HYPOTHESIS TESTING WITH CONDITIONAL QUERIES: LEARNABILITY AND THE VALUE OF INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model evaluations may fix all tests before observing any responses or select later tests using earlier responses. We study this choice in a conditional-query model on a finite outcome space $\mathcal{X}$ with $|\mathcal{X}| = N$. We first ask which pairs of distribution classes can be reliably distinguished. We then ask how many additional queries are required to match an adaptive tester when all queried events must be fixed in advance. We show that learnability holds if and only if the two classes have positive separation in their pairwise conditional probabilities. When this separation is zero, the optimal worst-case error is exactly $1/2$ at every finite query budget. For any T-query adaptive policy and any $\rho \in (0, 1)$, we construct a randomized non-adaptive procedure using $O\big(N^2(T + \log(1/\rho))\big)$ pair queries chosen before any response is observed. Its simulated transcript is within ρ in total variation of the adaptive transcript, uniformly over all distributions in the model. We also construct a matching family with constant adaptive query complexity and $\Omega_\varepsilon(N^2)$ non-adaptive query complexity. Consequently, the worst-case fixed-error adaptivity gap is $\Theta_\varepsilon(N^2)$. Thus interaction can reduce the required number of tests by a quadratic factor, but the apparent exponential branching of an interactive evaluation does not yield an exponential query advantage.

1 INTRODUCTION

Model evaluations take both static and interactive forms. Widely used benchmark suites typically fix their items before a model is evaluated (Hendrycks et al., 2021; Srivastava et al., 2023; Liang et al., 2023). Dynamic benchmark projects instead create or refresh data using model behavior over successive development cycles (Kiela et al., 2021; Nie et al., 2020; Potts et al., 2021). Behavioral testing broadens what benchmark items probe, while interactive platforms evaluate open-ended or multi-turn behavior (Ribeiro et al., 2020; Thrush et al., 2022; Zheng et al., 2023). Live benchmarks update questions over time to limit contamination (White et al., 2025). These approaches change evaluation design at different timescales. We focus on adaptivity within a single evaluation. A static benchmark fixes all tests before observing any response. An interactive evaluation lets earlier responses determine which tests are asked later. A static design avoids sequential rounds, but it may need to prepare tests for many possible response histories. An interactive evaluation can wait to see which history occurs before selecting the next test. Interaction may therefore reduce the number of tests at the cost of sequential evaluation. How large can this reduction be? In particular, can a fixed test suite match any interactive evaluation with a polynomial, rather than exponential, increase in size?

Conditional distribution testing studies algorithms that may sample from a distribution after conditioning on a chosen event (Chakraborty et al., 2013; Canonne et al., 2015; 2014). Existing results establish problem-specific upper and lower bounds for identity, equivalence, uniformity, support-size, and tolerant testing (Acharya et al., 2018; Falahatgar et al., 2015; Narayanan, 2021). The area is surveyed by Canonne (2020). Most closely related to our efficiency question, Kamath & Tzamos (2019) develop non-adaptive conditional-sampling algorithms, and Chakraborty et al. (2024) study bounded adaptivity for equivalence testing.

The timing of experimental choices also appears in active hypothesis testing. This line begins with sequential experiment design and studies how actions chosen from past observations affect testing performance (Chernoff, 1959; Naghshvar & Javidi, 2013b; Nitinawarat et al., 2013). In particular,

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

Naghshvar & Javidi (2013a) compare adaptive and non-adaptive action selection for a specified finite collection of hypotheses, actions, and observation laws. Their guarantees depend on the given observation model.

Adaptive question selection has a long history in psychometrics and computerized testing (Weiss, 1982; Gershon, 2005; Ghosh & Lan, 2021). Item-response models have also been used to analyze the difficulty and discriminative value of NLP evaluation examples (Lalor et al., 2016; Rodriguez et al., 2021; Maia Polo et al., 2024). Recent model-evaluation work uses predicted item difficulty or current model performance to choose questions more efficiently (Truong et al., 2025; Zhuang et al., 2025; Ding et al., 2026), and active estimators reduce the number of benchmark items needed for score estimation (Wu et al., 2026). These works either fix a particular testing objective or a specified hypothesis-and-action model. They do not characterize learnability for arbitrary pairs of distribution classes under conditional queries, or determine the worst-case cost of fixing all queries in advance. We address both questions.

We study this question through conditional distribution testing on a finite outcome space $\mathcal{X} = [N]$. The N outcomes are the behavior categories distinguished by an evaluation, such as answer choices, tool actions, or judge labels. The unknown object is a full-support distribution P on $\mathcal{X}$. A binary testing target specifies two hypothesis classes, $\mathcal{P}_0$ and $\mathcal{P}_1$, with the promise that $P \in \mathcal{P}_0 \cup \mathcal{P}_1$. The tester must determine which class contains P. We compare adaptive and non-adaptive testers under the same conditional-query model. Both may select any nonempty event $A \subseteq \mathcal{X}$ and receive a sample from $P(\cdot \mid A)$. The only difference is when the queried events are chosen. A non-adaptive tester fixes all events before observing any samples, whereas an adaptive tester may select each event from the observed history.

The model raises two questions. First, which pairs $(\mathcal{P}_0, \mathcal{P}_1)$ are learnable at all, in the sense that their worst-case error can be driven arbitrarily close to zero as the query budget grows? Second, among learnable pairs, how many additional queries are needed when all queried events must be fixed in advance? Since $\mathcal{X}$ has $2^N - 1$ nonempty events, a non-adaptive tester could sample every event sufficiently many times and then replay any finite-round adaptive policy. This elementary argument gives an exponential upper bound, but does not reveal the true cost of removing adaptive rounds. We seek both an exact learnability criterion and the smallest worst-case increase in query count. We call the ratio between the optimal non-adaptive and adaptive query complexities at a common error level the *adaptivity gap*.

Our contributions are summarized below.

- **An exact characterization of conditional-query learnability.** Define the pairwise conditional map Θ by $\Theta_{ij}(P) = P(i \mid \{i, j\})$. A pair of hypothesis classes is learnable if and only if its two images under Θ have positive pairwise separation. When this separation is zero, the optimal worst-case error under either adaptive or non-adaptive testing is exactly $1/2$ for every finite query budget.
- **A universal non-adaptive simulation bound.** For any T-round adaptive policy and any $\rho \in (0, 1)$, we construct a randomized non-adaptive procedure using $O\big(N^2(T + \log(1/\rho))\big)$ pair queries whose simulated transcript is within ρ in total variation of the adaptive transcript, uniformly over all full-support distributions. The entire pair sequence is chosen before any response is observed, and the bound does not depend on the cardinality or pairwise separation of the hypothesis classes.
- **Tightness of the non-adaptive simulation bound.** We construct a family whose adaptive query complexity is independent of N at fixed error, while every non-adaptive test requires $\Omega_\varepsilon(N^2)$ queries. This example shows that the quadratic dependence on N in the preceding simulation bound is unavoidable and gives the exact worst-case rate

$$\mathsf{G}_N(\varepsilon) = \Theta_\varepsilon(N^2).$$

Our results replace the exponential enumeration argument with a tight quadratic guarantee. A static conditional benchmark can match any T-query adaptive evaluation using $O\big(N^2(T + \log(1/\rho))\big)$ pair queries chosen in advance, up to transcript error ρ, and the lower bound shows that the quadratic dependence on N can be unavoidable. Thus, within this model, interaction can reduce the required number of tests by a quadratic factor, but cannot produce a larger worst-case gain from adaptive

selection alone. The apparent exponential branching of an interactive evaluation does not translate into exponential query savings.

2 PROBLEM SETUP

We model an evaluated system by an unknown distribution P over a finite set of observable outcomes. The evaluation must distinguish between two alternatives, $P \in \mathcal{P}_0$ and $P \in \mathcal{P}_1$, using samples from conditional distributions selected by the tester. Choosing every conditioning event before observing any responses gives the non-adaptive protocol, while allowing later events to depend on earlier observations gives the adaptive protocol. This section defines both protocols and the quantities used to compare them.

Hypotheses and conditional queries. Let $\mathcal{X}$ be a finite outcome space with $|\mathcal{X}| = N$, and let

$$\Delta_+(\mathcal{X}) := \{P \in \Delta(\mathcal{X}) : P(x) > 0 \text{ for every } x \in \mathcal{X}\}.$$

The testing problem is specified by two nonempty hypothesis classes $\mathcal{P}_0, \mathcal{P}_1 \subseteq \Delta_+(\mathcal{X})$. These classes may be arbitrary subsets of $\Delta_+(\mathcal{X})$ and need not be finite. Under a hidden label $b \in \{0, 1\}$, the unknown distribution P is selected from $\mathcal{P}_b$. The tester observes P only through conditional queries and must estimate b.

The query family is the collection of nonempty events $\mathcal{A} := 2^{\mathcal{X}} \setminus \{\varnothing\}$. A query $A \in \mathcal{A}$ returns a fresh observation from

$$P_A(x) := P(x \mid A) = \frac{P(x)\mathbf{1}\{x \in A\}}{P(A)}.$$

Full support ensures that every nonempty event is a valid conditional query.

Adaptive and non-adaptive testers. An adaptive tester may select each event after observing the responses to its earlier queries. After $t - 1$ queries, the complete observed history is

$$H_{t-1} := (A_1, X_1, \ldots, A_{t-1}, X_{t-1}).$$

A randomized adaptive policy selects

$$A_t \sim \pi_t(\cdot \mid H_{t-1}), \qquad X_t \sim P_{A_t},$$

where $\pi_t(\cdot \mid H_{t-1})$ is a probability distribution on $\mathcal{A}$. After T queries, a possibly randomized decision rule outputs $\widehat{b} \in \{0, 1\}$. A non-adaptive tester instead draws the entire query vector $A_{1:T}$ before observing any samples. The queries may be randomized, correlated, and repeated, but their joint distribution cannot depend on the observations.

Risk, learnability, and the adaptivity gap. The minimax risk records the smallest worst-case error achievable with a fixed query budget. The adaptive minimax risk at horizon T is

$$R_T^{\mathrm{ad}}(\mathcal{P}_0, \mathcal{P}_1) := \inf_{\text{adaptive testers}} \max_{b \in \{0,1\}} \sup_{P \in \mathcal{P}_b} \Pr_P(\widehat{b} \neq b).$$

The non-adaptive risk $R_T^{\mathrm{na}}(\mathcal{P}_0, \mathcal{P}_1)$ is defined by restricting the infimum to non-adaptive testers. Therefore

$$R_T^{\mathrm{ad}}(\mathcal{P}_0, \mathcal{P}_1) \leq R_T^{\mathrm{na}}(\mathcal{P}_0, \mathcal{P}_1) \leq \frac{1}{2}.$$

Learnability asks whether this risk can be made arbitrarily small as the query budget grows. We call $(\mathcal{P}_0, \mathcal{P}_1)$ adaptively conditionally learnable if $R_T^{\mathrm{ad}}(\mathcal{P}_0, \mathcal{P}_1) \to 0$ as $T \to \infty$. Non-adaptive conditional learnability is defined in the same way using R_T^{na}. For a target error $\varepsilon \in (0, 1/2)$, define

$$T_{\mathrm{ad}}^\star(\varepsilon; \mathcal{P}_0, \mathcal{P}_1) := \inf\left\{T \geq 0 : R_T^{\mathrm{ad}}(\mathcal{P}_0, \mathcal{P}_1) \leq \varepsilon\right\},$$

and define $T_{\mathrm{na}}^\star(\varepsilon; \mathcal{P}_0, \mathcal{P}_1)$ analogously. The infimum of an empty set is ∞.

For a class pair with finite adaptive query complexity at error ε, its fixed-error adaptivity gap is

$$\mathrm{Gap}_\varepsilon(\mathcal{P}_0, \mathcal{P}_1) := \frac{T_{\mathrm{na}}^\star(\varepsilon; \mathcal{P}_0, \mathcal{P}_1)}{T_{\mathrm{ad}}^\star(\varepsilon; \mathcal{P}_0, \mathcal{P}_1)}.$$

For a fixed outcome-space size N, the worst-case adaptivity gap is

$$\mathsf{G}_N(\varepsilon) := \sup_{\substack{\mathcal{P}_0,\mathcal{P}_1 \subseteq \Delta_+(\mathcal{X}) \\ 1 \le T^\star_{\mathrm{ad}}(\varepsilon;\mathcal{P}_0,\mathcal{P}_1) < \infty}} \mathrm{Gap}_\varepsilon(\mathcal{P}_0,\mathcal{P}_1).$$

We count conditional queries and place no restriction on the computation used to choose the queries or form the final decision.

3 Exact Characterization of Conditional-Query Learnability

This section gives an exact criterion for conditional-query learnability. The response law to a query on $\{i,j\}$ depends on P only through the relative weight $P(i)/P(j)$, and the collection of all such ratios determines P. We show that two hypothesis classes are learnable precisely when their images under this pairwise representation have a positive uniform separation. Sufficiency follows by estimating every pair in advance. Necessity follows by showing that pairwise closeness controls every conditional query and therefore every finite adaptive transcript.

Pairwise representation and main theorem. Fix an arbitrary labeling $\mathcal{X} = \{1,\dots,N\}$ and let $M := \binom{N}{2}$. For $1 \le i < j \le N$, define

$$\theta_{ij}(P) := P(i \mid \{i,j\}) = \frac{P(i)}{P(i)+P(j)}, \qquad \Theta(P) := \big(\theta_{ij}(P)\big)_{i<j} \in (0,1)^M.$$

For $i < j$, these coordinates determine the ratio

$$\frac{P(i)}{P(j)} = \frac{\theta_{ij}(P)}{1-\theta_{ij}(P)}.$$

The reverse ratio is its reciprocal. Together with the normalization $\sum_i P(i) = 1$, these ratios determine P. The same pairwise normalization underlies classical probabilistic models for paired comparisons (Bradley & Terry, 1952; Hunter, 2004). For $P, Q \in \Delta_+(\mathcal{X})$, write

$$d_{\mathrm{pair}}(P,Q) := \|\Theta(P) - \Theta(Q)\|_\infty.$$

The pairwise separation between the two hypothesis classes is

$$\Delta_{\mathrm{pair}}(\mathcal{P}_0,\mathcal{P}_1) := \inf_{P\in\mathcal{P}_0,\, Q\in\mathcal{P}_1} d_{\mathrm{pair}}(P,Q).$$

Theorem 3.1 (Exact learnability criterion)**.** *Suppose $N \ge 2$. The following statements are equivalent.*

1. $\Delta_{\mathrm{pair}}(\mathcal{P}_0,\mathcal{P}_1) > 0$.
2. $R^{\mathrm{na}}_T(\mathcal{P}_0,\mathcal{P}_1) \to 0$ *as* $T \to \infty$.
3. $R^{\mathrm{ad}}_T(\mathcal{P}_0,\mathcal{P}_1) \to 0$ *as* $T \to \infty$.

If $\Delta_{\mathrm{pair}}(\mathcal{P}_0,\mathcal{P}_1) = 0$, then for every finite T,

$$R^{\mathrm{ad}}_T(\mathcal{P}_0,\mathcal{P}_1) = R^{\mathrm{na}}_T(\mathcal{P}_0,\mathcal{P}_1) = \frac{1}{2}.$$

If $\Delta := \Delta_{\mathrm{pair}}(\mathcal{P}_0,\mathcal{P}_1) > 0$, then for every $\varepsilon \in (0,1/2)$,

$$T^\star_{\mathrm{na}}(\varepsilon;\mathcal{P}_0,\mathcal{P}_1) \le M \left\lceil \frac{8}{\Delta^2} \log \frac{2M}{\varepsilon} \right\rceil.$$

Sufficiency: positive separation gives a non-adaptive test. Suppose $\Delta := \Delta_{\mathrm{pair}}(\mathcal{P}_0,\mathcal{P}_1) > 0$. Query every pair the same number of times and let $\widehat{\Theta}$ collect the empirical frequencies. For $b \in \{0,1\}$, define

$$D_b := \inf_{Q\in\mathcal{P}_b} \|\widehat{\Theta} - \Theta(Q)\|_\infty,$$

and output the label with smaller D_b. If the true distribution $P \in \mathcal{P}_b$ satisfies $\|\widehat{\Theta} - \Theta(P)\|_\infty < \Delta/4$, then $D_b < \Delta/4$. The triangle inequality gives $D_{1-b} \ge \Delta - \Delta/4 = 3\Delta/4$. Hoeffding's inequality and a union bound over the M pairs give this event with probability at least $1-\varepsilon$ using the number of queries stated in the theorem. Thus positive separation already yields a non-adaptive test.

Necessity: pairwise closeness controls every query. Fix $P, Q \in \Delta_+(\mathcal{X})$ and let $d := d_{\mathrm{pair}}(P, Q)$. Pairwise closeness first bounds the response laws under every conditional query. Applying this bound at each round then controls the law of the entire adaptive transcript:

$$\begin{aligned} d_{\mathrm{pair}}(P, Q) = d \quad &\Longrightarrow \quad \max_{\varnothing \neq A \subseteq \mathcal{X}} \mathrm{TV}(P_A, Q_A) \leq (N-1)d, \\ &\Longrightarrow \quad \mathrm{TV}\big(\mathsf{L}_P(H_T), \mathsf{L}_Q(H_T)\big) \leq T(N-1)d. \end{aligned}$$

From pairs to arbitrary events. Fix a nonempty event A and write $p = P_A$ and $q = Q_A$. For distinct $i, j \in A$, conditioning once more on $\{i, j\}$ gives

$$\frac{|p_i q_j - q_i p_j|}{(p_i + p_j)(q_i + q_j)} = \left| \frac{p_i}{p_i + p_j} - \frac{q_i}{q_i + q_j} \right| \leq d.$$

Let $I_+ := \{i : p_i \geq q_i\}$ and $I_- := A \setminus I_+$. Total variation can be written as the mass transported across this cut:

$$\begin{aligned} \mathrm{TV}(p, q) &= \sum_{i \in I_+} \sum_{j \in I_-} (p_i q_j - q_i p_j) \\ &\leq d \sum_{\{i,j\} \subseteq A} (p_i + p_j)(q_i + q_j) \\ &= d \left(1 + (|A| - 2) \sum_{i \in A} p_i q_i \right) \leq (|A| - 1)d. \end{aligned}$$

Consequently,

$$\max_{\varnothing \neq A \subseteq \mathcal{X}} \mathrm{TV}(P_A, Q_A) \leq (N-1) d_{\mathrm{pair}}(P, Q). \tag{1}$$

From events to adaptive transcripts. Couple two executions of the same tester under P and Q. As long as their histories agree, use the same internal randomness, so the tester selects the same next event in both executions. Couple the two conditional observations optimally. Equation (1) bounds the probability that they first disagree at any round by $(N-1)d_{\mathrm{pair}}(P, Q)$. A union bound over T rounds therefore gives

$$\mathrm{TV}\big(\mathsf{L}_P(H_T), \mathsf{L}_Q(H_T)\big) \leq T(N-1) d_{\mathrm{pair}}(P, Q).$$

Zero separation forces risk $1/2$. Suppose $\Delta_{\mathrm{pair}}(\mathcal{P}_0, \mathcal{P}_1) = 0$ and fix any finite T-query adaptive tester. The conclusion is immediate for $T = 0$. For $T \geq 1$ and every $\eta > 0$, there exist $P \in \mathcal{P}_0$ and $Q \in \mathcal{P}_1$ with $d_{\mathrm{pair}}(P, Q) < \eta / [T(N-1)]$. Their transcript distributions are then less than η apart in total variation. For any decision rule, its error probabilities under P and Q sum to at least $1 - \eta$, so its worst-case error is at least $(1 - \eta)/2$. Letting η tend to zero and comparing with a fair random guess gives the exact minimax risk $1/2$. Appendix A supplies the concentration details and the full formal proof.

Thus pairwise separation exactly characterizes conditional-query learnability. We now compare the query counts of adaptive and non-adaptive tests on the learnable problems.

4 Non-Adaptive Simulation of Adaptive Queries

The proof builds on the pairwise coupling-from-the-past (CFTP) sampler of Fotakis et al. (2022), using the CFTP principle of Propp & Wilson (1996) and the read-once approach of Wilson (2000). Our analysis exploits uniform sampling over all pairs to control the total cost of simulating an adaptive transcript independently of the unknown distribution. Fix one round after the adaptive policy has selected an event A. We have only pair samples collected before A was known, and we need one draw from P_A. Each pair sample defines a random map on A that preserves P_A. Composing these maps backward until all starting states have the same output gives an exact sample from P_A. We first describe this one-round construction, then bound its cost and apply it successively to all T adaptive rounds. Appendix B gives a detailed comparison with the prior sampling results.

Theorem 4.1 (Universal transcript simulation). *Let $N \geq 2$, let $T \geq 1$, and let π be any randomized adaptive policy using T conditional queries. Write $M = \binom{N}{2}$. For every $\rho \in (0,1)$, there is a randomized non-adaptive procedure that makes at most*

$$L := \left\lceil 8M \left(T + \log \frac{1}{\rho} \right) \right\rceil \tag{2}$$

pair queries and produces a T-round simulated transcript $\widetilde{H}_T$ such that, for every $P \in \Delta_+(\mathcal{X})$,

$$\mathrm{TV}\big(\mathsf{L}_P(\widetilde{H}_T), \mathsf{L}_P(H_T^\pi)\big) \leq \rho.$$

Consequently, any decision rule with error at most ε under π has error at most $\varepsilon + \rho$ when applied to the simulated transcript.

Construction. The procedure first collects a fixed random sequence of pair samples. After a later event A is selected, a stored sample (e, y) defines $\Phi_{e,y}^A : A \to A$ by

$$\Phi_{e,y}^A(x) := \begin{cases} y, & e \subseteq A \text{ and } x \in e, \\ x, & \text{otherwise,} \end{cases} \qquad x \in A.$$

It merges the endpoints of e at y when $e \subseteq A$ and is otherwise the identity. Algorithm 1 composes unused maps until every possible starting state has the same output. Samples are never reused; if the stream is exhausted, a fixed rule completes a valid transcript without querying P.

Algorithm 1 Simulating a T-round adaptive policy from pair samples chosen in advance

Require: policy π, stream length L
Precommitted data collection
Draw $E_{1:L}$ i.i.d. uniformly from $\binom{\mathcal{X}}{2}$, submit all queries, and store $Y_s \sim P_{E_s}$.
Offline reconstruction
Set $s \leftarrow 1$ and $\widetilde{H}_0 \leftarrow \varnothing$.
for $t = 1, \ldots, T$ **do**
  Draw $A_t \sim \pi_t(\cdot \mid \widetilde{H}_{t-1})$.
  Set $G \leftarrow \mathrm{id}_{A_t}$.
  while G is not constant on A_t **do**
    if $s > L$ **then**
      return a valid T-round completion under π.
    end if
    $G \leftarrow G \circ \Phi_{E_s,Y_s}^{A_t}$ and $s \leftarrow s + 1$.
  end while
  Let $\widetilde{X}_t$ be the common value of G on A_t.
  Set $\widetilde{H}_t \leftarrow (\widetilde{H}_{t-1}, A_t, \widetilde{X}_t)$.
end for
return $\widetilde{H}_T$

Step 1: one pair sample preserves the target law. Fix A and one unused pair sample (e, Y). If $e \not\subseteq A$, then $\Phi_{e,Y}^A$ is the identity. If $e \subseteq A$, let $X \sim P_A$ be independent of $Y \sim P_e$. For $z \in e$,

$$\Pr\big(\Phi_{e,Y}^A(X) = z\big) = P_A(e)P_e(z) = \frac{P(e)}{P(A)} \frac{P(z)}{P(e)} = P_A(z).$$

For $z \in A \setminus e$, the mass is unchanged. Thus every unused map preserves P_A. Equivalently, conditional on $X \in e$, the law of X is exactly the law P_e used to redraw the pair. This uses the nested-conditioning relation $P_A(\cdot \mid e) = P_e$ for $e \subseteq A$; it need not hold for a generic prompt-response kernel.

Step 2: coalescence produces an exact sample. Write $\Phi_1^A, \Phi_2^A, \ldots$ for the successive maps used for A and set

$$C_n^A := \Phi_1^A \circ \Phi_2^A \circ \cdots \circ \Phi_n^A, \qquad \tau_A := \inf\{n : C_n^A \text{ is constant on } A\}.$$

Interpret Φ_1^A as the update from time -1 to the present, Φ_2^A as the update from time -2 to time -1, and so on. Thus C_n^A maps a state at time $-n$ to the fixed present. For analysis, start with an independent $Z_n \sim P_A$ at time $-n$. Repeated application of Step 1 gives $C_n^A(Z_n) \sim P_A$.

If C_n^A is constant, its output no longer depends on the unavailable initial state Z_n. The order of composition ensures more: if $C_n^A \equiv w$, then $C_n^A \circ \Phi_{n+1}^A \equiv w$. We are extending the starting time farther into the past while keeping the present fixed, so an already determined present value cannot change. Let W_A be this common value when coalescence first occurs. On $\{\tau_A \le n\}$, $C_n^A(Z_n) = W_A$. Step 3 shows that $\tau_A < \infty$ almost surely, and therefore $\mathrm{TV}(\mathsf{L}(W_A), P_A) \le \Pr(\tau_A > n) \to 0$. Hence $W_A \sim P_A$.

Step 3: how many pair samples are needed. The right-composition order above is convenient for exactness but awkward to count. For each fixed n, the maps are i.i.d., so reversing them does not change their joint law. We may therefore run the same maps forward and track the set of possible states. If k states remain, the next map reduces this number to $k-1$ exactly when its uniform pair has both endpoints in that set, an event of probability

$$p_k = \frac{\binom{k}{2}}{M}.$$

It follows that, for $|A| \ge 2$,

$$\tau_A \overset{d}{=} \sum_{k=2}^{|A|} G_k, \qquad G_k \sim \mathrm{Geom}\left(\frac{\binom{k}{2}}{M}\right) \text{ independently.} \tag{3}$$

Consequently,

$$\mathbb{E}\tau_A < 2M, \qquad \mathbb{E}\exp\left(\frac{\tau_A}{8M}\right) \le e.$$

The final transition, from two possible states to one, already has expected waiting time $M = \binom{N}{2}$. This explains the quadratic scale. Appendix B gives the geometric-waiting-time and exponential-moment calculations.

Step 4: from one response to the adaptive transcript. Let H_t^∞ be the reconstruction from an infinite pair-sample stream and let S_t be the total number of samples consumed through round t. After round $t-1$, the unused suffix is still i.i.d. and independent of the simulated history because S_{t-1} is a stopping time. Conditional on the next selected event A_t, Steps 1–2 therefore give

$$X_t^\infty \mid H_{t-1}^\infty, A_t \sim P_{A_t}.$$

Induction gives $H_T^\infty \overset{d}{=} H_T^\pi$. The conditional exponential-moment calculation in Appendix B also gives

$$\Pr\left[S_T > 8M\left(T + \log\frac{1}{\rho}\right)\right] \le \rho.$$

Couple the finite and infinite reconstructions using the same policy randomness and first L pair samples. They agree on $\{S_T \le L\}$, so with the choice of L in equation (2),

$$\mathrm{TV}\big(\mathsf{L}_P(\widetilde{H}_T), \mathsf{L}_P(H_T^\pi)\big) \le \Pr(S_T > L) \le \rho.$$

This proves Theorem 4.1.

Corollary 4.2 (Quadratic upper bound on the adaptivity gap)**.** *For every fixed* $\varepsilon \in (0, 1/2)$, $\mathsf{G}_N(\varepsilon) = O_\varepsilon(N^2)$.

Repeat an ε-error adaptive tester $O_\varepsilon(1)$ times so that its majority vote has error at most $\varepsilon/2$, then apply Theorem 4.1 with $\rho = \varepsilon/2$. This proves the corollary. The exact finite-error bound and the case of an unattained minimax infimum appear in Appendix B. The next section shows that the quadratic factor cannot be improved.

5 A MATCHING QUADRATIC LOWER BOUND

Corollary 4.2 shows that an $O(N^2)$ increase in query count is always sufficient to remove adaptive rounds. We now show that this factor is sometimes necessary. The unknown distribution will be indexed by a class label $b \in \{0,1\}$ and two additional indices $(j,k) \in [m]^2$. The tester must output b; it need not output j or k. Queries to two fixed sets determine j and k but have the same law under both labels. Once those indices are known, one particular pair distinguishes the labels.

Construction of the hard family. Fix $m \geq 2$, let $N = 4m$, and partition the outcome space into

$$I = \{a_1, \ldots, a_m\}, \quad J = \{b_1, \ldots, b_m\}, \quad U = \{u_1, \ldots, u_m\}, \quad V = \{v_1, \ldots, v_m\}.$$

For every $b \in \{0,1\}$ and $(j,k) \in [m]^2$, we define a distribution $P_{b,j,k}$. Write $B_{jk} := \{u_j, v_k\}$. The construction is chosen to produce the following three conditional laws.

Query	Conditional behavior	What it determines
I	$P_{b,j,k}(a_j \mid I) = 3/4$ for both labels	index j
J	$P_{b,j,k}(b_k \mid J) = 3/4$ for both labels	index k
B_{jk}	$P_{0,j,k}(u_j \mid B_{jk}) = 3/4$, $P_{1,j,k}(u_j \mid B_{jk}) = 1/4$	label b

The following masses realize these laws. On I, set

$$P_{b,j,k}(a_j) = \frac{3}{16}, \qquad P_{b,j,k}(a_r) = \frac{1}{16(m-1)} \quad (r \neq j),$$

and define the masses on J analogously with b_k in place of a_j. Each of I and J has total mass $1/4$, independently of the label.

For the remaining points, let $\eta = m^{-4}$ and $Z = 2m - 2 + 4\eta$. Every point in $U \cup V$ outside $\{u_j, v_k\}$ has mass $1/(2Z)$. On B_{jk}, define

$$\big(P_{0,j,k}(u_j), P_{0,j,k}(v_k)\big) = \left(\frac{3\eta}{2Z}, \frac{\eta}{2Z}\right), \qquad \big(P_{1,j,k}(u_j), P_{1,j,k}(v_k)\big) \quad = \left(\frac{\eta}{2Z}, \frac{3\eta}{2Z}\right).$$

All masses are positive and sum to one. The two distributions with the same (j,k) differ only at u_j and v_k. The label signal is hidden in one very light pair B_{jk}. Conditioning exactly on B_{jk} cancels its small total mass and reveals the constant $3/4$-versus-$1/4$ bias. If a query contains any additional ordinary atom, that atom dilutes the label-dependent discrepancy to $O(\eta)$. We choose $\eta = m^{-4}$ so that even $O(m^2)$ off-target queries contribute only $o(1)$ total KL divergence; therefore a non-adaptive tester essentially has to guess which of the m^2 target pairs to query. Define

$$\mathcal{P}_b^{(m)} := \{P_{b,j,k} : j,k \in [m]\}.$$

Theorem 5.1 (Matching quadratic lower bound)**.** *For every $m \geq 2$ and $\varepsilon \in (0, 1/2)$, write $n_\varepsilon := \lceil 8 \log(3/\varepsilon) \rceil$. There is a constant $c_\varepsilon > 0$ depending only on ε such that*

$$T^\star_{\mathrm{ad}}(\varepsilon; \mathcal{P}_0^{(m)}, \mathcal{P}_1^{(m)}) \leq 3n_\varepsilon, \qquad c_\varepsilon m^2 \leq T^\star_{\mathrm{na}}(\varepsilon; \mathcal{P}_0^{(m)}, \mathcal{P}_1^{(m)}) \leq (m^2+2)n_\varepsilon. \tag{4}$$

Hence this family has adaptivity gap $\Theta_\varepsilon(m^2) = \Theta_\varepsilon(N^2)$. After padding the construction with at most three common atoms, $\mathsf{G}_N(\varepsilon) = \Omega_\varepsilon(N^2)$ for every $N \geq 8$.

Step 1: locate the target pair adaptively. Query I independently n_ε times and let $\widehat{j}$ be the unique index whose atom appears more than $n_\varepsilon/2$ times, using an arbitrary fallback if no such index exists. Since a_j appears with probability $3/4$, Hoeffding's inequality gives

$$\Pr(\widehat{j} \neq j) \leq e^{-n_\varepsilon/8}.$$

The same procedure on J produces $\widehat{k}$ with the same error bound. After observing these samples, query $B_{\widehat{j}\widehat{k}}$ another n_ε times and output zero if $u_{\widehat{j}}$ appears more than $n_\varepsilon/2$ times. Conditional on $\widehat{j} = j$ and $\widehat{k} = k$, this final decision has error at most $e^{-n_\varepsilon/8}$. Since this quantity is at most $\varepsilon/3$, a union bound proves the adaptive bound in equation (4).

Step 2: only the target pair has a constant separation. Fix $s = (j, k)$, write $P_{b,s} := P_{b,j,k}$, and abbreviate $P_0 = P_{0,s}$, $P_1 = P_{1,s}$, and $B_s = B_{jk}$. On the target pair,

$$\mathrm{KL}\big((P_0)_{B_s} \,\|\, (P_1)_{B_s}\big) = \mathrm{KL}\big((3/4, 1/4) \,\|\, (1/4, 3/4)\big) = \frac{1}{2}\log 3 =: d_0.$$

For $A \neq B_s$, the two conditional laws either coincide or A contains a common atom whose mass dominates the total discrepancy on u_j and v_k. The normalization calculation in Appendix C, along with the fact that every conditional likelihood ratio lies between $1/9$ and 9, gives

$$\mathrm{TV}\big((P_0)_A, (P_1)_A\big) \leq 16\eta, \qquad \mathrm{KL}\big((P_0)_A \,\|\, (P_1)_A\big) \leq 32(\log 9)\eta =: C_0\eta \quad (A \neq B_s). \tag{5}$$

Thus a query to B_s has a constant KL divergence, whereas every other query has divergence of order m^{-4}.

Step 3: some target pair receives at most T/m^2 queries. Consider any randomized non-adaptive tester using T queries. For each $s = (j, k)$, let N_s count how many submitted queries are exactly B_s. For every realized query vector, $\sum_{s\in[m]^2} N_s \leq T$. The distribution of this query vector is the same for every unknown $P_{b,j,k}$. Averaging over the m^2 indices therefore gives one s such that $\mathbb{E}N_s \leq T/m^2$. This is the point at which nonadaptivity is used: after observing responses from I and J, an adaptive tester may choose a different target pair for each (j, k).

Step 4: worst-case accuracy requires $T = \Omega(m^2)$. For the selected s, the KL chain rule and equation (5) give

$$\mathrm{KL}\big(\mathsf{L}_{P_{0,s}}(H_T) \,\|\, \mathsf{L}_{P_{1,s}}(H_T)\big) \leq d_0\mathbb{E}N_s + C_0\eta T \leq T\left(\frac{d_0}{m^2} + \frac{C_0}{m^4}\right).$$

If the tester has worst-case error at most ε, then its error on both $P_{0,s}$ and $P_{1,s}$ is at most ε. Their transcript laws must therefore have total variation distance at least $1 - 2\varepsilon$. Pinsker's inequality, one of the standard relations between statistical divergences (Ali & Silvey, 1966; Reid & Williamson, 2011; Sason & Verdú, 2016), yields

$$T \geq \frac{2(1-2\varepsilon)^2}{\dfrac{\log 3}{2m^2} + \dfrac{32\log 9}{m^4}} \geq c_\varepsilon m^2. \tag{6}$$

For the matching upper bound, query I, J, and every B_{rs} exactly n_ε times, then apply the same majority rules to the stored samples. This uses $(m^2 + 2)n_\varepsilon$ queries and completes equation (4). Appendix C verifies the off-target bound, a valid value of c_ε, and padding to every $N \geq 8$.

6 Scope and Limitations

The upper and lower bounds identify the exact worst-case quadratic cost of removing adaptivity in this model. The model assumes a finite full-support outcome space and exact conditional sampling from every nonempty event. It therefore applies directly when an evaluation has a finite outcome representation and can implement the required conditional queries. Query complexity counts conditional samples, not the cost of specifying or realizing a conditioning event. Finally, the lower bound is worst-case and does not imply that typical evaluation tasks attain a quadratic gain from adaptive selection.

AI use statement

The authors conceived the problem and developed the theory. Generative AI assisted with proof development and checking, literature search, and editing. The authors verified all AI-assisted content and take full responsibility for the paper.

Reproducibility statement

Sections 2–3 state the model and assumptions; the appendix gives complete proofs of all results.

REFERENCES

Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. A chasm between identity and equivalence testing with conditional queries. *Theory of Computing*, 14(19):1–46, 2018. doi: 10.4086/toc.2018.v014a019.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B*, 28(1):131–142, 1966. doi: 10.1111/j.2517-6161.1966.tb00626.x.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing Library, Graduate Surveys*, 9:1–100, 2020. doi: 10.4086/toc.gs.2020.009.

Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1174–1192, 2014. doi: 10.1137/1.9781611973402.87.

Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015. doi: 10.1137/130945508.

Diptarka Chakraborty, Sourav Chakraborty, Gunjan Kumar, and Kuldeep S. Meel. Equivalence testing: The power of bounded adaptivity. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 3592–3600, 2024.

Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th Innovations in Theoretical Computer Science Conference*, 2013. doi: 10.1145/2422436.2422497.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959. doi: 10.1214/aoms/1177706205.

Xuanwen Ding, Chengjun Pan, Zejun Li, Jiwen Zhang, Siyuan Wang, and Zhongyu Wei. AutoJudger: An agent-driven framework for efficient benchmarking of MLLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 15009–15034, 2026.

Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. In *Proceedings of the 28th Conference on Learning Theory*, volume 40, pp. 607–636, 2015.

Dimitris Fotakis, Alkis Kalavasis, and Christos Tzamos. Perfect sampling from pairwise comparisons. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Richard C. Gershon. Computer adaptive testing. *Journal of Applied Measurement*, 6(1):109–127, 2005.

Aritra Ghosh and Andrew S. Lan. BOBCAT: Bilevel optimization-based computerized adaptive testing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2410–2417, 2021. doi: 10.24963/ijcai.2021/332.

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

David R. Hunter. MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32(1):384–406, 2004. doi: 10.1214/aos/1079120141.

Gautam Kamath and Christos Tzamos. Anaconda: A non-adaptive conditional sampling algorithm for distribution testing. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 679–693, 2019.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4110–4124, 2021. doi: 10.18653/v1/2021.naacl-main.324.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657, 2016. doi: 10.18653/v1/D16-1062.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 34303–34326, 2024.

Mohammad Naghshvar and Tara Javidi. Sequentiality and adaptivity gains in active hypothesis testing. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):768–782, 2013a.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013b. doi: 10.1214/13-AOS1144.

Shyam Narayanan. On tolerant distribution testing in the conditional sampling model. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms*, pp. 357–373, 2021. doi: 10.1137/1.9781611976465.23.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020. doi: 10.18653/v1/2020.acl-main.441.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multi-hypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 2388–2404, 2021. doi: 10.18653/v1/2021.acl-long.186.

James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1–2):223–252, 1996.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 4486–4503, 2021. doi: 10.18653/v1/2021.acl-long.346.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. doi: 10.1109/TIT.2016.2603151.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. Dynatask: A framework for creating dynamic AI benchmark tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 174–181, 2022. doi: 10.18653/v1/2022.acl-demo.17.

Sang T. Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 60238–60265, 2025.

David J. Weiss. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4):473–492, 1982. doi: 10.1177/014662168200600408.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025.

David Bruce Wilson. How to couple from the past using a read-once source of randomness. *Random Structures & Algorithms*, 16(1):85–113, 2000.

Skyler Wu, Yash Nair, and Emmanuel J. Candès. Efficient evaluation of LLM performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Yan Zhuang, Qi Liu, Zachary Pardos, Patrick C. Kyllonen, Jiyun Zu, Zhenya Huang, Shijin Wang, and Enhong Chen. Position: AI evaluation should learn from how we test humans. In *Proceedings of the 42nd International Conference on Machine Learning, Position Paper Track*, 2025.

CLARIFYING THE CONTRIBUTION

This paper gives precise characterizations of learnability and the value of interaction under conditional queries, with distinct roles for the individual results and the established mathematical tools. Theorem 3.1 uses the pairwise conditional representation, concentration inequalities, and transcript coupling to characterize uniform learnability of arbitrary pairs of distribution classes by positive separation in their pairwise conditional probabilities. It also shows that zero separation forces the optimal worst-case error to equal $1/2$ at every finite query budget. Theorem 4.1 builds on the pairwise CFTP exact-sampling mechanism of Fotakis et al. (2022). The coalescence structure under uniform sampling over all pairs, together with exponential-moment control of total sample consumption, allows pair samples collected in advance to simulate the entire transcript of any adaptive policy within total variation error ρ, using $O(N^2(T + \log(1/\rho)))$ queries independently of the minimum atom probability. Theorem 5.1 constructs a testing problem in which an adaptive tester first locates the relevant indices and then selects the target event. Conditional KL divergence and an averaging argument give a matching lower bound against non-adaptive testers allowed to query arbitrary events. Together, these results show that, in the exact conditional-query model for a fixed distribution, interaction does not change which pairs of hypothesis classes can be uniformly distinguished, but can change the number of queries required. At a fixed error level, the worst-case efficiency gain is exactly $\Theta_\varepsilon(N^2)$ in the explicit outcome-space size N.

A PROOF OF THE LEARNABILITY CRITERION

We first prove the stability relation used in Section 3.

Lemma A.1 (Pairwise stability of conditional queries)**.** *For any $P, Q \in \Delta_+(\mathcal{X})$ and any nonempty $A \subseteq \mathcal{X}$,*

$$\mathrm{TV}(P_A, Q_A) \le (|A| - 1) d_{\mathrm{pair}}(P, Q).$$

Consequently, equation (1) *holds.*

Proof. The claim is immediate when $|A| = 1$. Suppose $|A| = m \ge 2$ and write $p = P_A$, $q = Q_A$, and $d = d_{\mathrm{pair}}(P, Q)$. For every distinct $i, j \in A$, conditioning p and q again on $\{i, j\}$ gives the same pairwise conditionals as conditioning P and Q directly. Hence

$$\frac{|p_i q_j - q_i p_j|}{(p_i + p_j)(q_i + q_j)} \le d. \tag{A.1}$$

Let $I_+ := \{i \in A : p_i \ge q_i\}$ and $I_- := A \setminus I_+$. Then

$$\begin{aligned}
\mathrm{TV}(p, q) &= \sum_{i \in I_+} \sum_{j \in I_-} (p_i q_j - q_i p_j) \\
&\le d \sum_{i \in I_+} \sum_{j \in I_-} (p_i + p_j)(q_i + q_j) \\
&\le d \sum_{\{i,j\} \subseteq A} (p_i + p_j)(q_i + q_j).
\end{aligned}$$

The final sum equals

$$(m - 1) \sum_i p_i q_i + \sum_{i \ne j} p_i q_j = 1 + (m - 2) \sum_i p_i q_i \le m - 1.$$

This proves the lemma. □

Lemma A.2 (Adaptive transcript stability)**.** *Fix any randomized adaptive tester using T queries. If $\mathsf{L}_P(H_T)$ and $\mathsf{L}_Q(H_T)$ denote its transcript laws under P and Q, then*

$$\mathrm{TV}\big(\mathsf{L}_P(H_T), \mathsf{L}_Q(H_T)\big) \le T(N - 1) d_{\mathrm{pair}}(P, Q).$$

Proof. Couple the two executions one round at a time. As long as their histories agree, use the same private randomness to select the same event A_t, then couple the two observations optimally. Lemma A.1 bounds the conditional probability that the observations first differ at round t by $(N - 1) d_{\mathrm{pair}}(P, Q)$. A union bound over the T rounds bounds the probability that the transcripts differ. The coupling characterization of total variation gives the result. □

Proof of Theorem 3.1. First suppose $\Delta_{\mathrm{pair}}(\mathcal{P}_0, \mathcal{P}_1) = 0$. The case $T = 0$ follows directly from the absence of observations. For $T \geq 1$ and every $\eta > 0$, choose $P \in \mathcal{P}_0$ and $Q \in \mathcal{P}_1$ such that

$$d_{\mathrm{pair}}(P, Q) < \frac{\eta}{T(N-1)}.$$

Lemma A.2 makes the two transcript laws closer than η in total variation. For any decision rule, the sum of its error probabilities under P and Q is at least $1 - \eta$. Its worst-case error is therefore at least $(1 - \eta)/2$. Letting η tend to zero gives an adaptive minimax risk of at least $1/2$. A fair random guess achieves $1/2$, so the adaptive risk is exactly $1/2$. The same conclusion follows for non-adaptive testing from $R_T^{\mathrm{ad}} \leq R_T^{\mathrm{na}} \leq 1/2$.

Now suppose $\Delta := \Delta_{\mathrm{pair}}(\mathcal{P}_0, \mathcal{P}_1) > 0$ and fix $\delta \in (0, 1/2)$. Query every pair

$$n := \left\lceil \frac{8}{\Delta^2} \log \frac{2M}{\delta} \right\rceil$$

times. Let $\widehat{\theta}_{ij}$ be the empirical frequency of outcome i from queries to $\{i, j\}$, and let $\widehat{\Theta} = (\widehat{\theta}_{ij})_{i<j}$. Hoeffding's inequality and a union bound give

$$\Pr_P \left(\|\widehat{\Theta} - \Theta(P)\|_\infty \geq \frac{\Delta}{4} \right) \leq 2M \exp\left(-\frac{n\Delta^2}{8} \right) \leq \delta.$$

For $b \in \{0, 1\}$, define

$$D_b := \inf_{Q \in \mathcal{P}_b} \|\widehat{\Theta} - \Theta(Q)\|_\infty,$$

and output a label minimizing D_b. If $P \in \mathcal{P}_b$ and $\|\widehat{\Theta} - \Theta(P)\|_\infty < \Delta/4$, then $D_b < \Delta/4$ while $D_{1-b} \geq 3\Delta/4$. The decision is therefore correct with probability at least $1 - \delta$, uniformly over both classes. The test is non-adaptive and uses Mn queries. This proves the finite-sample bound and all three equivalences. □

B Proof of the Universal Transcript Simulation

Relation to prior perfect-sampling results. For $A = \mathcal{X}$, our endpoint-merging update and backward composition coincide with Algorithm 1 of Fotakis et al. (2022) when its pair distribution is uniform over all pairs. The single-sample exactness mechanism therefore follows existing CFTP constructions. Their Theorem 2 bounds the sampling cost through the minimum atom probability and the spectral gap of the induced Markov chain; their Theorem 3 obtains a distribution-structure-independent guarantee under an additional bounded-probability-ratio assumption. Our analysis instead uses the complete, uniformly sampled pair family to obtain a cost bound valid for every full-support distribution. In the forward image process, when k states remain, the probability of a merger is exactly $\binom{k}{2}/\binom{N}{2}$, regardless of the distribution. Lemma B.2 consequently gives an exact geometric-waiting-time decomposition and a uniform exponential-moment bound for every conditioning event A. To obtain Theorem 4.1, we then simulate events chosen from the preceding simulated history, using the independence of the unconsumed sample stream, and control the total consumption across all rounds. This yields the fixed budget $O(N^2(T + \log(1/\rho)))$ for the entire transcript, without a minimum-mass or probability-ratio assumption and without an additional $\log T$ factor. The matching lower bound in Theorem 5.1 further shows that the quadratic overhead is unavoidable for conditional hypothesis testing, even when non-adaptive testers may query arbitrary events.

Fix $P \in \Delta_+(\mathcal{X})$. Let $\omega_s = (E_s, Y_s)$, $s \geq 1$, be an i.i.d. source with

$$E_s \sim \mathrm{Unif}\left(\binom{\mathcal{X}}{2} \right), \qquad Y_s \mid E_s = e \sim P_e.$$

For a nonempty event A, let $F_s^A := \Phi_{E_s, Y_s}^A$ be the random map defined in Section 4.

Lemma B.1 (Stationarity of the pairwise maps)**.** *If $X \sim P_A$ is independent of F_s^A, then $F_s^A(X) \sim P_A$.*

Proof. Condition on $E_s = e$. If $e \not\subseteq A$, the map is the identity. If $e \subseteq A$, it fixes every state outside e and replaces a state in e by an independent draw from P_e. For $z \in e$, the resulting mass at z is

$$P_A(e)P_e(z) = \frac{P(e)}{P(A)}\frac{P(z)}{P(e)} = P_A(z).$$

The mass of each $z \in A \setminus e$ is unchanged. Thus every conditional map kernel preserves P_A, and averaging over E_s proves the claim. □

Interpret F_1^A as the map from time -1 to time 0, F_2^A as the map from time -2 to time -1, and so on. Define

$$C_n^A := F_1^A \circ F_2^A \circ \cdots \circ F_n^A, \qquad \tau_A := \inf\{n \geq 0 : C_n^A \text{ is constant on } A\}.$$

When $\tau_A < \infty$, write W_A for the common value of $C_{\tau_A}^A$.

Lemma B.2 (Exactness and coalescence cost)**.** *For every nonempty $A \subseteq \mathcal{X}$, $\tau_A < \infty$ almost surely and $W_A \sim P_A$. If $a := |A| \geq 2$, then*

$$\tau_A \overset{d}{=} \sum_{k=2}^{a} G_k, \qquad G_k \sim \mathrm{Geom}\left(\frac{\binom{k}{2}}{M}\right) \text{ independently}. \tag{B.1}$$

Consequently,

$$\mathbb{E}\tau_A = 2M\left(1 - \frac{1}{a}\right) < 2M, \tag{B.2}$$

and

$$\mathbb{E}\exp\left(\frac{\tau_A}{8M}\right) \leq e. \tag{B.3}$$

For $a = 1$, the same conclusions hold with $\tau_A = 0$.

Proof. For each fixed n, reversal invariance of the i.i.d. maps gives

$$F_1^A \circ \cdots \circ F_n^A \overset{d}{=} F_n^A \circ \cdots \circ F_1^A.$$

Consider the forward image process $S_0 = A$ and $S_r = F_r^A(S_{r-1})$. If $|S_{r-1}| = k$, its size decreases by one exactly when both endpoints of E_r lie in S_{r-1}. This event has probability

$$p_k := \frac{\binom{k}{2}}{M}.$$

If the pair contains at most one state in the current image, the map either fixes that image or replaces one state by another and its size stays k. Thus the forward time to reach a singleton is the sum of independent geometric waiting times in equation (B.1). The reversal identity shows that this hitting time has the same distribution as τ_A. It is finite almost surely.

Lemma B.1 also proves exactness by the standard coupling-from-the-past argument (Propp & Wilson, 1996). For completeness, let $Z_n \sim P_A$ be independent of the first n maps. Applying the maps from the oldest to the most recent gives $C_n^A(Z_n) \sim P_A$. On $\{\tau_A \leq n\}$, this value equals W_A for every Z_n. Hence

$$\mathrm{TV}(\mathsf{L}(W_A), P_A) \leq \Pr(\tau_A > n),$$

and the right-hand side tends to zero.

The mean follows from

$$\sum_{k=2}^{a} \frac{1}{\binom{k}{2}} = 2\left(1 - \frac{1}{a}\right).$$

It remains to prove the exponential moment. Set $s = 1/(8M)$. If $G \sim \mathrm{Geom}(p)$ on $\{1, 2, \ldots\}$, then

$$\mathbb{E}e^{sG} = \frac{pe^s}{1 - (1-p)e^s} = \frac{e^s}{1-u}, \qquad u := \frac{(1-p)(e^s - 1)}{p}.$$

For every $p = p_k$, we have $s \leq p/8$. Since $s \leq 1/8$ and $e^s - 1 \leq 2s$, it follows that $u \leq 1/4$. Therefore

$$\log \mathbb{E}e^{sG} = s - \log(1-u) \leq s + \frac{4}{3}u \leq \frac{4s}{p}.$$

Using independence in equation (B.1),

$$\begin{aligned}\log \mathbb{E}e^{s\tau_A} &\le 4s \sum_{k=2}^{a} \frac{1}{p_k} \\ &= \frac{1}{2} \sum_{k=2}^{a} \frac{1}{\binom{k}{2}} \\ &= 1 - \frac{1}{a} \le 1,\end{aligned}$$

which proves equation (B.3). □

Proof of Theorem 4.1. First allow an infinite i.i.d. source $(\omega_s)_{s\ge 1}$. Draw the private randomness of π independently. At round t, the policy chooses A_t from the simulated history. Starting at the next unused source element, run the construction above until its maps coalesce on A_t, and set $\widetilde{X}_t = W_{A_t}$.

The number of elements read is a stopping time for the fresh source segment. The strong Markov property of an i.i.d. sequence implies that the unused tail is again i.i.d. and independent of the maps already consumed. This is the read-once coupling-from-the-past construction of Wilson (2000). By Lemma B.2, conditionally on the simulated history, $\widetilde{X}_t \sim P_{A_t}$. Induction over t therefore shows that the infinite-stream simulated transcript has exactly the same law as H_T^{π}.

Let τ_t be the number of source elements consumed at round t and let $S_T = \sum_{t=1}^{T} \tau_t$. Equation (B.3) holds conditionally on every past simulated history, uniformly over the selected event. Iterating conditional expectations gives

$$\mathbb{E} \exp\left(\frac{S_T}{8M}\right) \le e^T.$$

Markov's inequality now yields

$$\begin{aligned}\Pr\left[S_T > 8M\left(T + \log\frac{1}{\rho}\right)\right] &\le \exp\left(-T - \log\frac{1}{\rho}\right) e^T \\ &= \rho. \end{aligned} \tag{B.4}$$

The finite simulation chooses the L pairs in equation (2) before observing any response and runs the same construction until either all T rounds are simulated or the source is exhausted. In the latter case a fixed completion rule returns a valid T-round transcript. Couple it to the infinite construction using the same first L source elements. Their transcripts agree unless $S_T > L$, an event of probability at most ρ by equation (B.4). The coupling characterization of total variation proves the theorem. Applying the same decision rule changes its error by at most ρ. □

Proof of Corollary 4.2. Suppose a T-query adaptive tester has worst-case error at most $\varepsilon < 1/2$. For any $\delta \in (0, 1/2)$, let $K_{\varepsilon,\delta}$ be the smallest odd integer at least

$$\frac{2}{(1-2\varepsilon)^2} \log\frac{2}{\delta}.$$

Run $K = K_{\varepsilon,\delta}$ independent copies and take a majority vote. Hoeffding's inequality gives error at most

$$\exp\left(-\frac{K(1-2\varepsilon)^2}{2}\right) \le \frac{\delta}{2}.$$

This amplified tester uses KT adaptive queries. Apply Theorem 4.1 with $\rho = \delta/2$. The simulated tester has error at most δ and uses at most

$$\left\lceil 8M\left(K_{\varepsilon,\delta}T + \log\frac{2}{\delta}\right)\right\rceil \tag{B.5}$$

pair queries.

For the fixed-error gap, set $T = T^{\star}_{\mathrm{ad}}(\varepsilon; \mathcal{P}_0, \mathcal{P}_1)$ and $\delta = \varepsilon$. If the minimax infimum at horizon T is not attained, choose a T-query adaptive tester with error below any fixed $\bar{\varepsilon} \in (\varepsilon, 1/2)$ and apply the same argument with $K_{\bar{\varepsilon},\varepsilon}$. Since the definition of $\mathsf{G}_N(\varepsilon)$ restricts to $T \ge 1$, division by T gives a constant depending only on ε times $M = \binom{N}{2}$. □

C Proof of the Quadratic Lower Bound

Proof of Theorem 5.1. Fix $m \geq 2$ and $\varepsilon \in (0, 1/2)$, and write $n_\varepsilon = \lceil 8 \log(3/\varepsilon) \rceil$. The masses in the construction are positive. Moreover,

$$P_{b,j,k}(I) = P_{b,j,k}(J) = \frac{1}{4}, \qquad P_{b,j,k}(U \cup V) = \frac{2m - 2 + 4\eta}{2Z} = \frac{1}{2},$$

so each $P_{b,j,k}$ is a full-support probability distribution.

Conditional on I, the distinguished atom a_j has probability $3/4$. Let $\widehat{j}$ be the unique index whose atom appears more than $n_\varepsilon/2$ times among n_ε samples from $P_{b,j,k}(\cdot \mid I)$, with a fixed fallback when no such index exists. Hoeffding's inequality gives

$$\Pr(\widehat{j} \neq j) \leq \Pr\big(\mathrm{Bin}(n_\varepsilon, 3/4) \leq n_\varepsilon/2\big) \leq e^{-n_\varepsilon/8}.$$

The same procedure using J gives $\Pr(\widehat{k} \neq k) \leq e^{-n_\varepsilon/8}$. Conditional on $\widehat{j} = j$ and $\widehat{k} = k$, thresholding the frequency of u_j in n_ε samples from B_{jk} at $1/2$ has error at most $e^{-n_\varepsilon/8}$ under either label. Since $3e^{-n_\varepsilon/8} \leq \varepsilon$, the adaptive tester uses $3n_\varepsilon$ queries and has worst-case error at most ε. A non-adaptive tester obtains the same error by querying I and J n_ε times each and every pair $B_{rs} = \{u_r, v_s\}$ n_ε times before observing any response. It then uses the samples from $B_{\widehat{j}\widehat{k}}$. This proves both upper bounds in equation (4).

It remains to prove the non-adaptive lower bound. Fix $s = (j, k)$, write $P_{b,s} := P_{b,j,k}$, and abbreviate $P_0 = P_{0,s}$, $P_1 = P_{1,s}$, and $B_s = B_{jk}$. The target pair satisfies

$$d_0 := \mathrm{KL}\big((P_0)_{B_s} \,\|\, (P_1)_{B_s}\big) = \mathrm{KL}\big((3/4, 1/4) \,\|\, (1/4, 3/4)\big) = \frac{1}{2} \log 3.$$

We next bound every other query. For a nonempty event A, let $\mu_A = (P_0)_A$ and $\nu_A = (P_1)_A$. The unconditional likelihood ratio $P_0(x)/P_1(x)$ belongs to $\{1/3, 1, 3\}$, which implies

$$\frac{1}{9} \leq \frac{\mu_A(x)}{\nu_A(x)} \leq 9 \qquad (x \in A).$$

For positive finite measures α and β on A, set $a = \alpha(A)$ and $b = \beta(A)$. Then

$$\begin{aligned} 2\,\mathrm{TV}(\alpha_A, \beta_A) &= \sum_{x \in A} \left| \frac{\alpha(x)}{a} - \frac{\beta(x)}{b} \right| \\ &\leq \frac{\sum_{x \in A} |\alpha(x) - \beta(x)|}{a} + \frac{|a - b|}{a}. \end{aligned}$$

Since $|a - b| \leq \sum_{x \in A} |\alpha(x) - \beta(x)|$, and the same inequality holds after exchanging a and b,

$$\mathrm{TV}(\alpha_A, \beta_A) \leq \frac{\sum_{x \in A} |\alpha(x) - \beta(x)|}{\min\{\alpha(A), \beta(A)\}}.$$

The distributions P_0 and P_1 differ only at u_j and v_k, with unconditional ℓ_1 discrepancy $2\eta/Z$. If A contains neither active atom, or if it is a singleton active atom, then $\mu_A = \nu_A$. In every other case with $A \neq B_s$, the event A contains a common atom. Every common atom has mass at least $1/[16(m-1)]$ under both labels. For inactive atoms in $U \cup V$, this follows from $Z \leq 8(m-1)$. Since $Z \geq 2(m-1)$, the preceding normalization inequality gives

$$\mathrm{TV}(\mu_A, \nu_A) \leq \frac{2\eta/Z}{1/[16(m-1)]} \leq 16\eta \qquad (A \neq B_s).$$

If two distributions μ and ν have likelihood ratios in $[1/R, R]$, then

$$\begin{aligned} \mathrm{KL}(\mu \| \nu) + \mathrm{KL}(\nu \| \mu) &= \sum_x (\mu(x) - \nu(x)) \log \frac{\mu(x)}{\nu(x)} \\ &\leq 2(\log R)\, \mathrm{TV}(\mu, \nu). \end{aligned}$$

Consequently,

$$\mathrm{KL}(\mu_A \| \nu_A) \le 32(\log 9)\eta =: C_0\eta \qquad (A \neq B_s).$$

Now fix an arbitrary randomized non-adaptive tester using T queries. Its query vector $A_{1:T}$ has the same law under every distribution in the two classes. For $s \in [m]^2$, let $N_s = \sum_{t=1}^{T} \mathbf{1}\{A_t = B_s\}$. Since $\sum_{s\in[m]^2} N_s \le T$ for every realized query vector, there is an $s \in [m]^2$ for which $\mathbb{E}N_s \le T/m^2$. Conditional on the query vector, the responses are independent. The KL chain rule therefore gives

$$\begin{aligned}
&\mathrm{KL}\big(\mathsf{L}_{P_{0,s}}(H_T) \,\|\, \mathsf{L}_{P_{1,s}}(H_T)\big) \\
&\quad = \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{KL}\big((P_{0,s})_{A_t} \,\|\, (P_{1,s})_{A_t}\big)\right] \\
&\quad \le d_0 \mathbb{E}N_s + C_0\eta T \le T\left(\frac{d_0}{m^2} + \frac{C_0}{m^4}\right).
\end{aligned}$$

If the tester has worst-case error at most ε, data processing through its decision rule implies that these transcript laws have total variation at least $1 - 2\varepsilon$. Pinsker's inequality yields

$$2(1-2\varepsilon)^2 \le T\left(\frac{d_0}{m^2} + \frac{C_0}{m^4}\right).$$

For $m \ge 2$,

$$\frac{d_0}{m^2} + \frac{C_0}{m^4} \le \frac{\frac{1}{2}\log 3 + 8\log 9}{m^2}.$$

Thus equation (6) holds with

$$c_\varepsilon := \frac{2(1-2\varepsilon)^2}{\frac{1}{2}\log 3 + 8\log 9}.$$

Since a zero-query test has worst-case error at least $1/2$, the adaptive query complexity of this family is at least one. Combining this fact with the three bounds in equation (4) shows that its adaptivity gap is $\Theta_\varepsilon(m^2) = \Theta_\varepsilon(N^2)$.

Finally, let $N \ge 8$, set $m = \lfloor N/4 \rfloor$, and write $r = N - 4m$. The case $r = 0$ is already covered. If $r > 0$, add atoms $c_1, \ldots, c_r$ and define

$$\widetilde{P}_{b,j,k}(x) = \frac{7}{8} P_{b,j,k}(x) \quad (x \in I \cup J \cup U \cup V), \qquad \widetilde{P}_{b,j,k}(c_\ell) = \frac{1}{8r}.$$

Let $\widetilde{\mathcal{P}}_b^{(N)} = \{\widetilde{P}_{b,j,k} : j,k \in [m]\}$. The masses sum to $7/8 + 1/8 = 1$, so these are full-support distributions on an outcome space of size N. Conditional laws on events contained in the original four blocks are unchanged, so the adaptive upper bound and the target-pair divergence remain unchanged.

Every off-target event that distinguishes the labels contains a common atom. Each such atom has mass at least $7/(128m)$. Indeed, this follows from the original masses for atoms in $I \cup J$, from $Z \le 8m$ for inactive atoms in $U \cup V$, and from $r \le 3$ and $1/(8r) \ge 7/(128m)$ for the new atoms. The unconditional ℓ_1 discrepancy is

$$\frac{7}{8}\frac{2\eta}{Z} = \frac{7\eta}{4Z} \le \frac{7\eta}{4m},$$

where $Z \ge m$. Hence every off-target conditional total variation is at most 32η. The conditional likelihood ratios remain in $[1/9, 9]$, so every off-target KL divergence is at most $64(\log 9)\eta$. For any randomized non-adaptive tester with T queries, let $\widetilde{N}_s$ count the queries equal to B_s. Since $\sum_s \widetilde{N}_s \le T$, there is an $s \in [m]^2$ such that $\mathbb{E}\widetilde{N}_s \le T/m^2$. The KL chain rule then gives

$$\mathrm{KL}\big(\mathsf{L}_{\widetilde{P}_{0,s}}(H_T) \,\|\, \mathsf{L}_{\widetilde{P}_{1,s}}(H_T)\big) \le T\left(\frac{d_0}{m^2} + \frac{64\log 9}{m^4}\right).$$

If the tester has worst-case error at most ε, Pinsker's inequality implies

$$2(1-2\varepsilon)^2 \le T\left(\frac{d_0}{m^2} + \frac{64\log 9}{m^4}\right),$$

and therefore $T = \Omega_\varepsilon(m^2)$. The adaptive tester above still uses at most $3n_\varepsilon$ queries. Since $m = \lfloor N/4 \rfloor \ge N/8$, the padded family has adaptivity gap $\Omega_\varepsilon(N^2)$. This proves the theorem. □