996) | 64 | 0.106±0.014 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.1±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| ✓ | 128 | 0.208±0.018 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.2±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| | 256 | 0.384±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.4±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| | 512 | 0.632±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.6±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| | 1024 | 0.864±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 9.1±0.0 | 11.1±0.0 | 7.1±0.0 |
| P41 ($\delta$=0.988) | 64 | 0.304±0.021 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.3±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| ✓ | 128 | 0.520±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.5±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| | 256 | 0.812±0.017 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.8±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| | 512 | 0.962±0.009 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| P2 ($\delta$=0.982) | 64 | 0.418±0.022 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.4±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| ✓ | 128 | 0.662±0.021 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.7±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| | 256 | 0.878±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| | 512 | 0.992±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 9.3±0.1 | 11.2±0.0 | 7.2±0.0 |
| P60 ($\delta$=0.951) | 64 | 0.822±0.017 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 10.1±0.1 | 12.2±0.1 | 7.9±0.1 |
| ✓ | 128 | 0.970±0.008 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 10.1±0.1 | 12.2±0.1 | 7.9±0.1 |
| | 256 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 10.1±0.1 | 12.2±0.1 | 7.9±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 10.1±0.1 | 12.2±0.1 | 7.9±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.6±0.0 | 10.1±0.1 | 12.2±0.1 | 7.9±0.1 |
| P38 ($\delta$=0.950) | 64 | 0.832±0.017 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 10.4±0.1 | 12.3±0.1 | 8.2±0.1 |
| ✓ | 128 | 0.968±0.008 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.5±0.0 | 10.4±0.1 | 12.3±0.1 | 8.2±0.1 |
| | 256 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.8±0.0 | 10.4±0.1 | 12.3±0.1 | 8.2±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 10.4±0.1 | 12.3±0.1 | 8.2±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 10.4±0.1 | 12.3±0.1 | 8.2±0.1 |
| P52 ($\delta$=0.944) | 64 | 0.852±0.016 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.5±0.1 | 12.7±0.1 | 8.1±0.1 |
| ✗ | 128 | 0.974±0.007 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 10.5±0.1 | 12.7±0.1 | 8.1±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 10.5±0.1 | 12.7±0.1 | 8.1±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.6±0.0 | 10.5±0.1 | 12.7±0.1 | 8.1±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.9±0.0 | 10.5±0.1 | 12.7±0.1 | 8.1±0.1 |
| P5 ($\delta$=0.940) | 64 | 0.874±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 10.5±0.1 | 12.4±0.1 | 8.1±0.1 |
| ✓ | 128 | 0.984±0.006 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.5±0.1 | 12.4±0.1 | 8.1±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.5±0.1 | 12.4±0.1 | 8.1±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.5±0.1 | 12.4±0.1 | 8.1±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.5±0.1 | 12.4±0.1 | 8.1±0.1 |
| P45 ($\delta$=0.940) | 64 | 0.874±0.015 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.9±0.0 | 10.4±0.1 | 12.5±0.1 | 8.1±0.1 |
| ✗ | 128 | 0.992±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.4±0.1 | 12.5±0.1 | 8.1±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.4±0.1 | 12.5±0.1 | 8.1±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.4±0.1 | 12.5±0.1 | 8.1±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 10.4±0.1 | 12.5±0.1 | 8.1±0.1 |
| P43 ($\delta$=0.934) | 64 | 0.936±0.011 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 10.7±0.1 | 11.6±0.1 | 8.5±0.1 |
| ✗ | 128 | 0.996±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.3±0.0 | 10.7±0.1 | 11.6±0.1 | 8.5±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.7±0.0 | 10.7±0.1 | 11.6±0.1 | 8.5±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 10.7±0.1 | 11.6±0.1 | 8.5±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 10.7±0.1 | 11.6±0.1 | 8.5±0.1 |
| P34 ($\delta$=0.887) | 64 | 0.970±0.008 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 11.9±0.2 | 14.0±0.1 | 9.2±0.1 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 11.9±0.2 | 14.0±0.1 | 9.2±0.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 11.9±0.2 | 14.0±0.1 | 9.2±0.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 11.9±0.2 | 14.0±0.1 | 9.2±0.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 11.9±0.2 | 14.0±0.1 | 9.2±0.1 |
| P6 ($\delta$=0.880) | 64 | 0.992±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.9±0.0 | 12.5±0.2 | 14.7±0.2 | 9.9±0.2 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.1±0.0 | 12.5±0.2 | 14.7±0.2 | 9.9±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.2±0.0 | 12.5±0.2 | 14.7±0.2 | 9.9±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.4±0.0 | 12.5±0.2 | 14.7±0.2 | 9.9±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 12.5±0.2 | 14.7±0.2 | 9.9±0.2 |
| P22 ($\delta$=0.875) | 64 | 0.990±0.004 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 12.5±0.2 | 15.1±0.1 | 10.1±0.2 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.3±0.0 | 12.5±0.2 | 15.1±0.1 | 10.1±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 12.5±0.2 | 15.1±0.1 | 10.1±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 12.5±0.2 | 15.1±0.1 | 10.1±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 12.5±0.2 | 15.1±0.1 | 10.1±0.2 |
| P10 ($\delta$=0.872) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.5±0.0 | 12.7±0.2 | 14.4±0.1 | 10.3±0.2 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 12.7±0.2 | 14.4±0.1 | 10.3±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.2±0.0 | 12.7±0.2 | 14.4±0.1 | 10.3±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 12.7±0.2 | 14.4±0.1 | 10.3±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.6±0.0 | 12.7±0.2 | 14.4±0.1 | 10.3±0.2 |
| P28 ($\delta$=0.861) | 64 | 0.996±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 13.5±0.2 | 15.8±0.1 | 11.5±0.2 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.2±0.0 | 13.5±0.2 | 15.8±0.1 | 11.5±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.6±0.0 | 13.5±0.2 | 15.8±0.1 | 11.5±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.9±0.0 | 13.5±0.2 | 15.8±0.1 | 11.5±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 13.5±0.2 | 15.8±0.1 | 11.5±0.2 |

Table 17 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| P26 ($\delta$=0.858) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.5±0.0 | 13.4±0.2 | 15.8±0.2 | 11.2±0.2 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 13.4±0.2 | 15.8±0.2 | 11.2±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.2±0.0 | 13.4±0.2 | 15.8±0.2 | 11.2±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 13.4±0.2 | 15.8±0.2 | 11.2±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.6±0.0 | 13.4±0.2 | 15.8±0.2 | 11.2±0.2 |
| P25 ($\delta$=0.856) | 64 | 0.996±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.1±0.0 | 13.3±0.2 | 15.2±0.2 | 10.9±0.2 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.4±0.0 | 13.3±0.2 | 15.2±0.2 | 10.9±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.8±0.0 | 13.3±0.2 | 15.2±0.2 | 10.9±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.2±0.0 | 13.3±0.2 | 15.2±0.2 | 10.9±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.6±0.0 | 13.3±0.2 | 15.2±0.2 | 10.9±0.2 |
| P31 ($\delta$=0.822) | 64 | 0.996±0.003 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.5±0.0 | 14.0±0.2 | 14.6±0.2 | 10.7±0.2 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 14.0±0.2 | 14.6±0.2 | 10.7±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.5±0.0 | 14.0±0.2 | 14.6±0.2 | 10.7±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 14.0±0.2 | 14.6±0.2 | 10.7±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.5±0.0 | 14.0±0.2 | 14.6±0.2 | 10.7±0.2 |
| P21 ($\delta$=0.817) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.2±0.0 | 14.6±0.2 | 16.3±0.2 | 12.1±0.2 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 14.6±0.2 | 16.3±0.2 | 12.1±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.1±0.0 | 14.6±0.2 | 16.3±0.2 | 12.1±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.7±0.0 | 14.6±0.2 | 16.3±0.2 | 12.1±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.2±0.0 | 14.6±0.2 | 16.3±0.2 | 12.1±0.2 |
| P1 ($\delta$=0.800) | 64 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 15.0±0.3 | 16.6±0.2 | 11.3±0.3 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 15.0±0.3 | 16.6±0.2 | 11.3±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 15.0±0.3 | 16.6±0.2 | 11.3±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 15.0±0.3 | 16.6±0.2 | 11.3±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 15.0±0.3 | 16.6±0.2 | 11.3±0.3 |
| P24 ($\delta$=0.790) | 64 | 0.998±0.002 | **1.000±0.000** | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 15.3±0.3 | 17.3±0.2 | 11.8±0.2 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.1±0.0 | 15.3±0.3 | 17.3±0.2 | 11.8±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.7±0.0 | 15.3±0.3 | 17.3±0.2 | 11.8±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.2±0.0 | 15.3±0.3 | 17.3±0.2 | 11.8±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.6±0.0 | 15.3±0.3 | 17.3±0.2 | 11.8±0.2 |
| P71 ($\delta$=0.781) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.6±0.0 | 15.5±0.3 | 16.0±0.3 | 11.9±0.2 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.8±0.0 | 15.5±0.3 | 16.0±0.3 | 11.9±0.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 15.5±0.3 | 16.0±0.3 | 11.9±0.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 15.5±0.3 | 16.0±0.3 | 11.9±0.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 15.5±0.3 | 16.0±0.3 | 11.9±0.2 |
| P53 ($\delta$=0.771) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 16.6±0.3 | 18.2±0.2 | 14.7±0.3 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.5±0.0 | 16.6±0.3 | 18.2±0.2 | 14.7±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.9±0.0 | 16.6±0.3 | 18.2±0.2 | 14.7±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.0±0.0 | 16.6±0.3 | 18.2±0.2 | 14.7±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.0±0.0 | 16.6±0.3 | 18.2±0.2 | 14.7±0.3 |
| P56 ($\delta$=0.757) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 17.4±0.4 | 18.0±0.3 | 13.0±0.3 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.1±0.0 | 17.4±0.4 | 18.0±0.3 | 13.0±0.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.2±0.0 | 17.4±0.4 | 18.0±0.3 | 13.0±0.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 17.4±0.4 | 18.0±0.3 | 13.0±0.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 17.4±0.4 | 18.0±0.3 | 13.0±0.3 |
| P11 ($\delta$=0.691) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 5.3±0.0 | 20.3±0.3 | 21.1±0.3 | 22.9±0.6 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.7±0.0 | 20.3±0.3 | 21.1±0.3 | 23.0±0.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.9±0.0 | 20.3±0.3 | 21.1±0.3 | 23.0±0.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 20.3±0.3 | 21.1±0.3 | 23.0±0.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 20.3±0.3 | 21.1±0.3 | 23.0±0.6 |
| P46 ($\delta$=0.679) | 64 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 21.1±0.4 | 21.2±0.3 | 18.9±0.4 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 21.1±0.4 | 21.2±0.3 | 18.9±0.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 21.1±0.4 | 21.2±0.3 | 18.9±0.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 21.1±0.4 | 21.2±0.3 | 18.9±0.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 21.1±0.4 | 21.2±0.3 | 18.9±0.4 |
| P15 ($\delta$=0.666) | 64 | 1.000±0.000 | 0.996±0.003 | 0.996±0.003 | 0.998±0.002 | 0.998±0.002 | 4.8±0.0 | 21.3±0.4 | 22.9±0.4 | 19.8±0.4 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.3±0.0 | 21.5±0.4 | 23.1±0.4 | 19.9±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.7±0.0 | 21.5±0.4 | 23.1±0.4 | 19.9±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.9±0.0 | 21.5±0.4 | 23.1±0.4 | 19.9±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 21.5±0.4 | 23.1±0.4 | 19.9±0.5 |
| P42 ($\delta$=0.655) | 64 | 1.000±0.000 | 0.998±0.002 | 0.996±0.003 | 0.998±0.002 | 0.994±0.003 | 2.7±0.0 | 22.0±0.5 | 23.3±0.4 | 16.8±0.4 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.9±0.0 | 22.1±0.5 | 23.5±0.4 | 17.1±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 22.1±0.5 | 23.5±0.4 | 17.1±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 22.1±0.5 | 23.5±0.4 | 17.1±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 22.1±0.5 | 23.5±0.4 | 17.1±0.5 |
| P68 ($\delta$=0.645) | 64 | 1.000±0.000 | 0.998±0.002 | 0.998±0.002 | 0.998±0.002 | 0.998±0.002 | 3.3±0.0 | 23.0±0.5 | 24.2±0.4 | 18.1±0.4 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.7±0.0 | 23.1±0.5 | 24.3±0.4 | 18.2±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 23.1±0.5 | 24.3±0.4 | 18.2±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 23.1±0.5 | 24.3±0.4 | 18.2±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 23.1±0.5 | 24.3±0.4 | 18.2±0.5 |
| P51 ($\delta$=0.642) | 64 | 1.000±0.000 | 0.994±0.003 | 0.998±0.002 | 0.994±0.003 | 0.998±0.002 | 2.3±0.0 | 22.8±0.5 | 24.3±0.4 | 17.0±0.5 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.4±0.0 | 23.1±0.6 | 24.4±0.4 | 17.1±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.7±0.0 | 23.1±0.6 | 24.4±0.4 | 17.1±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.2±0.0 | 23.1±0.6 | 24.4±0.4 | 17.1±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 23.1±0.6 | 24.4±0.4 | 17.1±0.5 |
| P16 ($\delta$=0.602) | 64 | 1.000±0.000 | 0.992±0.004 | 0.994±0.003 | 0.992±0.004 | 0.998±0.002 | 3.3±0.0 | 26.5±0.6 | 26.7±0.5 | 20.7±0.5 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.7±0.0 | 26.8±0.6 | 26.9±0.5 | 20.8±0.5 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.1±0.0 | 26.8±0.6 | 26.9±0.5 | 20.8±0.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 26.8±0.6 | 26.9±0.5 | 20.8±0.5 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.6±0.0 | 26.8±0.6 | 26.9±0.5 | 20.8±0.5 |
| P62 ($\delta$=0.573) | 64 | 0.998±0.002 | 0.972±0.007 | 0.984±0.006 | 0.976±0.007 | 0.980±0.006 | 4.1±0.0 | 27.4±0.5 | 28.2±0.5 | 26.1±0.6 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.2±0.0 | 28.6±0.6 | 28.9±0.5 | 26.9±0.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.4±0.0 | 28.6±0.6 | 28.9±0.5 | 26.9±0.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.7±0.0 | 28.6±0.6 | 28.9±0.5 | 26.9±0.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.9±0.0 | 28.6±0.6 | 28.9±0.5 | 26.9±0.6 |
| P32 ($\delta$=0.566) | 64 | 0.992±0.004 | 0.958±0.009 | 0.952±0.010 | 0.980±0.006 | 0.974±0.007 | 2.0±0.0 | 29.0±0.7 | 30.6±0.6 | 21.8±0.6 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 31.0±0.8 | 33.0±0.8 | 23.2±0.7 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 31.0±0.8 | 33.0±0.8 | 23.2±0.7 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 31.0±0.8 | 33.0±0.8 | 23.2±0.7 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 31.0±0.8 | 33.0±0.8 | 23.2±0.7 |
| P50 ($\delta$=0.545) | 64 | 0.994±0.003 | 0.984±0.006 | 0.982±0.006 | 0.976±0.007 | 0.974±0.007 | 3.3±0.0 | 30.4±0.6 | 31.8±0.5 | 27.2±0.6 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.5±0.0 | 31.1±0.6 | 32.5±0.6 | 28.3±0.7 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 31.1±0.6 | 32.5±0.6 | 28.3±0.7 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.5±0.0 | 31.1±0.6 | 32.5±0.6 | 28.3±0.7 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.2±0.0 | 31.1±0.6 | 32.5±0.6 | 28.3±0.7 |
| P17 ($\delta$=0.538) | 64 | 0.982±0.006 | 0.938±0.011 | 0.928±0.012 | 0.968±0.008 | 0.952±0.010 | 3.2±0.0 | 29.7±0.7 | 31.5±0.6 | 22.9±0.6 |

Table 17 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.7±0.0 | 32.7±0.8 | 34.8±0.8 | 25.5±0.8 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 32.7±0.8 | 34.8±0.8 | 25.5±0.8 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 32.7±0.8 | 34.8±0.8 | 25.5±0.8 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 32.7±0.8 | 34.8±0.8 | 25.5±0.8 |
| P14 ($\delta$=0.504) | 64 | 0.976±0.007 | 0.920±0.012 | 0.890±0.014 | 0.962±0.009 | 0.862±0.015 | 5.1±0.0 | 34.2±0.6 | 36.9±0.6 | 35.0±0.8 |
| ✗ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 6.1±0.0 | 37.7±0.8 | 41.4±0.8 | 41.4±1.0 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.9±0.0 | 37.7±0.8 | 41.4±0.8 | 41.6±1.0 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.7±0.0 | 37.7±0.8 | 41.4±0.8 | 41.6±1.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 8.4±0.0 | 37.7±0.8 | 41.4±0.8 | 41.6±1.0 |
| P20 ($\delta$=0.502) | 64 | 0.960±0.009 | 0.880±0.015 | 0.898±0.014 | 0.934±0.011 | 0.910±0.013 | 3.7±0.0 | 32.7±0.7 | 34.1±0.6 | 26.3±0.7 |
| ✗ | 128 | 1.000±0.000 | 0.998±0.002 | 1.000±0.000 | 0.998±0.002 | 1.000±0.000 | 4.7±0.0 | 38.1±0.9 | 38.7±0.9 | 31.2±1.0 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.1 | 38.3±1.0 | 38.7±0.9 | 31.2±1.0 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 7.2±0.0 | 38.3±1.0 | 38.7±0.9 | 31.2±1.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 8.1±0.0 | 38.3±1.0 | 38.7±0.9 | 31.2±1.0 |
| P3 ($\delta$=0.493) | 64 | 0.972±0.007 | 0.936±0.011 | 0.932±0.011 | 0.960±0.009 | 0.728±0.020 | 5.8±0.0 | 31.9±0.6 | 33.7±0.5 | 36.0±0.8 |
| ✓ | 128 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.982±0.006 | 6.0±0.0 | 34.9±0.8 | 36.8±0.7 | 48.6±1.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 34.9±0.8 | 36.8±0.7 | 50.3±1.3 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 34.9±0.8 | 36.8±0.7 | 50.3±1.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 34.9±0.8 | 36.8±0.7 | 50.3±1.3 |
| P40 ($\delta$=0.458) | 64 | 0.926±0.012 | 0.760±0.019 | 0.788±0.018 | 0.888±0.014 | 0.834±0.017 | 2.0±0.0 | 33.9±0.8 | 34.5±0.7 | 26.8±0.8 |
| ✗ | 128 | 0.998±0.002 | 0.986±0.005 | 0.990±0.004 | 0.986±0.005 | 0.984±0.006 | 2.0±0.0 | 44.9±1.1 | 43.8±1.0 | 35.3±1.2 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 46.5±1.3 | 45.0±1.2 | 37.3±1.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 46.5±1.3 | 45.0±1.2 | 37.3±1.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 46.5±1.3 | 45.0±1.2 | 37.3±1.4 |
| P49 ($\delta$=0.458) | 64 | 0.910±0.013 | 0.716±0.020 | 0.690±0.021 | 0.868±0.015 | 0.814±0.017 | 2.0±0.0 | 34.5±0.8 | 35.8±0.8 | 27.4±0.8 |
| ✗ | 128 | 1.000±0.000 | 0.976±0.007 | 0.972±0.007 | 0.978±0.007 | 0.978±0.007 | 2.0±0.0 | 47.7±1.2 | 50.3±1.2 | 37.4±1.3 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 49.9±1.4 | 53.0±1.4 | 39.7±1.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 49.9±1.4 | 53.0±1.4 | 39.7±1.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 49.9±1.4 | 53.0±1.4 | 39.7±1.4 |
| P59 ($\delta$=0.430) | 64 | 0.888±0.014 | 0.772±0.019 | 0.782±0.018 | 0.868±0.015 | 0.780±0.019 | 4.5±0.0 | 37.4±0.7 | 37.7±0.7 | 37.3±0.8 |
| ✗ | 128 | 1.000±0.000 | 0.986±0.005 | 0.986±0.005 | 0.994±0.003 | 0.986±0.005 | 5.0±0.0 | 47.4±1.1 | 46.8±1.0 | 46.8±1.1 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.4±0.0 | 48.7±1.2 | 48.0±1.1 | 48.2±1.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.6±0.0 | 48.7±1.2 | 48.0±1.1 | 48.2±1.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.9±0.0 | 48.7±1.2 | 48.0±1.1 | 48.2±1.2 |
| P33 ($\delta$=0.428) | 64 | 0.910±0.013 | 0.808±0.018 | 0.786±0.018 | 0.874±0.015 | 0.538±0.022 | 5.3±0.0 | 37.9±0.7 | 38.1±0.6 | 39.8±1.0 |
| ✗ | 128 | 0.998±0.002 | 0.996±0.003 | 0.992±0.004 | 0.992±0.004 | 0.942±0.010 | 5.6±0.0 | 46.0±1.0 | 47.0±1.0 | 60.9±1.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.8±0.0 | 46.4±1.0 | 47.8±1.0 | 65.8±1.6 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 46.4±1.0 | 47.8±1.0 | 65.8±1.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 6.0±0.0 | 46.4±1.0 | 47.8±1.0 | 65.8±1.6 |
| P29 ($\delta$=0.392) | 64 | 0.760±0.019 | 0.520±0.022 | 0.538±0.022 | 0.776±0.019 | 0.634±0.022 | 2.9±0.0 | 36.7±1.0 | 37.9±0.9 | 30.3±1.0 |
| ✓ | 128 | 0.986±0.005 | 0.926±0.012 | 0.936±0.011 | 0.948±0.010 | 0.940±0.011 | 3.4±0.0 | 60.6±1.5 | 59.7±1.4 | 50.2±1.6 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.8±0.0 | 67.4±1.8 | 65.7±1.7 | 56.7±1.9 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 67.4±1.8 | 65.7±1.7 | 56.7±1.9 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 67.4±1.8 | 65.7±1.7 | 56.7±1.9 |
| P37 ($\delta$=0.380) | 64 | 0.746±0.019 | 0.482±0.022 | 0.486±0.022 | 0.746±0.019 | 0.604±0.022 | 2.0±0.0 | 36.5±1.0 | 38.1±0.9 | 29.3±1.0 |
| ✗ | 128 | 0.972±0.007 | 0.894±0.014 | 0.902±0.013 | 0.928±0.012 | 0.914±0.013 | 2.0±0.0 | 62.0±1.5 | 62.4±1.4 | 51.0±1.6 |
| | 256 | 1.000±0.000 | 0.998±0.002 | 0.998±0.002 | 1.000±0.000 | 0.998±0.002 | 2.0±0.0 | 72.0±1.9 | 71.7±1.9 | 60.8±2.1 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 72.4±2.0 | 72.1±1.9 | 61.2±2.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 2.0±0.0 | 72.4±2.0 | 72.1±1.9 | 61.2±2.2 |
| P12 ($\delta$=0.364) | 64 | 0.748±0.019 | 0.508±0.022 | 0.554±0.022 | 0.742±0.020 | 0.586±0.022 | 3.0±0.0 | 40.7±0.9 | 40.6±0.8 | 38.3±0.9 |
| ✗ | 128 | 0.984±0.006 | 0.914±0.013 | 0.950±0.010 | 0.940±0.011 | 0.938±0.011 | 3.1±0.0 | 62.9±1.4 | 60.5±1.3 | 57.9±1.4 |
| | 256 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.994±0.003 | 1.000±0.000 | 3.2±0.0 | 70.8±1.7 | 65.5±1.6 | 64.5±1.8 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.4±0.0 | 70.8±1.7 | 65.5±1.6 | 64.5±1.8 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.6±0.0 | 70.8±1.7 | 65.5±1.6 | 64.5±1.8 |
| P30 ($\delta$=0.362) | 64 | 0.702±0.020 | 0.464±0.022 | 0.492±0.022 | 0.680±0.021 | 0.580±0.022 | 2.3±0.0 | 39.9±1.1 | 39.6±1.0 | 31.4±1.0 |
| ✗ | 128 | 0.960±0.009 | 0.854±0.016 | 0.884±0.014 | 0.920±0.012 | 0.886±0.014 | 2.5±0.0 | 64.1±1.5 | 63.0±1.5 | 53.2±1.7 |
| | 256 | 1.000±0.000 | 0.998±0.002 | 0.998±0.002 | 0.994±0.003 | 0.998±0.002 | 2.8±0.0 | 78.6±2.1 | 74.1±1.9 | 65.9±2.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 78.9±2.1 | 74.5±2.0 | 66.3±2.3 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 78.9±2.1 | 74.5±2.0 | 66.3±2.3 |
| P9 ($\delta$=0.331) | 64 | 0.588±0.022 | 0.418±0.022 | 0.462±0.022 | 0.680±0.021 | 0.542±0.022 | 3.2±0.0 | 40.9±1.1 | 41.3±0.9 | 32.7±1.0 |
| ✓ | 128 | 0.932±0.011 | 0.780±0.019 | 0.832±0.017 | 0.876±0.015 | 0.824±0.017 | 3.5±0.0 | 65.9±1.6 | 65.0±1.5 | 53.8±1.7 |
| | 256 | 1.000±0.000 | 0.988±0.005 | 0.994±0.003 | 0.992±0.004 | 0.988±0.005 | 3.7±0.0 | 85.8±2.2 | 80.2±2.0 | 72.1±2.4 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 88.2±2.4 | 81.5±2.2 | 74.7±2.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 4.0±0.0 | 88.2±2.4 | 81.5±2.2 | 74.7±2.6 |
| P65 ($\delta$=0.330) | 64 | 0.576±0.022 | 0.424±0.022 | 0.458±0.022 | 0.688±0.021 | 0.536±0.022 | 3.9±0.0 | 38.9±1.1 | 39.3±0.9 | 32.2±1.0 |
| ✗ | 128 | 0.904±0.013 | 0.782±0.018 | 0.822±0.017 | 0.876±0.015 | 0.802±0.018 | 4.5±0.0 | 64.2±1.6 | 62.5±1.5 | 52.9±1.7 |
| | 256 | 0.998±0.002 | 0.984±0.006 | 0.998±0.002 | 0.990±0.004 | 0.986±0.005 | 4.9±0.0 | 84.8±2.3 | 80.4±2.2 | 74.6±2.5 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.3±0.0 | 88.1±2.6 | 80.9±2.3 | 77.7±2.8 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.7±0.0 | 88.1±2.6 | 80.9±2.3 | 77.7±2.8 |
| P72 ($\delta$=0.288) | 64 | 0.462±0.022 | 0.238±0.019 | 0.230±0.019 | 0.530±0.022 | 0.350±0.021 | 2.3±0.0 | 36.1±1.5 | 37.0±1.4 | 28.1±1.3 |
| ✗ | 128 | 0.796±0.018 | 0.570±0.022 | 0.568±0.022 | 0.756±0.019 | 0.614±0.022 | 2.7±0.0 | 71.8±2.1 | 72.1±2.0 | 57.1±2.2 |
| | 256 | 0.984±0.006 | 0.942±0.010 | 0.944±0.010 | 0.960±0.009 | 0.926±0.012 | 3.1±0.0 | 113.1±2.8 | 113.5±2.8 | 97.8±3.2 |
| | 512 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.994±0.003 | 1.000±0.000 | 3.6±0.0 | 125.3±3.5 | 125.3±3.5 | 115.0±4.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.9±0.0 | 125.3±3.5 | 125.3±3.5 | 115.0±4.1 |
| P63 ($\delta$=0.280) | 64 | 0.442±0.022 | 0.250±0.019 | 0.280±0.020 | 0.516±0.022 | 0.346±0.021 | 3.0±0.0 | 41.8±1.3 | 42.6±1.1 | 36.6±1.2 |
| ✗ | 128 | 0.832±0.017 | 0.620±0.022 | 0.668±0.021 | 0.796±0.018 | 0.668±0.021 | 3.0±0.0 | 74.4±1.8 | 73.9±1.7 | 64.7±1.8 |
| | 256 | 0.994±0.003 | 0.958±0.009 | 0.966±0.008 | 0.954±0.009 | 0.956±0.009 | 3.0±0.0 | 109.2±2.6 | 104.2±2.5 | 98.0±2.8 |
| | 512 | 1.000±0.000 | 0.998±0.002 | 0.998±0.002 | 0.998±0.002 | 0.998±0.002 | 3.0±0.0 | 117.2±3.1 | 110.8±3.0 | 107.2±3.4 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 118.2±3.2 | 111.8±3.1 | 108.2±3.5 |
| P55 ($\delta$=0.233) | 64 | 0.276±0.020 | 0.158±0.016 | 0.158±0.016 | 0.432±0.022 | 0.274±0.020 | 3.6±0.0 | 39.8±1.6 | 39.9±1.5 | 37.6±1.4 |
| ✗ | 128 | 0.600±0.022 | 0.414±0.022 | 0.412±0.022 | 0.664±0.021 | 0.514±0.022 | 4.1±0.0 | 77.4±2.3 | 76.8±2.3 | 64.4±2.1 |
| | 256 | 0.938±0.011 | 0.828±0.017 | 0.826±0.017 | 0.874±0.015 | 0.846±0.016 | 4.5±0.0 | 132.3±3.2 | 132.9±3.2 | 114.3±3.5 |
| | 512 | 0.998±0.002 | 0.992±0.004 | 0.992±0.004 | 0.994±0.003 | 0.988±0.005 | 4.8±0.0 | 163.6±4.2 | 164.8±4.3 | 146.4±4.7 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.0±0.0 | 167.2±4.6 | 168.4±4.6 | 151.5±5.1 |
| P19 ($\delta$=0.228) | 64 | 0.250±0.019 | 0.124±0.015 | 0.130±0.015 | 0.408±0.022 | 0.226±0.019 | 2.9±0.0 | 35.9±2.1 | 38.8±1.7 | 28.1±1.6 |
| ✗ | 128 | 0.506±0.022 | 0.302±0.021 | 0.322±0.021 | 0.616±0.022 | 0.386±0.022 | 3.6±0.0 | 72.9±2.9 | 73.0±2.6 | 56.8±2.8 |
| | 256 | 0.854±0.016 | 0.702±0.020 | 0.728±0.020 | 0.828±0.017 | 0.710±0.020 | 4.6±0.0 | 138.9±3.6 | 139.0±3.6 | 117.6±4.1 |
| | 512 | 0.988±0.005 | 0.968±0.008 | 0.980±0.006 | 0.982±0.006 | 0.960±0.009 | 5.5±0.0 | 192.7±5.0 | 189.2±4.9 | 175.2±5.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 5.9±0.0 | 205.6±5.8 | 197.6±5.5 | 192.3±6.5 |
| P23 ($\delta$=0.224) | 64 | 0.290±0.020 | 0.156±0.016 | 0.152±0.016 | 0.382±0.022 | 0.234±0.019 | 3.0±0.0 | 44.9±1.8 | 45.3±1.6 | 42.1±1.5 |
| ✗ | 128 | 0.620±0.022 | 0.370±0.022 | 0.370±0.022 | 0.634±0.022 | 0.428±0.022 | 3.0±0.0 | 75.0±2.3 | 75.3±2.2 | 65.3±2.1 |
| | 256 | 0.944±0.010 | 0.802±0.018 | 0.798±0.018 | 0.886±0.014 | 0.770±0.019 | 3.0±0.0 | 135.6±3.3 | 135.6±3.3 | 119.2±3.5 |
| | 512 | 0.996±0.003 | 0.988±0.005 | 0.988±0.005 | 0.996±0.003 | 0.984±0.006 | 3.0±0.0 | 172.0±4.5 | 171.9±4.4 | 164.4±4.9 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 3.0±0.0 | 177.4±4.9 | 177.3±4.9 | 172.3±5.6 |
| P36 ($\delta$=0.224) | 64 | 0.236±0.019 | 0.142±0.016 | 0.136±0.015 | 0.428±0.022 | 0.202±0.018 | 4.3±0.0 | 38.1±1.8 | 39.2±1.6 | 32.5±1.7 |
| ✗ | 128 | 0.460±0.022 | 0.334±0.021 | 0.334±0.021 | 0.612±0.022 | 0.382±0.022 | 5.4±0.0 | 72.0±2.6 | 73.1±2.5 | 62.5±2.6 |
| | 256 | 0.828±0.017 | 0.720±0.020 | 0.720±0.020 | 0.812±0.017 | 0.702±0.020 | 6.4±0.0 | 136.9±3.7 | 137.8±3.6 | 121.1±3.9 |

Table 17 – continued

| Problem | $N$ | Bonf. | CITE | W-CITE | KR | MMC | $\bar{K}_N$ | $\bar{\tau}_{\mathrm{CITE}}$ | $\bar{\tau}_{\mathrm{W\text{-}CITE}}$ | $\bar{\tau}_{\mathrm{MMC}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 512 | 0.984±0.006 | 0.970±0.008 | 0.970±0.008 | 0.980±0.006 | 0.960±0.009 | 7.1±0.0 | 189.4±5.1 | 189.3±5.0 | 181.5±5.6 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 1.000±0.000 | 7.6±0.0 | 202.9±6.1 | 203.2±6.1 | 200.3±6.9 |
| P64 ($\delta$=0.215) | 64 | 0.242±0.019 | 0.098±0.013 | 0.102±0.014 | 0.358±0.021 | 0.198±0.018 | 2.1±0.0 | 35.4±2.4 | 38.5±2.1 | 29.3±1.8 |
| ✓ | 128 | 0.508±0.022 | 0.276±0.020 | 0.292±0.020 | 0.552±0.022 | 0.360±0.021 | 2.1±0.0 | 77.6±3.0 | 77.2±2.7 | 61.1±2.9 |
| | 256 | 0.878±0.015 | 0.638±0.021 | 0.676±0.021 | 0.782±0.018 | 0.624±0.022 | 2.2±0.0 | 141.7±3.8 | 140.7±3.6 | 116.6±4.3 |
| | 512 | 0.992±0.004 | 0.946±0.010 | 0.954±0.009 | 0.968±0.008 | 0.932±0.011 | 2.4±0.0 | 206.8±5.3 | 198.2±5.1 | 193.1±6.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.998±0.002 | 0.998±0.002 | 2.6±0.0 | 230.4±6.8 | 217.7±6.3 | 223.0±7.7 |
| P39 ($\delta$=0.207) | 64 | 0.186±0.017 | 0.098±0.013 | 0.096±0.013 | 0.340±0.021 | 0.024±0.007 | 6.2±0.0 | 46.1±1.8 | 47.8±1.7 | 30.6±4.9 |
| ✗ | 128 | 0.538±0.022 | 0.366±0.022 | 0.332±0.021 | 0.638±0.021 | 0.056±0.010 | 6.5±0.0 | 84.8±2.2 | 83.1±2.2 | 72.4±7.6 |
| | 256 | 0.890±0.014 | 0.818±0.017 | 0.780±0.019 | 0.876±0.015 | 0.130±0.015 | 6.9±0.0 | 138.6±2.9 | 141.2±3.0 | 146.6±9.3 |
| | 512 | 0.996±0.003 | 0.996±0.003 | 0.986±0.005 | 0.992±0.004 | 0.384±0.022 | 7.3±0.0 | 172.7±4.2 | 181.8±4.4 | 313.9±10.2 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.790±0.018 | 7.7±0.0 | 174.8±4.5 | 187.6±4.9 | 523.6±12.5 |
| P58 ($\delta$=0.205) | 64 | 0.208±0.018 | 0.110±0.014 | 0.116±0.014 | 0.334±0.021 | 0.024±0.007 | 5.0±0.0 | 47.0±1.9 | 47.9±1.5 | 32.7±6.1 |
| ✗ | 128 | 0.506±0.022 | 0.340±0.021 | 0.346±0.021 | 0.628±0.022 | 0.088±0.013 | 5.4±0.0 | 81.7±2.3 | 81.4±2.2 | 77.5±5.1 |
| | 256 | 0.910±0.013 | 0.776±0.019 | 0.758±0.019 | 0.878±0.015 | 0.342±0.021 | 5.7±0.0 | 142.0±3.2 | 140.0±3.2 | 168.6±5.0 |
| | 512 | 1.000±0.000 | 0.996±0.003 | 0.994±0.003 | 0.996±0.003 | 0.814±0.017 | 6.0±0.0 | 182.5±4.4 | 184.3±4.5 | 288.0±6.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.996±0.003 | 6.0±0.0 | 184.1±4.5 | 186.6±4.7 | 354.0±8.2 |
| P69 ($\delta$=0.188) | 64 | 0.132±0.015 | 0.060±0.011 | 0.152±0.016 | 0.280±0.020 | 0.010±0.004 | 5.4±0.0 | 43.3±2.5 | 45.9±1.4 | 23.4±9.0 |
| ✗ | 128 | 0.414±0.022 | 0.240±0.019 | 0.494±0.022 | 0.502±0.022 | 0.014±0.005 | 5.7±0.0 | 85.0±2.7 | 81.1±1.8 | 41.7±14.1 |
| | 256 | 0.862±0.015 | 0.712±0.020 | 0.912±0.013 | 0.830±0.017 | 0.018±0.006 | 5.9±0.0 | 153.1±3.2 | 127.6±2.8 | 71.2±22.6 |
| | 512 | 0.992±0.004 | 0.984±0.006 | 0.998±0.002 | 0.982±0.006 | 0.024±0.007 | 6.0±0.0 | 200.2±4.3 | 143.3±3.5 | 142.7±42.1 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.062±0.011 | 6.0±0.0 | 207.0±4.9 | 144.2±3.6 | 516.9±60.6 |
| P27 ($\delta$=0.179) | 64 | 0.074±0.012 | 0.030±0.008 | 0.052±0.010 | 0.190±0.018 | 0.004±0.003 | 6.7±0.0 | 44.4±4.3 | 50.9±2.1 | 7.0±0.0 |
| ✗ | 128 | 0.366±0.022 | 0.216±0.018 | 0.304±0.021 | 0.476±0.022 | 0.004±0.003 | 6.9±0.0 | 91.7±2.5 | 90.2±2.0 | 7.0±0.0 |
| | 256 | 0.836±0.017 | 0.720±0.020 | 0.834±0.017 | 0.814±0.017 | 0.004±0.003 | 7.0±0.0 | 158.8±2.9 | 149.0±2.7 | 7.0±0.0 |
| | 512 | 1.000±0.000 | 0.996±0.003 | 1.000±0.000 | 0.990±0.004 | 0.004±0.003 | 7.0±0.0 | 207.7±4.3 | 177.4±3.7 | 7.0±0.0 |
| | 1024 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 1.000±0.000 | 0.004±0.003 | 7.0±0.0 | 209.0±4.4 | 177.4±3.7 | 7.0±0.0 |
| P7 ($\delta$=0.142) | 64 | 0.088±0.013 | 0.040±0.009 | 0.026±0.007 | 0.160±0.016 | 0.042±0.009 | 3.1±0.0 | 37.6±3.3 | 38.0±3.7 | 29.7±3.0 |
| ✗ | 128 | 0.238±0.019 | 0.100±0.013 | 0.072±0.012 | 0.340±0.021 | 0.136±0.015 | 3.1±0.0 | 76.7±5.1 | 77.8±5.8 | 75.9±4.2 |
| | 256 | 0.534±0.022 | 0.324±0.021 | 0.212±0.018 | 0.578±0.022 | 0.336±0.021 | 3.2±0.0 | 162.4±5.3 | 159.2±6.6 | 145.0±5.2 |
| | 512 | 0.864±0.015 | 0.684±0.021 | 0.560±0.022 | 0.814±0.017 | 0.684±0.021 | 3.4±0.0 | 272.6±6.8 | 291.5±7.4 | 262.2±7.3 |
| | 1024 | 0.994±0.003 | 0.958±0.009 | 0.888±0.014 | 0.964±0.008 | 0.946±0.010 | 3.6±0.0 | 397.8±10.8 | 451.0±11.7 | 390.5±11.4 |
| P61 ($\delta$=0.128) | 64 | 0.074±0.012 | 0.040±0.009 | 0.032±0.008 | 0.168±0.017 | 0.098±0.013 | 2.0±0.0 | 28.9±4.0 | 38.5±3.8 | 21.4±2.1 |
| ✓ | 128 | 0.160±0.016 | 0.082±0.012 | 0.074±0.012 | 0.286±0.020 | 0.140±0.016 | 2.0±0.0 | 65.9±6.4 | 75.3±6.1 | 44.8±4.7 |
| | 256 | 0.384±0.022 | 0.174±0.017 | 0.166±0.017 | 0.390±0.022 | 0.204±0.018 | 2.0±0.0 | 137.6±8.4 | 146.2±8.1 | 91.3±7.8 |
| | 512 | 0.758±0.019 | 0.388±0.022 | 0.388±0.022 | 0.610±0.022 | 0.372±0.022 | 2.0±0.0 | 277.7±10.6 | 285.6±10.1 | 226.4±12.2 |
| | 1024 | 0.974±0.007 | 0.808±0.018 | 0.812±0.017 | 0.890±0.014 | 0.766±0.019 | 2.0±0.0 | 522.4±13.8 | 524.0±13.4 | 502.3±15.9 |
| P48 ($\delta$=0.119) | 64 | 0.052±0.010 | 0.016±0.006 | 0.034±0.008 | 0.176±0.017 | 0.070±0.011 | 2.4±0.0 | 38.0±6.0 | 42.8±2.9 | 26.7±2.8 |
| ✗ | 128 | 0.118±0.014 | 0.042±0.009 | 0.078±0.012 | 0.242±0.019 | 0.100±0.013 | 2.6±0.0 | 76.0±7.8 | 76.2±5.4 | 50.4±5.7 |
| | 256 | 0.256±0.020 | 0.106±0.014 | 0.198±0.018 | 0.382±0.022 | 0.148±0.016 | 3.0±0.0 | 145.9±9.4 | 150.5±6.9 | 99.6±9.5 |
| | 512 | 0.574±0.022 | 0.304±0.021 | 0.474±0.022 | 0.580±0.022 | 0.312±0.021 | 3.4±0.0 | 296.3±10.7 | 285.0±8.8 | 250.4±13.1 |
| | 1024 | 0.898±0.014 | 0.716±0.020 | 0.876±0.015 | 0.840±0.016 | 0.658±0.021 | 3.6±0.0 | 557.8±14.0 | 484.7±12.3 | 512.4±16.1 |
| P18 ($\delta$=0.114) | 64 | 0.042±0.009 | 0.036±0.008 | 0.038±0.009 | 0.148±0.016 | 0.074±0.012 | 4.2±0.0 | 31.1±5.1 | 33.7±4.0 | 25.3±2.9 |
| ✗ | 128 | 0.104±0.014 | 0.066±0.011 | 0.104±0.014 | 0.236±0.019 | 0.106±0.014 | 4.8±0.0 | 61.2±6.8 | 75.0±5.2 | 45.3±5.0 |
| | 256 | 0.252±0.019 | 0.162±0.016 | 0.230±0.019 | 0.390±0.022 | 0.206±0.018 | 5.0±0.0 | 138.0±8.1 | 140.0±6.5 | 119.4±8.3 |
| | 512 | 0.504±0.022 | 0.338±0.021 | 0.500±0.022 | 0.588±0.022 | 0.322±0.021 | 5.0±0.0 | 263.5±10.9 | 273.4±9.0 | 212.7±11.6 |
| | 1024 | 0.890±0.014 | 0.706±0.020 | 0.890±0.014 | 0.866±0.015 | 0.648±0.021 | 5.0±0.0 | 512.8±15.0 | 473.9±12.6 | 480.5±16.9 |
| P13 ($\delta$=0.114) | 64 | 0.026±0.007 | 0.002±0.002 | 0.002±0.002 | 0.064±0.011 | 0.002±0.002 | 3.0±0.0 | 49.0±0.0 | 48.0±0.0 | 47.0±0.0 |
| ✗ | 128 | 0.060±0.011 | 0.014±0.005 | 0.020±0.006 | 0.104±0.014 | 0.016±0.006 | 3.0±0.0 | 95.7±10.5 | 95.1±7.9 | 89.8±7.8 |
| | 256 | 0.232±0.019 | 0.082±0.012 | 0.106±0.014 | 0.280±0.020 | 0.084±0.012 | 3.0±0.0 | 189.1±8.3 | 182.0±7.1 | 176.7±7.9 |
| | 512 | 0.644±0.021 | 0.298±0.020 | 0.410±0.022 | 0.560±0.022 | 0.262±0.020 | 3.0±0.0 | 344.3±9.5 | 343.9±8.2 | 331.8±11.0 |
| | 1024 | 0.952±0.010 | 0.782±0.018 | 0.868±0.015 | 0.866±0.015 | 0.716±0.020 | 3.0±0.0 | 584.8±11.7 | 541.6±11.0 | 588.6±12.7 |
| P57 ($\delta$=0.109) | 64 | 0.022±0.007 | 0.012±0.005 | 0.018±0.006 | 0.122±0.015 | 0.002±0.002 | 7.5±0.0 | 39.5±4.6 | 38.9±3.5 | 64.0±0.0 |
| ✗ | 128 | 0.106±0.014 | 0.064±0.011 | 0.062±0.011 | 0.240±0.019 | 0.002±0.002 | 7.9±0.0 | 87.4±4.9 | 83.0±5.8 | 64.0±0.0 |
| | 256 | 0.270±0.020 | 0.194±0.018 | 0.150±0.016 | 0.448±0.022 | 0.002±0.002 | 8.0±0.0 | 161.5±6.4 | 148.7±7.7 | 64.0±0.0 |
| | 512 | 0.602±0.022 | 0.480±0.022 | 0.388±0.022 | 0.702±0.020 | 0.002±0.002 | 8.0±0.0 | 290.1±8.1 | 295.4±9.8 | 64.0±0.0 |
| | 1024 | 0.950±0.010 | 0.884±0.014 | 0.758±0.019 | 0.922±0.012 | 0.002±0.002 | 8.0±0.0 | 496.1±12.5 | 517.0±13.8 | 64.0±0.0 |
| P44 ($\delta$=0.073) | 64 | 0.002±0.002 | 0.000±0.000 | 0.000±0.000 | 0.028±0.007 | 0.002±0.002 | 4.4±0.0 | — | — | 7.0±0.0 |
| ✗ | 128 | 0.014±0.005 | 0.002±0.002 | 0.000±0.000 | 0.044±0.009 | 0.006±0.003 | 4.8±0.0 | 127.0±0.0 | — | 75.3±35.5 |
| | 256 | 0.036±0.008 | 0.010±0.004 | 0.000±0.000 | 0.106±0.014 | 0.006±0.003 | 5.3±0.0 | 178.6±19.8 | — | 75.3±35.5 |
| | 512 | 0.146±0.016 | 0.056±0.010 | 0.008±0.004 | 0.234±0.019 | 0.012±0.005 | 5.9±0.0 | 364.4±20.0 | 385.2±54.8 | 225.7±71.7 |
| | 1024 | 0.454±0.022 | 0.246±0.019 | 0.034±0.008 | 0.512±0.022 | 0.028±0.007 | 6.5±0.0 | 701.6±20.3 | 694.4±56.4 | 529.3±85.1 |
| P4 ($\delta$=0.071) | 64 | 0.000±0.000 | 0.000±0.000 | 0.000±0.000 | 0.040±0.009 | 0.000±0.000 | 7.1±0.0 | — | — | — |
| ✗ | 128 | 0.004±0.003 | 0.000±0.000 | 0.002±0.002 | 0.054±0.010 | 0.000±0.000 | 7.5±0.0 | — | 89.0±0.0 | — |
| | 256 | 0.044±0.009 | 0.016±0.006 | 0.026±0.007 | 0.146±0.016 | 0.000±0.000 | 7.8±0.0 | 197.2±13.8 | 186.0±13.9 | — |
| | 512 | 0.210±0.018 | 0.110±0.014 | 0.210±0.018 | 0.358±0.021 | 0.000±0.000 | 8.0±0.0 | 381.2±13.6 | 367.8±9.2 | — |
| | 1024 | 0.608±0.022 | 0.416±0.022 | 0.678±0.021 | 0.680±0.021 | 0.000±0.000 | 8.0±0.0 | 678.7±15.5 | 634.8±12.1 | — |
| P66 ($\delta$=0.070) | 64 | 0.034±0.008 | 0.016±0.006 | 0.014±0.005 | 0.120±0.015 | 0.052±0.010 | 2.0±0.0 | 23.4±6.2 | 31.1±8.6 | 25.4±3.8 |
| ✗ | 128 | 0.072±0.012 | 0.026±0.007 | 0.028±0.007 | 0.120±0.015 | 0.066±0.011 | 2.0±0.0 | 52.0±11.4 | 62.9±10.0 | 41.1±6.3 |
| | 256 | 0.088±0.013 | 0.042±0.009 | 0.044±0.009 | 0.186±0.017 | 0.078±0.012 | 2.0±0.0 | 104.8±17.6 | 103.8±14.3 | 66.0±11.1 |
| | 512 | 0.222±0.019 | 0.084±0.012 | 0.118±0.014 | 0.262±0.020 | 0.114±0.014 | 2.0±0.0 | 254.7±26.0 | 288.1±20.8 | 171.1±22.7 |
| | 1024 | 0.454±0.022 | 0.174±0.017 | 0.232±0.019 | 0.442±0.022 | 0.172±0.017 | 2.0±0.0 | 546.2±34.7 | 549.0±28.6 | 383.8±36.8 |
| P35 ($\delta$=0.068) | 64 | 0.028±0.007 | 0.018±0.006 | 0.014±0.005 | 0.114±0.014 | 0.042±0.009 | 3.1±0.0 | 27.2±4.5 | 30.9±5.3 | 21.8±2.7 |
| ✗ | 128 | 0.054±0.010 | 0.028±0.007 | 0.022±0.007 | 0.146±0.016 | 0.052±0.010 | 3.6±0.0 | 51.9±9.8 | 57.3±11.9 | 35.5±6.2 |
| | 256 | 0.068±0.011 | 0.048±0.010 | 0.028±0.007 | 0.188±0.017 | 0.064±0.011 | 4.1±0.0 | 108.8±15.5 | 86.5±18.1 | 63.1±11.8 |
| | 512 | 0.162±0.016 | 0.082±0.012 | 0.042±0.009 | 0.286±0.020 | 0.100±0.013 | 4.4±0.0 | 227.3±25.3 | 178.4±33.2 | 176.8±24.0 |
| | 1024 | 0.334±0.021 | 0.180±0.017 | 0.072±0.012 | 0.444±0.022 | 0.176±0.017 | 4.7±0.0 | 538.3±34.7 | 439.2±57.1 | 434.2±36.2 |
| P67 ($\delta$=0.026) | 64 | 0.010±0.004 | 0.004±0.003 | 0.004±0.003 | 0.050±0.010 | 0.026±0.007 | 3.4±0.0 | 34.0±1.0 | 34.0±1.0 | 31.2±5.1 |
| ✗ | 128 | 0.008±0.004 | 0.006±0.003 | 0.006±0.003 | 0.072±0.012 | 0.028±0.007 | 3.7±0.0 | 54.3±20.3 | 54.0±20.0 | 35.4±6.3 |
| | 256 | 0.020±0.006 | 0.010±0.004 | 0.010±0.004 | 0.086±0.013 | 0.030±0.008 | 3.9±0.0 | 93.8±26.6 | 92.2±25.8 | 43.9±10.3 |
| | 512 | 0.022±0.007 | 0.012±0.005 | 0.016±0.006 | 0.108±0.014 | 0.032±0.008 | 4.0±0.0 | 132.2±44.1 | 172.9±42.8 | 62.6±21.1 |
| | 1024 | 0.046±0.009 | 0.012±0.005 | 0.018±0.006 | 0.154±0.016 | 0.034±0.008 | 4.0±0.0 | 132.2±44.1 | 219.3±59.9 | 91.1±34.7 |
| P70 ($\delta$=0.004) | 64 | 0.008±0.004 | 0.004±0.003 | 0.006±0.003 | 0.032±0.008 | 0.008±0.004 | 3.6±0.0 | 47.0±7.0 | 51.0±6.4 | 41.5±4.4 |
| ✗ | 128 | 0.006±0.003 | 0.006±0.003 | 0.008±0.004 | 0.044±0.009 | 0.014±0.005 | 3.9±0.0 | 65.0±18.4 | 63.0±12.8 | 56.6±7.7 |
| | 256 | 0.002±0.002 | **0.006±0.003** | 0.008±0.004 | 0.050±0.010 | 0.014±0.005 | 4.4±0.0 | 65.0±18.4 | 63.0±12.8 | 56.6±7.7 |
| | 512 | 0.006±0.003 | 0.006±0.003 | 0.012±0.005 | 0.054±0.010 | 0.014±0.005 | 4.8±0.0 | 65.0±18.4 | 188.3±80.0 | 56.6±7.7 |
| | 1024 | 0.012±0.005 | 0.008±0.004 | 0.018±0.006 | 0.058±0.010 | 0.016±0.006 | 5.0±0.0 | 210.8±146.3 | 361.2±110.3 | 130.5±74.2 |

# I Reproducibility Statement

**Hardware and inference server.** Real-LLM experiments are conducted using NVIDIA A100 80 GB GPUs with vLLM as the inference server.

**Data and code access.** All code will be open-sourced upon publication. Datasets and models are publicly available (Table 18). ER-REASON requires PhysioNet credentialed access.

| Name | Reference |
|---|---|
| *Models* | |
| Qwen3-30B-A3B-Instruct-2507 [Qwen Team, 2025] | `https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507` |
| gpt-oss-20b [OpenAI, 2025] | `https://huggingface.co/openai/gpt-oss-20b` |
| *Datasets* | |
| AIME 2026 [Dekoninck et al., 2026] | `https://huggingface.co/datasets/MathArena/aime_2026` |
| FrontierScience-Olympiad [Wang et al., 2026] | `https://huggingface.co/datasets/openai/frontierscience` |
| ER-REASON [Mehandru et al., 2025] | `https://physionet.org/content/er-reason/1.0.0/` |

Table 18: List of models and datasets used in the LLM self-consistency experiments.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly present the result and insight of this paper.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The conclusion section clearly states the limitation.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This paper rigorously provide the assumption and full proofs of the theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The details are described in the main body and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We submit all of our codes as a supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are shown in the experiments section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included the error bars in our experimental plot, while it is too small and hard to see it because our result is very stable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experimental resources are shown in the experiments section and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: These results are fundamental in nature and focus on theory; they do not have a direct impact on society.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our models and met

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited every papers and works used in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We mention our new code and attach them as supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not use crowdsourcing and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We use LLMs for a standard way, and do not use LLMs for non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.