Simulating Field Experiments for Method Testing

Anonymous Authors[1]

Abstract

Field experiments (A/B tests) are often the most credible benchmark for methods (algorithms) in societal systems, but their cost and latency bottleneck rapid methodological progress. LLM-based persona simulation offers a cheap synthetic alternative, yet it is unclear whether replacing humans with personas preserves the benchmark interface that adaptive methods optimize against. We give a zero-error characterization: when (i) methods observe only the aggregate outcome (aggregate-only observation) and (ii) evaluation depends only on the submitted artifact and not on the method's identity or provenance (method-blind evaluation), swapping humans for personas is indistinguishable on the method interface from a literal panel change, i.e., from changing only the evaluation population (e.g., New York to Jakarta). We also give quantitative stability: approximate method-blindness implies approximate panel-change representation, while provenance sensitivity of size η creates an $\eta/2$ lower bound against any artifact-only representation, with an adaptive repeated-round version. Finally, we move from validity to usefulness: we define an information-theoretic discriminability of the induced aggregate channel and show that making persona benchmarking as decision-relevant as a field experiment is fundamentally a sample-size question, yielding explicit bounds on the number of independent evaluation units required to reliably distinguish meaningfully different methods at a chosen resolution.

1. Introduction

One of the recurring lessons from machine learning is that progress in methods depends heavily on cheap and reliable evaluation benchmarks (Blum & Hardt, 2015; Zaharia et al., 2018; Bommasani et al., 2023; Miller, 2024). In many societal domains, the closest analogue to such a benchmark is a field experiment or A/B test: pricing policies, marketplace interventions, matching rules, and behavioral treatments are often judged by how they perform when deployed on a real population (Harrison & List, 2004; Kohavi et al., 2020). But field experiments are expensive and slow, so they bottleneck method development (Diamond, 1986; Paluck & Cialdini, 2014; Samek, 2019).

LLM-based persona simulation offers a tempting synthetic alternative. A language model can be conditioned on respondent profiles, generate micro-level judgments or behaviors, and those micro-responses can then be aggregated into a benchmark score (Toubia et al., 2025). This makes large-scale evaluation cheap. At the same time, recent work has emphasized that LLM personas should not be treated as general substitutes for studying human behaviors (Peng et al., 2025; Li et al., 2025). That broader limitation still leaves open a narrower question that is specific to benchmarking:

When can LLM-based persona simulation serve as a drop-in substitute for a field experiment (or A/B test) as a benchmark for comparing methods?

This is not a score-matching question: persona and human evaluations need not numerically agree, just as the same method can score differently in New York and Jakarta. The relevant issue is whether replacing humans with personas preserves the *benchmark interface* seen by an adaptive method, i.e., whether swapping evaluators is indistinguishable, on the method interface, from changing only the evaluation population. This is an identification question, distinct from population-transfer validity—we characterize when the swap behaves like a panel change, not whether conclusions transfer back to a target human population.

In this paper, we give a quantitative characterization of when this benchmark-level equivalence holds, by identifying two benchmark-hygiene conditions:

(i) *aggregate-only observation (AO):* Each method observes only the final aggregate score (and not the individual-level responses or individual identities),

(ii) *method-blind evaluation (MB):* For each method, the distribution of the returned score depends only on what

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

was submitted and not on which training/optimization procedure produced it.

We show that these two conditions are jointly *necessary and sufficient* for swapping human evaluation for persona evaluation to be equivalent, on the method interface, to changing only the evaluation population, e.g., changing from the New York population to the Jakarta population. This zero-error statement is the special case of a robustness result: if method-blindness holds only up to ε per round, then the adaptive transcript law is within $T\varepsilon$ total variation of a literal-panel-change representation through horizon T. Conversely, if method identity or provenance changes the first-round feedback law by η, no artifact-only panel-change representation can approximate all methods better than $\eta/2$, and repeated provenance-sensitive rounds can amplify this lower bound.

In addition to this identification result, we extend the discussion to *usefulness*. Once (AO)+(MB) makes persona benchmarking identification-valid, the remaining question is whether the induced aggregate channel $Q_{\mathrm{pers}}(\cdot \mid w)$ is informative enough to distinguish and optimize meaningfully different methods. We formalize this through a discriminability parameter and show that it yields an explicit sample-complexity rule for the number of independent persona evaluations required for reliable method comparison. In this sense, beyond enforcing (AO)+(MB), "persona quality" becomes a measurable budget question.

The rest of the paper is organized as follows. Section A reviews related work. Section 2 formalizes the benchmarking setup and states the (AO) and (MB) conditions. Section 3 gives the identification and robustness discussion, proving that persona vs. human evaluation is indistinguishable on the method interface from a literal panel change if and only if (AO) and (MB) hold in the zero-error case, and quantifying approximate violations. Section 4 extends the identification discussion to a usefulness discussion that relates to sample complexity, and Section 5 gives two empirical illustrations. Due to page limitations, we defer the discussion of related works to Appendix A.

2. Setup: Benchmarking experimentation

2.1. Configurations and artifacts.

Let Θ denote the space of tunable configuration variables; this includes all controllable degrees of freedom that specify a system or a procedure (e.g., model weights, prompts/context, hyperparameters, decoding rules, tool policies, memory policies, data curation choices, or post-processing rules). A single candidate *configuration* is a choice $\theta \in \Theta$. A submitted *artifact* is what you hand to the benchmark, i.e., the externally visible object that the benchmark evaluates. We model this via an *artifact map* $g : \Theta \to \mathcal{W}$ and write

$$w = g(\theta) \in \mathcal{W}.$$

Possible choice of the artifact space $\mathcal{W}$ encompass: (i) a single output for a fixed input, (ii) an output distribution (stochastic method), (iii) a full interaction method mapping contexts to actions/outputs, (iv) a rollout distribution of an agent interacting with tools or environments.

Method as a configuration optimizer. We model method benchmarking as a "submit–observe" loop. A *method* (algorithm) may be either *non-adaptive* (a single-shot submission, or a fixed distribution over submissions) or *adaptive* (updating submissions based on past benchmark feedback). Concretely, at each round $t = 1, 2, \ldots, T$, the method chooses a configuration $\theta_t \in \Theta$ (equivalently an artifact $w_t := g(\theta_t) \in \mathcal{W}$), submits w_t to an *evaluator*, and receives a feedback observation o_t taking values in some observation space $\mathcal{O}$. Formally, a method (or method) $\mathcal{A}$ is modeled as a (possibly randomized) procedure that induces a decision kernel π_t over configurations at each round t, such that

$$\theta_t \sim \pi_t(\cdot \mid H_{t-1}, S),$$

where H_{t-1} is the method's observable history before round t, i.e.,

$$H_{t-1} := \{(\theta_\tau, o_\tau)\}_{\tau=1}^{t-1}$$

and S denotes any *side information* available before benchmarking begins (e.g., offline datasets, pretrained weights, logs, simulators).

Note that such a definition of a method contains:

- **Offline alignment** (e.g., DPO (Rafailov et al., 2023), SFT (Wei et al., 2022)) as the special case where π_t does not depend on H_{t-1} (or where $T = 1$).
- **Online alignment** (e.g., RLHF (Ouyang et al., 2022)) as the case where π_t adapts to past feedback $o_{1:t-1}$.
- **Hyperparameter tuning / AutoML / architecture search** (He et al., 2021; Ren et al., 2021) as black-box optimization over Θ using benchmark scores, where θ_t encodes a full training-and-deployment recipe, and π_t implements a sequential search procedure.
- **Prompt and system configuration search** (Pryzant et al., 2023; Kang & Yoganarasimhan, 2025) as the case where Θ indexes prompts, system messages, tool-use policies, retrieval and memory settings, and post-processing rules.
- **Data-centric training recipe search** (Zha et al., 2025) as the case where Θ includes dataset construction and curation choices (filtering, reweighting, mixing, synthetic-data generation policies).

2.2. Evaluation as panel × instrument × aggregation

We model an evaluation as a two-stage process: many *micro-level judgments* are first produced, and these are then compressed into a single *aggregate feedback signal* that the method actually observes. Under this model, an *evaluation setup* (or simply an *evaluator*) is fully specified as

$$(P, I, \Gamma, m),$$

which we will treat as a primitive object throughout.

Panel (P). A *panel* is a population of evaluators, either humans or LLM personas. Formally, let $\mathcal{P}$ denote the panel space and let P be a distribution over $\mathcal{P}$. Each evaluation call draws a fresh micro-panel of m independent evaluators

$$p_1, \ldots, p_m \overset{\text{i.i.d.}}{\sim} P.$$

Micro-instrument (I). Given an artifact $w \in \mathcal{W}$ and a panel member $p \in \mathcal{P}$, the *micro-instrument* produces an individual response. Formally, this is a conditional distribution

$$I(\cdot \mid w, p) \quad \text{over a micro-response space } \mathcal{Z}. \tag{1}$$

where the micro-response space $\mathcal{Z}$ can be a Likert score, a binary preference, or a short textual judgment. Each evaluator responds independently:

$$Z_\ell \sim I(\cdot \mid w, p_\ell), \qquad \ell = 1, \ldots, m.$$

Aggregation into observed feedback (Γ, m). Finally, the benchmark aggregates the m micro-responses into a single observable output. This is captured by a deterministic aggregation map

$$\Gamma : \mathcal{Z}^m \to \mathcal{O}, \tag{2}$$

where $\mathcal{O}$ is the feedback observation space. Typical examples include the mean score, a majority vote, or a pass/fail indicator. Putting these pieces together, a single evaluation call on artifact w returns the aggregate

$$o \;=\; \Gamma(Z_1, \ldots, Z_m) \in \mathcal{O}. \tag{3}$$

What's observed by the method. Although evaluation involves panel members and micro-responses, the method never observes them. What it sees is only the induced distribution of the aggregate feedback $o \in \mathcal{O}$. The tuple (P, I, Γ, m) therefore defines a Markov kernel on $\mathcal{O}$:

$$Q_{P,I}(A \mid w) \;:=\; \mathbb{P}\big(\Gamma(Z_1, \ldots, Z_m) \in A \,\big|\, w\big), \quad A \subseteq \mathcal{O} \text{ measurable.} \tag{4}$$

Intuitively, $Q_{P,I}(\cdot \mid w)$ is the distribution of the single observable feedback produced by the whole pipeline "sample panel → elicit micro-responses → aggregate" when the submitted artifact is w. All identification arguments in the sequel are necessarily about this reduced-form object.

2.3. Persona benchmark vs. human benchmark

Up to this point, we have described an *evaluation setup* abstractly as a tuple (P, I, Γ, m). Now we instantiate this abstraction in the two cases we want to compare: evaluation by humans versus evaluation by LLM personas.

Human benchmark. In a *human benchmark*, the panel distribution P_{hum} samples human evaluators, and the micro-instrument I_{hum} is the procedure that elicits a micro-response from a human (e.g., a rating, a preference, or a short written judgment). Together with the same aggregation map Γ and micro-panel size m, this induces an observable feedback kernel

$$Q_{\text{hum}}(\cdot \mid w).$$

Persona benchmark. In a *persona benchmark*, the panel distribution P_{pers} samples persona profiles (e.g., demographic or attitudinal descriptors), and the micro-instrument I_{pers} is implemented by an LLM judge conditioned on the sampled persona. Using the *same* aggregation Γ and micro-panel size m yields a second observable feedback kernel

$$Q_{\text{pers}}(\cdot \mid w).$$

What matters for the method. Although these two pipelines differ internally (humans vs. personas; human judgments vs. LLM judgments), the method only interacts with each benchmark through the induced distribution of the *aggregate* feedback. In other words, for the method, the relevant objects are precisely the two reduced-form kernels $Q_{\text{hum}}(\cdot \mid w)$ and $Q_{\text{pers}}(\cdot \mid w)$.

2.4. Key conditions

In asking the key question in this paper, "*when can we treat persona evaluation as a clean benchmark interface for comparing methods?*", we need to discuss two "benchmark hygiene" conditions that clarify what information the method does (and does not) get access to, and whether the benchmark behaves like a well-defined environment independent of who is playing.

Definition 1 (Aggregate-only observation (AO))**.** At each round t, the method observes only the aggregate feedback $o_t \in \mathcal{O}$. It does *not* observe the micro-level tuple $(p_1, \ldots, p_m, Z_1, \ldots, Z_m)$, any panel identifiers, or any additional side-channel information beyond o_t.

Intuitively, (AO) says the method sees exactly what a standard leaderboard would show: one score (or label) per submission. This prevents method's gaming behavior that relies on recognizing individual panelists/personas or exploiting micro-level structure that would be invisible in the intended benchmark interface.

Before stating the second condition, we define a probability measure $\mathbb{P}^{\mathcal{A}}$: for a fixed benchmark implementation, running a method $\mathcal{A}$ induces a probability measure $\mathbb{P}^{\mathcal{A}}$ over the interaction transcript $(w_1, o_1, w_2, o_2, \dots)$. Under (AO), we can define the method's interaction transcript after $t-1$ rounds as

$$\widetilde{H}_{t-1} := \{(w_\tau, o_\tau)\}_{\tau=1}^{t-1},$$

and its information before choosing w_t is the σ-field

$$\mathcal{I}_{t-1} := \sigma(S, \widetilde{H}_{t-1}).$$

Definition 2 (Method-blind evaluation (MB))**.** There exists a Markov kernel $Q(\cdot \mid w)$ on $\mathcal{O}$ such that for every method $\mathcal{A}$, every round t, and every measurable $A \subseteq \mathcal{O}$,

$$\mathbb{P}^{\mathcal{A}}\big(o_t \in A \,\big|\, \mathcal{I}_{t-1},\, w_t\big) = Q(A \mid w_t) \quad \text{a.s.}$$

Intuitively, (MB) is the minimal condition for calling this evaluation setup a *benchmark environment* at all: the evaluator should not care about the identity of the training procedure or other metadata, and care only about what was submitted (the artifact). In other words, the benchmark interaction is fully summarized by the reduced-form kernel $Q(\cdot \mid w)$, which is fixed across methods.

3. Identification: When Is Persona Benchmarking "Just Panel Change"?

From the method's point of view, each benchmark is a black box: it takes an artifact $w \in \mathcal{W}$ and returns a random aggregate feedback value $o \in \mathcal{O}$. All the internal structure (panel draws, micro-judgments, and aggregation) has already been compressed into the reduced-form kernels

$$Q_{\text{pers}}(\cdot \mid w) \quad \text{and} \quad Q_{\text{hum}}(\cdot \mid w),$$

defined in (4). In this section, we utilize this intuition to answer the following question:

> *When is swapping human evaluation for persona evaluation, as seen through the method's interface, nothing more than changing the panel P?*

We first formalize what it means to "only change the evaluation population" in the panel–instrument–aggregation model (a *literal panel change*), and then define the corresponding *interface-level* notion directly: the actual protocols must admit an auxiliary representation as a literal panel change on all adaptive transcript laws.

Throughout this section, the phrase *aggregate-only interface* means the information structure in which, at round t, the method's pre-decision information is

$$\mathcal{I}_{t-1} := \sigma(S, \widetilde{H}_{t-1}), \qquad \widetilde{H}_{t-1} := \{(w_\tau, o_\tau)\}_{\tau=1}^{t-1},$$

so the method observes only the aggregate feedback history. This is exactly the interface postulated by aggregate-only observation (AO). We first characterize panel change on this aggregate-only interface. We then promote the corresponding protocol-level notion of panel change to a definition and prove that it is equivalent to $(AO) + (MB)$.

This yields the identification result in the form we actually want: on the aggregate-only interface, panel-change representation is equivalent to method-blind evaluation (MB); therefore, for the actual human and persona protocols, the swap is indistinguishable from a literal panel change if and only if (AO) and (MB) both hold. We then give the quantitative version: approximate method-blindness implies approximate panel-change representation, with error growing at most linearly in the interaction horizon.

3.1. Literal panel-change representation

Definition 3 (Literal panel change)**.** Fix a panel space $\mathcal{P}$, micro-response space $\mathcal{Z}$, aggregation map $\Gamma : \mathcal{Z}^m \to \mathcal{O}$ and micro-panel size $m \in \mathbb{N}$. Let $I(\cdot \mid w, p)$ be a micro-instrument on $\mathcal{Z}$. For two panel distributions P, P' on $\mathcal{P}$, define two benchmarks

$$B := (P, I, \Gamma, m), \qquad B' := (P', I, \Gamma, m).$$

We say B' is obtained from B by a *literal panel change* if the only difference between them is that P is replaced by P' (i.e., I, Γ, m are identical).

Definition 3 captures the classical "same survey, different respondents" intuition: two benchmarks have the same instrument (the question/rubric and how responses are generated), the same aggregation rule, and the same micro-panel size, but sample from a different population (e.g., one from New York and one from Jakarta). That is, the protocol is unchanged except for the distribution over who evaluates.

Swapping humans for personas is not literally a panel change in this narrow sense because the micro-instrument is implemented differently (humans vs. an LLM judge conditioned on a persona). The relevant question is whether, after looking only at the method-facing interaction, the two actual protocols admit an equivalent representation in which the only changed object is the panel distribution.

Definition 4 (Interface-equivalence to a literal panel change)**.** Two actual benchmark protocols B and B' are *interface-equivalent to a literal panel change* if there exist two auxiliary benchmarks

$$\overline{B} = (\overline{P}, \overline{I}, \overline{\Gamma}, \overline{m}), \qquad \overline{B}' = (\overline{P}', \overline{I}, \overline{\Gamma}, \overline{m}),$$

which differ only in the panel distribution, such that for every adaptive method $\mathcal{A}$ and every horizon T,

$$\mathbb{P}^{\mathcal{A}}_{B,T} = \mathbb{P}^{\mathcal{A}}_{\overline{B},T}, \qquad \mathbb{P}^{\mathcal{A}}_{B',T} = \mathbb{P}^{\mathcal{A}}_{\overline{B}',T}.$$

Definition 4 does not define panel change by postulating artifact-only kernels. Here $\mathbb{P}^{\mathcal{A}}_{B,T}$ denotes the method-facing transcript law through horizon T. Side-channel observations, if exposed, are part of this method-facing transcript; because the auxiliary literal-panel-change benchmarks output only aggregate feedback, such side channels must be absent for the equality to hold. The definition asks for an operational representation: every adaptive method must see the same transcript law as it would have seen under a literal change of the evaluation population.

3.2. Panel change on the aggregate-only interface, and protocol-level characterization

Lemma 3.1 (Panel change on the aggregate-only interface $\iff$ (MB))**.** *Let B_{hum} and B_{pers} denote the human and persona benchmarks, both operated on the aggregate-only interface. The following are equivalent:*

1. ***(Panel change on the aggregate-only interface).*** *There exists a pair of benchmarks*

$$\overline{B} = (\overline{P}, \overline{I}, \overline{\Gamma}, \overline{m}), \qquad \overline{B}' = (\overline{P}', \overline{I}, \overline{\Gamma}, \overline{m}),$$

that differ by a literal panel change in the sense of Definition 3, such that for every method $\mathcal{A}$ and every horizon T,

$$\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{hum}}}(dw_{1:T}, do_{1:T}) = \mathbb{P}^{\mathcal{A}}_{\overline{B}}(dw_{1:T}, do_{1:T}),$$
$$\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{pers}}}(dw_{1:T}, do_{1:T}) = \mathbb{P}^{\mathcal{A}}_{\overline{B}'}(dw_{1:T}, do_{1:T}).$$

Equivalently, on the aggregate-only interface, swapping humans for personas is indistinguishable from an ordinary change of the evaluation population.

2. ***(MB).*** *Each benchmark is method-blind in the sense of Condition 2, with respect to the aggregate-only interface.*

Lemma 3.2 (Panel change on the actual method interface $\iff$ (AO)+(MB))**.** *Let B_{hum} and B_{pers} denote the actual human and persona benchmarking protocols, with common artifact space $\mathcal{W}$ and feedback space $\mathcal{O}$. The following are equivalent:*

1. ***(Panel change).*** *The human $\leftrightarrow$ persona swap is interface-equivalent to a literal panel change in the sense of Definition 4.*

2. ***(AO)+(MB).*** *The protocols satisfy aggregate-only observation (AO) and each benchmark is method-blind (MB).*

Definition 5 (ε-method-blindness)**.** Fix a benchmark B operated on the aggregate-only interface. We say B is ε-method-blind up to horizon T if there exists a Markov kernel $Q_B : \mathcal{W} \to \Delta(\mathcal{O})$ such that, for every method $\mathcal{A}$, every $t \le T$, and almost every realized $(\mathcal{I}_{t-1}, w_t)$,

$$d_{\mathrm{TV}}(\mathcal{L}_B(o_t \mid \mathcal{I}_{t-1}, w_t), Q_B(\cdot \mid w_t)) \le \varepsilon.$$

The following theorem is the main robustness result. It turns (MB) from an exact condition into an audit target: per-round deviations from artifact-only evaluation accumulate at most linearly in the adaptive horizon.

Theorem 3.3 (Robust panel-change representation)**.** *Suppose the actual human and persona benchmarks satisfy aggregate-only observation. Suppose further that the human benchmark is $\varepsilon_{\mathrm{hum}}$-method-blind with kernel Q_{hum}, and the persona benchmark is $\varepsilon_{\mathrm{pers}}$-method-blind with kernel Q_{pers}, up to horizon T. Then there exist two auxiliary benchmarks*

$$\overline{B}_{\mathrm{hum}} = (\delta_0, \overline{I}, \overline{\Gamma}, 1), \qquad \overline{B}_{\mathrm{pers}} = (\delta_1, \overline{I}, \overline{\Gamma}, 1),$$

which differ only in panel distribution, such that for every adaptive method $\mathcal{A}$,

$$d_{\mathrm{TV}}\left(\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{hum}},T}, \mathbb{P}^{\mathcal{A}}_{\overline{B}_{\mathrm{hum}},T}\right) \le T\varepsilon_{\mathrm{hum}},$$

and

$$d_{\mathrm{TV}}\left(\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{pers}},T}, \mathbb{P}^{\mathcal{A}}_{\overline{B}_{\mathrm{pers}},T}\right) \le T\varepsilon_{\mathrm{pers}}.$$

In particular, exact (AO)+(MB) is the special case $\varepsilon_{\mathrm{hum}} = \varepsilon_{\mathrm{pers}} = 0$.

The empirical sections below focus on discriminability κ_Q, not on estimating ε: their protocols enforce (MB) by design. Theorem 3.3 specifies what an additional provenance-sensitivity audit would need to bound if (MB) were only approximate.

Proposition 3.4 (Provenance sensitivity lower-bounds distance from panel change)**.** *Suppose there exist two one-shot methods $\mathcal{A}_0, \mathcal{A}_1$ that submit the same artifact w, but the benchmark returns first-round feedback laws R_0 and R_1 satisfying*

$$d_{\mathrm{TV}}(R_0, R_1) = \eta.$$

Then for every artifact-only kernel $Q(\cdot \mid w)$,

$$\max\{d_{\mathrm{TV}}(R_0, Q(\cdot \mid w)), d_{\mathrm{TV}}(R_1, Q(\cdot \mid w))\} \ge \eta/2.$$

Consequently, the benchmark is at least $\eta/2$-far from any artifact-only, panel-change representation even at horizon $T = 1$.

More generally, suppose two T-round methods submit the same artifact sequence, but the benchmark induces transcript laws $R_0^{(T)}$ and $R_1^{(T)}$ with

$$d_{\mathrm{TV}}(R_0^{(T)}, R_1^{(T)}) = \eta_T.$$

Then every artifact-only panel-change representation is at least $\eta_T/2$-far from one of these two method-facing transcript laws. In particular, if a one-round provenance-sensitive gap of size η repeats independently in the extremal form $R_1 = \delta_0$ and $R_0 = (1-\eta)\delta_0 + \eta\delta_1$, then $\eta_T = 1-(1-\eta)^T$, so the lower bound is $\frac{1}{2}[1-(1-\eta)^T]$, which is $\Omega(T\eta)$ for $T\eta \le 1$ up to constants.

Thus Lemma 3.2 is the zero-error identification case, while Theorem 3.3 and Proposition 3.4 give the stability and impossibility statements that make the protocol conditions operational.

Lemma 3.2 also highlights two distinct failure modes:

If (MB) fails, the benchmark is not a well-defined oracle environment. When (MB) fails, there is no single kernel $Q(\cdot \mid w)$ that governs the returned score across methods. Equivalently, the evaluator's behavior depends on provenance/identity or on interaction history in a way that is not summarized by the submitted artifact. In such a case, the benchmark cannot be treated as an artifact-only oracle channel, so "panel change" is not an identified description of the swap. See Appendix C.2 for when MB can be justified.

If (AO) fails, the method interface changes (even if aggregate scores look the same). If the protocol reveals micro-level information (panel identities, raw votes, ordering, etc.), then two evaluator implementations can induce the *same* aggregate kernel on $\mathcal{O}$ but still be distinguishable and exploitable by an adaptive method. Hence the observable interaction is not characterized by the aggregate channel alone. See Appendix C.1 for when AO fails.

4. Beyond validity: When does a persona panel constitute a *useful* benchmark?

As we discussed in Section 3, (AO)+(MB) characterizes exactly when the human $\leftrightarrow$ persona swap is indistinguishable on the method interface from a literal panel change. But validity, which is an identification argument, is not necessarily equivalent to *usefulness*. In this section, we formalize when and when not a persona-based LLM simulation can be a perfectly valid, but much less useful, benchmark compared to field experiments. We show that 1) dataset size is what really matters for a persona-based benchmark, or any benchmark, to be considered a useful benchmark, and 2) how to find such a required dataset size empirically.

As a starting point, recall that, under (AO)+(MB), the method interacts with the persona benchmark only through the reduced-form kernel

$$Q_{\mathrm{pers}}(\cdot \mid w) \quad \text{on} \quad \mathcal{O},$$

which returns an aggregate feedback draw $o \sim Q_{\mathrm{pers}}(\cdot \mid w)$ for each submitted artifact w. Thus, we arrive at an important observation:

Whether a persona dataset will constitute a useful benchmark is a question about how informative the induced channel Q_{pers} is for distinguishing and optimizing *methods.*

Throughout this section, we will utilize this idea to derive the minimum persona dataset size for benchmarking.

4.1. Usefulness and "mountain-fog" analogy

Let

$$\mathcal{W}_0 \subseteq \mathcal{W}$$

denote *the region of interest* for the artifacts (e.g., a tunable region of artifacts around a baseline artifact). For scalar feedback $\mathcal{O} = \mathbb{R}$, define the *benchmark landscape* and *benchmark noise* by

$$\mu_{\mathrm{pers}}(w) := \mathbb{E}[\, o \mid w \,], \quad \sigma^2_{\mathrm{pers}}(w) := \mathrm{Var}(o \mid w), \quad (5)$$

where $o \sim Q_{\mathrm{pers}}(\cdot \mid w)$.

A useful analogy to describe the usefulness of a benchmark is the "mountain-fog" metaphor. The (unknown) benchmark landscape $w \mapsto \mu_{\mathrm{pers}}(w)$ is the *mountain*: it assigns to each artifact the expected aggregate score returned by the benchmark. The benchmark noise scale $w \mapsto \sigma_{\mathrm{pers}}(w)$ is the *fog*. This metaphor separates two different reasons a benchmark may not be useful. A benchmark can be a *flat* mountain, meaning that $\mu_{\mathrm{pers}}(w)$ changes only slightly across meaningfully different artifacts; or it can be *noisy* (thick fog), meaning that $\sigma_{\mathrm{pers}}(w)$ is large relative to the score differences the method is trying to detect.

4.2. Formalization of usefulness: discriminability

We formalize the "mountain-fog" idea by introducing a concept we call *discriminability*. Let $D_{\mathrm{KL}}(\cdot,\cdot)$ be Kullback–Leibler divergence on $\mathcal{O}$. Also, suppose that we can define a metric $d_{\mathcal{W}}$ on $\mathcal{W}_0$.

Let

$$\mathcal{S}_r := \{(w, w') \in \mathcal{W}_0 \times \mathcal{W}_0 : d_{\mathcal{W}}(w, w') \ge r\}.$$

where $r > 0$ is a resolution parameter that formalizes "minimal meaningful change" in artifacts, which is often *prespecified ex ante*. The pair $(d_{\mathcal{W}}, r)$ should be read as: *"we only require the benchmark to separate artifacts that differ by at least r under $d_{\mathcal{W}}$."* Choosing a smaller r is a stricter requirement (it asks the benchmark to resolve finer changes), and it can only make discriminability harder (the infimum ranges over a larger set, so κ_Q can only decrease). Thus r should reflect the smallest change that is substantively meaningful *for method development*, not the smallest change that

can be expressed syntactically. In Section 4.4, we discuss the guidelines of figuring out $d_{\mathcal{W}}$ and r.

Fix a probability measure ν on $\mathcal{S}_r$. For $(w, w') \sim \nu$, define the random variable

$$U := D_{\mathrm{KL}}\left(Q_{\text{pers}}\left(\cdot \mid w\right) \| Q_{\text{pers}}\left(\cdot \mid w'\right)\right).$$

One conservative notion of discriminability is the quantity we call the q-quantile of U:

$$\text{Quantile}_q(U) := \inf\{u : \nu(U \le u) \ge q\} \tag{6}$$

Note that $\text{Quantile}_q(U)$ can be rewritten as

$$\sup_{\substack{E \subset \mathcal{S}_r \\ \nu(E) \ge 1-q}} \inf_{(w,w') \in E} D_{\mathrm{KL}}\left(Q(\cdot \mid w) \| Q\left(\cdot \mid w'\right)\right)$$

This implies that, for at least $1 - q$ of the r-separated pairs (under ν), the KL separation is at least $\text{Quantile}_q(U)$. If this is near zero, it means that distinct artifacts in $\mathcal{W}_0$ are essentially indistinguishable through the benchmark interface; therefore, the benchmark provides little usable feedback for methods.

In practice, however, the KL divergence is challenging to estimate empirically. To avoid overloading the micro-panel size m from the benchmark setup, let N_{eval} denote the number of independent evaluation units used to score one artifact in a comparison. The following Lemma 4.1 provides a useful simplification under a homoscedastic Gaussian reduced-form model[1], resolving this challenge.

Lemma 4.1. *Suppose $\mathcal{O} = \mathbb{R}$ and that an aggregate evaluation based on N_{eval} independent evaluation units induces a Gaussian homoscedastic reduced-form kernel:*

$$Q_{\text{pers}}(\cdot \mid w) = \mathcal{N}\big(\mu_{\text{pers}}(w),\, \sigma^2/N_{\text{eval}}\big), \quad \sigma > 0 \tag{7}$$

where N_{eval} is the number of independent evaluation units used to score one artifact (e.g., persona ratings or latent-call draws). Then for any $w, w' \in \mathcal{W}$,

$$D_{\mathrm{KL}}\big(Q_{\text{pers}}(\cdot \mid w) \,\|\, Q_{\text{pers}}(\cdot \mid w')\big) = \frac{\big(\mu_{\text{pers}}(w) - \mu_{\text{pers}}(w')\big)^2}{2\sigma^2/N_{\text{eval}}} = \frac{\Delta(w,w')^2}{2\sigma^2/N_{\text{eval}}}. \tag{8}$$

The right-hand side of Equation (8) is closely related to the quantity we often call the pairwise signal-to-noise (SNR), which is defined as

$$\mathrm{SNR}(w, w') := \frac{\Delta(w,w')^2}{2\sigma^2}$$

[1]The standard quantitative Berry-Esseen results give a finite-sample error bound that decays at rate $O(1/\sqrt{N_{\text{eval}}})$ for bounded summands, making this Gaussian approximation increasingly accurate as the number of independent evaluation units grows (Berry, 1941).

This quantity is *empirically estimable*: it depends on the mean difference and the per-unit variance scale σ^2, which can be estimated from unit-level responses or from aggregate variance after multiplying by N_{eval}.

4.3. Sample complexity.

Under the homoscedastic Gaussian model in Lemma 4.1, KL separation equals N_{eval} times the per-evaluation-unit SNR, so the KL and SNR notions of discriminability coincide up to this known sample-size factor; henceforth we work with the SNR version.

We define *per-evaluation-unit q-robust discriminability* of a benchmark as

$$\kappa_Q(q) := \sup_{\substack{E \subset \mathcal{S}_r \\ \nu(E) \ge 1-q}} \inf_{(w,w') \in E} \mathrm{SNR}(w, w') \tag{9}$$

$\kappa_Q(q)$ has a direct operational interpretation: it is the *per-evaluation-unit information* available to distinguish two artifacts that differ by at least r in the benchmark interface.

Lemma 4.2 (Pairwise comparison sample complexity from discriminability)**.** *Work under the same homoscedastic Gaussian reduced-form model as in Lemma 4.1. N_{eval} is the number of independent evaluation units used to score each artifact. Define $\Delta(w, w') := \mu_{\text{pers}}(w) - \mu_{\text{pers}}(w')$.*

Then for $(W, W') \sim \nu$,

$$\mathbb{P}\big(\widehat{\mu}_{pers}(W) \le \widehat{\mu}_{pers}(W'),\ \Delta(W, W') > 0\big) \le q + \exp\left(-\frac{N_{\text{eval}}}{2}\kappa_Q(q)\right).$$

In particular, choosing

$$N_{\text{eval}} \ \ge\ \frac{2}{\kappa_Q(q)} \log \frac{1}{\delta}$$

gives $\mathbb{P}(\widehat{\mu}_{pers}(W) \le \widehat{\mu}_{pers}(W'), \Delta\,(W, W') > 0) \le q + \delta$.

Section 5 gives two empirical illustrations of this sample-complexity analysis, with full details in Appendices D and E.

4.4. Choice of r and $d_{\mathcal{W}}$

The definition of discriminability in (9) depends on two user-specified design choices: a metric $d_{\mathcal{W}}$ on the artifact space $\mathcal{W}_0$ and a resolution threshold $r > 0$. These are generally *method- and task-specific* design parameters: different method families explore different degrees of freedom in $\mathcal{W}$ and therefore induce different natural notions of distance and resolution. Operationally, $d_{\mathcal{W}}$ and r determine which pairs of artifacts the benchmark is required to reliably distinguish, and therefore they determine the relevant sample complexity via Lemma 4.2.

Below are practical guidelines for selecting them in a way that is both interpretable and robust.

Tie $d_{\mathcal{W}}$ to the developer's degrees of freedom. A good default is to define $d_{\mathcal{W}}$ via the natural parameterization that methods actually tune. If artifacts are produced by knobs $\theta \in \Theta$ through $w(\theta)$, and there is a natural distance d_Θ on Θ, one can induce a pseudo-metric on $\mathcal{W}$ by

$$d_{\mathcal{W}}(w(\theta), w(\theta')) := d_\Theta(\theta, \theta').$$

This makes r interpretable as a *step size in the space the method explores*. Examples include the scaled Euclidean distance on continuous hyperparameters or the edit distance on a structured prompt template.

Choose r as a *minimal meaningful iteration* unit. In most benchmarking use cases, there is a natural notion of the smallest "iteration" a developer expects to be worth distinguishing. The guiding principle is that r should be *large enough* that changes below r are not worth spending benchmark budget on, but *small enough* that improvements developers actually seek fall above r.

- Prompt/instruction tuning: r can be "one allowed edit" under a pre-specified edit set (add/remove one constraint, add one example, modify one rubric item). Under token-level edit distance, this corresponds to a small fixed number of edits.
- Hyperparameter tuning: choose a scaled metric so that a standard "one-step" change has size ≈ 1, then set $r = 1$. For instance, scale each coordinate by a typical tuning increment.
- Model or policy variants: set r to the smallest recipe change you would treat as a distinct method (e.g., one additional fine-tuning epoch, one dataset mixture adjustment above a threshold, a decoding rule change).

In short, $d_{\mathcal{W}}$ should encode *meaningful artifact differences* (preferably behavioral and invariant to cosmetic changes), while r should encode the smallest change that developers intend to reliably resolve. With these choices fixed, κ_Q becomes an operationally estimable quantity, and Lemma 4.2 translates it directly into the required persona data size for stable method comparison.

5. Experiments

5.1. Ad Effectiveness Prompt Optimization

The first experiment applies the calculation to the persona-simulation ad benchmark used to evaluate TEXTBO, a prompt-optimization based self-improving AI system (Kang & Yoganarasimhan, 2025). An artifact is a scenario, prompt, and generated ad image; the benchmark aggregates persona-conditioned effectiveness ratings. We estimate $\widehat{\kappa}_Q(q)$ for one-step prompt improvements and report the implied number of persona evaluations required to distinguish improved prompts at a chosen confidence level. The calculation implies that $N_{\text{eval}}^{\text{req}}$ ranges from 46 to 1303 and averages 469.25 across the eight scenarios, compared with the original $m = 200$ micro-panel. Appendix D gives the full experimental calculation and explains how (AO) and (MB) are enforced in this setup.

5.2. Sales-Conversation Stopping as an Alignment Benchmark

The second experiment extends the same interface view to a sales-conversation stopping alignment benchmark adapted from Manzoor et al. (2025). The policy must align its stopping decisions with the benchmark objective: conserve call time by quitting when conversion is unlikely, while not prematurely ending calls that would have produced a sale. An artifact is a silent `wait`/`quit` policy that observes transcript prefixes but does not control the dialogue. We generate held-out persona-conditioned latent calls, evaluate policies by hidden replay, and compute $\widehat{\kappa}_Q(q)$ and the corresponding required latent-call draw count for the stated comparison distribution over stopping-policy baselines. In the overall $\{B_1, B_2\}$ evaluation slice, the calculation gives $N_{\text{eval}}^{\text{req}} = 17{,}063$, exceeding the 1,029-buyer held-out panel and making the sample-size constraint visible. Appendix E gives the full benchmark construction, the (AO)/(MB) protocol check, and the calculation.

6. Conclusion

We characterized when LLM-persona panels can substitute for human field experiments as a *benchmark interface* for method development. In the zero-error case, the human–persona swap is indistinguishable on the method's interface from a literal panel change if and only if the protocols satisfy *aggregate-only observation* (AO) and *method-blind evaluation* (MB); the robustness theorem makes this operational by giving $T\varepsilon$-approximate panel-change representation under ε-method-blindness, while provenance sensitivity of size η produces an $\eta/2$ lower bound against any artifact-only representation, amplified to $\Omega(T\eta)$ by repeated rounds. Interface validity, however, is distinct from both population transfer (whether rankings carry to a target human population) and usefulness (whether the channel is informative enough to act on); for the latter, our discriminability κ_Q at resolution r yields a budget on the order of $\kappa_Q^{-1}(q)\log(1/\delta)$ independent evaluation units for reliable comparison, so beyond enforcing (AO)+(MB) the practical question becomes a measurable sample-size requirement.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Athey, S. and Imbens, G. W. The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, volume 1, pp. 73–140. North-Holland, 2017. doi: 10.1016/BS.HEFE.2016.10.003.

Berry, A. C. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

Blum, A. and Hardt, M. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1006–1014. PMLR, 2015. URL https://proceedings.mlr.press/v37/blum15.html.

Bogert, E., Lauharatanahirun, N., and Schecter, A. Human preferences toward algorithmic advice in a word association task. *Scientific Reports*, 12(1):14501, 2022. doi: 10.1038/s41598-022-18638-2.

Bommasani, R., Liang, P., and Lee, T. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

Buckley, C. and Voorhees, E. M. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, New York, NY, USA, 2000. ACM. doi: 10.1145/345508.345543.

Carterette, B. Robust test collections for retrieval evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–62, New York, NY, USA, 2007. ACM. doi: 10.1145/1277741.1277754.

Carterette, B., Allan, J., and Sitaraman, R. K. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 268–275, New York, NY, USA, 2006. ACM. doi: 10.1145/1148170.1148219.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. Chatbot arena: An open platform for evaluating LLMs by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, New York, 1972.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

de Quidt, J., Haushofer, J., and Roth, C. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302, November 2018a. doi: 10.1257/aer.20171330.

de Quidt, J., Haushofer, J., and Roth, C. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302, 2018b. doi: 10.1257/aer.20171330.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 123–132, New York, NY, USA, 2013. ACM. doi: 10.1145/2433396.2433413.

Diamond, J. M. Overview: Laboratory experiments, field experiments, and natural experiments. In Diamond, J. M. and Case, T. J. (eds.), *Community Ecology*, pp. 3–22. Harper & Row, New York, 1986.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023. doi: 10.1016/j.tics.2023.04.008.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. Questioning the survey responses of large language models, 2024. URL https://arxiv.org/abs/2306.07951. NeurIPS 2024.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 117–126, New York, NY, USA, 2015. ACM. doi: 10.1145/2746539.2746580.

Feldman, V., Frostig, R., and Hardt, M. The advantages of multiple classes for reducing overfitting from test set reuse. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1892–1900. PMLR, 2019. URL https://proceedings.mlr.press/v97/feldman19a.html.

Gui, G. and Kim, S. Leveraging llms to improve experimental design: A generative stratification approach. *arXiv preprint arXiv:2509.25709*, 2025.

Gui, G. and Toubia, O. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023.

Harrison, G. W. and List, J. A. Field experiments. *Journal of Economic Literature*, 42(4):1009–1055, 2004. doi: 10.1257/0022051043004577.

He, X., Zhao, K., and Chu, X. Automl: A survey of the state-of-the-art. *Knowledge-based systems*, 212:106622, 2021.

Heineman, D., Hofmann, V., Magnusson, I., Gu, Y., Smith, N. A., Hajishirzi, H., Lo, K., and Dodge, J. Signal and noise: A framework for reducing uncertainty in language model evaluation, 2025. URL https://arxiv.org/abs/2508.13144.

Horton, J. J., Filippas, A., and Manning, B. S. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023. URL https://www.nber.org/papers/w31122.

Kang, E. H. and Yoganarasimhan, H. Textbo: Bayesian optimization in language space for eval-efficient self-improving ai. *arXiv preprint arXiv:2511.12063*, 2025.

Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009. doi: 10.1007/s10618-008-0114-1.

Kohavi, R., Tang, D., and Xu, Y. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, UK, 2020. doi: 10.1017/9781108653985.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Levitt, S. D. and List, J. A. Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments. *American Economic Journal: Applied Economics*, 3(1):224–238, January 2011. doi: 10.1257/app.3.1.224.

Li, A., Chen, H., Namkoong, H., and Peng, T. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025.

Logg, J. M., Minson, J. A., and Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019. doi: 10.1016/j.obhdp.2018.12.005.

Madaan, L., Singh, A. K., Schaeffer, R., Poulton, A., Koyejo, S., Stenetorp, P., Narang, S., and Hupkes, D. Quantifying variance in evaluation benchmarks, 2024. URL https://arxiv.org/abs/2406.10229.

Manzoor, E., Ascarza, E., and Netzer, O. Learning when to quit in sales conversations. *arXiv preprint arXiv:2511.01181*, 2025.

Miller, E. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Osborne, M. R. and Bailey, E. R. Me vs. the machine? subjective evaluations of human- and ai-generated advice. *Scientific Reports*, 15(1):3980, 2025. doi: 10.1038/s41598-025-86623-6.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022.

Paluck, E. L. and Cialdini, R. B. Field research methods. In Judd, C. M. and Reis, H. T. (eds.), *Handbook of Research Methods in Social and Personality Psychology*, pp. 81–98. Cambridge University Press, 2 edition, 2014. doi: 10.1017/CBO9780511996481.008.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, New York, NY, USA, 2023. ACM. doi: 10.1145/3586183.3606763.

Peng, T., Gui, G., Merlau, D. J., Fan, G. J., Sliman, M. B., Brucks, M., Johnson, E. J., Morwitz, V., Althenayyan, A., Bellezza, S., et al. A mega-study of digital twins reveals strengths, weaknesses and opportunities for further improvement. *arXiv preprint arXiv:2509.19088*, 2025.

Pryzant, R., Iter, D., Li, J., Lee, Y. T., Zhu, C., and Zeng, M. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741, 2023.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., and Wang, X. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4):1–34, 2021.

Samek, A. Advantages and disadvantages of field experiments. In *Handbook of research methods and applications in experimental economics*, pp. 104–120. Edward Elgar Publishing, 2019.

Shavelson, R. J. and Webb, N. M. *Generalizability Theory: A Primer*. Sage Publications, Newbury Park, CA, 1991.

Shrout, P. E. and Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86 (2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., and Chen, H. Database report: Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6): 1446–1455, 2025. doi: 10.1287/mksc.2025.0262.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., et al. Accelerating the machine learning lifecycle with mlflow. *IEEE Data Eng. Bull.*, 41 (4):39–45, 2018.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025. doi: 10.1145/3711118. Article 129.

A. Related works

We position our contribution relative to work on measurement reliability, sparse-judgment evaluation, field experimentation, adaptive benchmark design, provenance-sensitive evaluation, persona simulation, and benchmark usefulness.

Measurement reliability and sampled facets. Classical measurement theory already treats evaluator variation as a design object rather than as a single undifferentiated noise term. Cronbach et al. (1972) develop generalizability theory, showing that dependability depends on sampled facets such as raters, items, and occasions, and Shavelson & Webb (1991) provide a practical framework for choosing those facets to achieve target reliability. Shrout & Fleiss (1979) introduce the standard ICC taxonomy for multi-rater reliability, while Dawid & Skene (1979) model annotator-specific error rates when the latent truth is unobserved. This literature motivates our panel-centered view of evaluation, but it does not ask when replacing one evaluator population with another preserves the benchmark interface seen by an adaptive method.

Evaluation stability under sparse judgments. In information retrieval, Buckley & Voorhees (2000) study evaluation-measure stability and show that the number of queries can materially affect ranking reliability. Carterette et al. (2006) show that minimal test collections can still support confident ranking, and Carterette (2007) study how judgments can be made more reusable for evaluating new systems. These papers are close in spirit to our usefulness discussion, but they study stability under limited judgments rather than evaluator substitution.

Field experiments and online experimentation. Field experiments are the closest real-world analogue of a benchmark in many societal systems. Harrison & List (2004) characterize field experiments by how subjects, tasks, stakes, information, and context depart from the laboratory, while Athey & Imbens (2017) synthesize the main design and inference issues in randomized experiments. In online systems, Kohavi et al. (2009) survey controlled experimentation on the web, Deng et al. (2013) show that variance reduction can substantially improve experiment sensitivity, and Kohavi et al. (2020) consolidate the platform-level lessons for trustworthy A/B testing at scale. Our paper differs by asking when an LLM-persona benchmark can replace this experimental benchmark at the level of the method interface, rather than by matching treatment effects or raw score levels.

Adaptive benchmarking and information disclosure. A separate line of work emphasizes that what the benchmark reveals is itself part of the evaluation protocol. Blum & Hardt (2015) formalize leaderboard overfitting and propose a reliable mechanism for repeated submissions, Dwork et al. (2015) provide the broader theory of validity under adaptive data analysis, and Feldman et al. (2019) show that some multiclass settings are more robust to test-set reuse than binary ones. This literature is the closest precursor to our aggregate-only observation condition: once methods adapt to feedback, disclosure policy changes the optimization problem.

Blinding, provenance, and evaluator dependence. Another adjacent literature shows that identical artifacts can be judged differently when evaluators see source labels or other contextual cues. Logg et al. (2019) show that people often respond differently to algorithmic and human advice, de Quidt et al. (2018b) analyze experimenter-demand effects as a general threat to behavioral measurement, and Dominguez-Olmedo et al. (2024) show that survey-style elicitation of LLMs is highly sensitive to ordering and labeling artifacts. Large-scale evaluation platforms such as Chatbot Arena therefore rely on anonymous pairwise comparison to suppress provenance cues (Chiang et al., 2024). This literature motivates our method-blind evaluation condition, which asks the benchmark to depend on the submitted artifact rather than on method identity or side-channel context.

Persona simulation and digital twins. Recent work on persona simulation asks whether LLMs can reproduce human behavior well enough for social-scientific use. Argyle et al. (2023) show that conditioned language models can sometimes reproduce subgroup-level survey patterns, Park et al. (2023) extend the idea to interactive generative agents with persistent memory and social behavior, and Horton et al. (2023) study LLMs as simulated economic agents in canonical experimental settings. At the same time, Dillion et al. (2023) caution that LLMs are at best conditional proxies for human participants, and Toubia et al. (2025) contribute a large-scale dataset for building and auditing digital twins. Relative to this literature, we ask a narrower question: not whether personas perfectly mimic humans, but when swapping humans for personas changes only the evaluator population on the method-facing benchmark interface.

Adjacent LLM-substitution uses. Gui & Toubia (2023) study causal confounding in LLM simulations of human experimental subjects and propose unblinding prompts to make counterfactual scenarios unambiguous; this concerns faithful subject simulation, not evaluator replacement as a benchmark-interface property or a necessary-and-sufficient (AO)+(MB) characterization. Li et al. (2025) show that scalable LLM persona generation can produce biased, non-representative synthetic populations, so their focus is whether personas are good "silicon samples" for downstream analyses rather than whether evaluator-panel replacement preserves the artifact-to-feedback interface. Gui & Kim (2025) use LLM predictions for Generative Stratification before randomization, improving design efficiency while preserving unbiasedness through within-stratum randomization; the LLM affects experimental design, not the evaluation panel faced by an adaptive optimizer.

Benchmark usefulness, variance, and discriminability. A growing evaluation literature studies benchmark usefulness as a signal-to-noise problem rather than as a binary notion of validity. Madaan et al. (2024) quantify variance in benchmark outcomes and show that evaluation fluctuations can be large enough to change conclusions, while Heineman et al. (2025) explicitly frame benchmark selection in terms of signal and noise. Our contribution is complementary: once evaluator substitution is identification-valid, we introduce a benchmark-internal discriminability measure that yields the persona sample size required for reliable method comparison.

Taken together, these literatures study reliability, ranking stability, experimental sensitivity, adaptive benchmark design, provenance effects, and persona realism, but they do not characterize when evaluator substitution is interface-equivalent to panel change or how many persona evaluations are needed to make that substitution decision-relevant.

B. Deferred theoretical discussions

B.1. Auxiliary representation lemmas for Section 3

Throughout the proofs below, fix a method $\mathcal{A}$. Any side information S available before benchmarking begins is absorbed into $\mathcal{I}_{t-1}$; when kernels are written pointwise, we suppress it from the notation.

Lemma B.1 (Literal panel changes factor through artifact-only kernels)**.** *Under aggregate-only observation (AO), the method observes only $o_t \in \mathcal{O}$ each round. Let $B = (P, I, \Gamma, m)$ and $B' = (P', I, \Gamma, m)$ differ by a literal panel change in the sense of Definition 3. Define the reduced-form kernels*

$$Q_{P,I}(A \mid w) := \mathbb{P}\big(\Gamma(Z_1, \dots, Z_m) \in A \mid w\big),$$
$$Q_{P',I}(A \mid w) := \mathbb{P}\big(\Gamma(Z'_1, \dots, Z'_m) \in A \mid w\big).$$

Then for every method $\mathcal{A}$ and every horizon T, the aggregate transcript laws factor through $Q_{P,I}$ and $Q_{P',I}$.

Proof of Lemma B.1. Fix any method $\mathcal{A}$ with submission kernels $\pi_t(\cdot \mid \mathcal{I}_{t-1})$ under (AO), and fix a round t. Under benchmark $B = (P, I, \Gamma, m)$, conditional on the submitted artifact w_t, the benchmark generates

$$p_{t,1}, \dots, p_{t,m} \overset{i.i.d.}{\sim} P, \qquad Z_{t,\ell} \sim I(\cdot \mid w_t, p_{t,\ell}) \text{ independently over } \ell, \qquad o_t = \Gamma(Z_{t,1}, \dots, Z_{t,m}).$$

By construction, given w_t this sampling uses only fresh benchmark randomness (fresh panel draw and micro-responses) and therefore does not depend on $\mathcal{I}_{t-1}$ nor on the identity of $\mathcal{A}$. Hence for every measurable $A \subseteq \mathcal{O}$,

$$\mathbb{P}_B^{\mathcal{A}}(o_t \in A \mid \mathcal{I}_{t-1}, w_t) = Q_{P,I}(A \mid w_t) \quad \text{a.s.}$$

This is exactly the (MB)-type conditional independence statement with kernel $Q_{P,I}$. Combining this with the fact that $w_t \sim \pi_t(\cdot \mid \mathcal{I}_{t-1})$ under (AO), the standard sequential composition / chain rule for Markov kernels yields, for every horizon T,

$$\mathbb{P}_B^{\mathcal{A}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, Q_{P,I}(do_t \mid w_t).$$

The same argument for $B' = (P', I, \Gamma, m)$ gives

$$\mathbb{P}_{B'}^{\mathcal{A}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, Q_{P',I}(do_t \mid w_t).$$

□

Lemma B.2 (Artifact-only kernels admit a literal panel-change representation)**.** *Let $Q, Q' : \mathcal{W} \to \Delta(\mathcal{O})$ be two artifact-to-feedback kernels. There exists a pair of auxiliary benchmarks $\overline{B} = (\overline{P}, \overline{I}, \overline{\Gamma}, \overline{m})$ and $\overline{B}' = (\overline{P}', \overline{I}, \overline{\Gamma}, \overline{m})$ that differ only in panel distribution and whose reduced-form kernels are Q and Q', respectively.*

Proof of Lemma B.2. We construct an *auxiliary* panel–instrument–aggregation representation. This construction is not meant to mirror the internal structure of the original human/persona protocols. Let

$$\overline{\mathcal{P}} := \{0, 1\}, \qquad \overline{\mathcal{Z}} := \mathcal{O}, \quad \overline{m} := 1,$$
$$\overline{\Gamma} : \overline{\mathcal{Z}} \to \mathcal{O} \text{ be the identity map } \overline{\Gamma}(o) = o.$$

Define a *single* micro-instrument $\overline{I}$ on $\overline{\mathcal{Z}} = \mathcal{O}$ by

$$\overline{I}(\cdot \mid w, 0) := Q(\cdot \mid w), \qquad \overline{I}(\cdot \mid w, 1) := Q'(\cdot \mid w).$$

Now define the two panel distributions

$$\overline{P} := \delta_0, \qquad \overline{P}' := \delta_1.$$

Then $\overline{B} := (\overline{P}, \overline{I}, \overline{\Gamma}, \overline{m})$ and $\overline{B}' := (\overline{P}', \overline{I}, \overline{\Gamma}, \overline{m})$ differ *only* in the panel distribution (δ_0 versus δ_1), hence are a literal panel change.

Moreover, under $\overline{B}$, each benchmark call on w samples $p = 0$ a.s., then outputs $o \sim \overline{I}(\cdot \mid w, 0) = Q(\cdot \mid w)$. Thus the reduced-form kernel of $\overline{B}$ is exactly Q. Similarly, the reduced-form kernel of $\overline{B}'$ is exactly Q'. □

B.2. Proof of Lemma 3.1 and Lemma 3.2 in Section 3

Proof of Lemma 3.1. We prove (2) ⇒ (1) and (1) ⇒ (2).

(2) ⇒ (1)**: (MB) implies panel change on the aggregate-only interface.** On the aggregate-only interface, AO holds by construction. Suppose each benchmark is method-blind (MB) on this interface. Apply Lemma B.3 to the human benchmark B_{hum} to obtain a kernel

$$Q_{\text{hum}} : \mathcal{W} \to \Delta(\mathcal{O})$$

such that for every method $\mathcal{A}$ and horizon T,

$$\mathbb{P}^{\mathcal{A}}_{B_{\text{hum}}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, Q_{\text{hum}}(do_t \mid w_t).$$

Likewise apply Lemma B.3 to B_{pers} to obtain

$$Q_{\text{pers}} : \mathcal{W} \to \Delta(\mathcal{O})$$

with

$$\mathbb{P}^{\mathcal{A}}_{B_{\text{pers}}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, Q_{\text{pers}}(do_t \mid w_t).$$

Apply Lemma B.2 with $Q := Q_{\text{hum}}$ and $Q' := Q_{\text{pers}}$. It yields a pair of benchmarks

$$\overline{B} = (\overline{P}, \overline{I}, \overline{\Gamma}, \overline{m}), \qquad \overline{B}' = (\overline{P}', \overline{I}, \overline{\Gamma}, \overline{m})$$

that differ by a literal panel change. By Lemma B.1, their transcript laws have the same artifact-only factorizations as the displayed laws above, and therefore satisfy

$$\mathbb{P}^{\mathcal{A}}_{\overline{B}} = \mathbb{P}^{\mathcal{A}}_{B_{\text{hum}}}, \qquad \mathbb{P}^{\mathcal{A}}_{\overline{B}'} = \mathbb{P}^{\mathcal{A}}_{B_{\text{pers}}}$$

for every method $\mathcal{A}$ and every horizon T. This is exactly item (1).

(1) ⇒ (2)**: panel change on the aggregate-only interface implies (MB).** Suppose item (1) holds. Then there exists a pair of benchmarks

$$\overline{B} = (\overline{P}, \overline{I}, \overline{\Gamma}, \overline{m}), \qquad \overline{B}' = (\overline{P}', \overline{I}, \overline{\Gamma}, \overline{m}),$$

that differ by a literal panel change and such that for every method $\mathcal{A}$ and every horizon T,

$$\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{hum}}} = \mathbb{P}^{\mathcal{A}}_{\overline{B}}, \qquad \mathbb{P}^{\mathcal{A}}_{B_{\mathrm{pers}}} = \mathbb{P}^{\mathcal{A}}_{\overline{B}'}.$$

Because the benchmarks are being operated on the aggregate-only interface, AO holds for $\overline{B}$ and $\overline{B}'$ on that interface. Therefore Lemma B.1 applies and yields kernels

$$\overline{Q}, \overline{Q}' : \mathcal{W} \to \Delta(\mathcal{O})$$

such that for every method $\mathcal{A}$ and every horizon T,

$$\mathbb{P}^{\mathcal{A}}_{\overline{B}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, \overline{Q}(do_t \mid w_t),$$

and

$$\mathbb{P}^{\mathcal{A}}_{\overline{B}'}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, \overline{Q}'(do_t \mid w_t).$$

Using the equalities in item (1), the same factorizations hold for B_{hum} and B_{pers}:

$$\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{hum}}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, \overline{Q}(do_t \mid w_t),$$

and

$$\mathbb{P}^{\mathcal{A}}_{B_{\mathrm{pers}}}(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, \overline{Q}'(do_t \mid w_t).$$

Apply Lemma B.4 to the two benchmarks. It follows that B_{hum} satisfies (MB) with kernel $\overline{Q}$ and B_{pers} satisfies (MB) with kernel $\overline{Q}'$. This proves item (2). □

Proof of Lemma 3.2. We prove (1) ⇒ (2) and (2) ⇒ (1).

(1) ⇒ (2)**: panel change on the actual method interface implies (AO)+(MB).** Suppose the swap is interface-equivalent to a literal panel change in the sense of Definition 4. The auxiliary literal-panel-change benchmarks expose only the aggregate output o_t each round; equality of the method-facing transcript laws to those auxiliary benchmarks therefore implies aggregate-only observation (AO) for the actual protocols. The same definition, now on the aggregate-only interface, implies panel change on the aggregate-only interface. Applying Lemma 3.1 yields (MB) for both benchmarks. Therefore (AO)+(MB) hold.

(2) ⇒ (1)**: (AO)+(MB) imply panel change on the actual method interface.** Suppose the actual protocols satisfy (AO) and (MB). Because AO holds, the actual observable interface is exactly the aggregate-only interface. Applying Lemma 3.1 yields panel change on that aggregate-only interface. This is exactly interface-equivalence to a literal panel change in the sense of Definition 4. □

Proof of Theorem 3.3. Construct the auxiliary benchmarks using the same dummy-panel representation as in Lemma B.2: let $\overline{\mathcal{P}} = \{0, 1\}$, $\overline{\mathcal{Z}} = \mathcal{O}$, $\overline{m} = 1$, let $\overline{\Gamma}$ be the identity map, and define

$$\overline{I}(\cdot \mid w, 0) := Q_{\mathrm{hum}}(\cdot \mid w), \qquad \overline{I}(\cdot \mid w, 1) := Q_{\mathrm{pers}}(\cdot \mid w).$$

Then $\overline{B}_{\mathrm{hum}} = (\delta_0, \overline{I}, \overline{\Gamma}, 1)$ and $\overline{B}_{\mathrm{pers}} = (\delta_1, \overline{I}, \overline{\Gamma}, 1)$ differ only in the panel distribution.

Fix one of the two benchmarks, write its approximation error as ε and its artifact-only kernel as Q, and fix an adaptive method $\mathcal{A}$. Because (AO) holds, if the actual and auxiliary histories agree through round $t-1$, then $\mathcal{A}$ submits the same w_t in both processes. By ε-method-blindness and the maximal-coupling characterization of total variation, conditional on this common $(\mathcal{I}_{t-1}, w_t)$ we can couple the actual feedback draw with the auxiliary draw from $Q(\cdot \mid w_t)$ so that they disagree

with probability at most ε. Iterating this coupling over $t = 1, \dots, T$, the probability that the two length-T transcripts ever differ is at most $T\varepsilon$ by a union bound. Total variation is bounded by the disagreement probability under any coupling, so

$$d_{\mathrm{TV}}\left(\mathbb{P}^{\mathcal{A}}_{B,T}, \mathbb{P}^{\mathcal{A}}_{\overline{B},T}\right) \le T\varepsilon.$$

Applying this argument once with $B = B_{\mathrm{hum}}$, $Q = Q_{\mathrm{hum}}$, $\varepsilon = \varepsilon_{\mathrm{hum}}$, and once with $B = B_{\mathrm{pers}}$, $Q = Q_{\mathrm{pers}}$, $\varepsilon = \varepsilon_{\mathrm{pers}}$, gives the two displayed bounds. □

Proof of Proposition 3.4. For any artifact-only kernel $Q(\cdot \mid w)$, the triangle inequality for total variation gives

$$\eta = d_{\mathrm{TV}}(R_0, R_1) \le d_{\mathrm{TV}}(R_0, Q(\cdot \mid w)) + d_{\mathrm{TV}}(Q(\cdot \mid w), R_1).$$

Therefore at least one of the two terms on the right is at least $\eta/2$, which proves the displayed lower bound. Any artifact-only panel-change representation would assign a single first-round feedback law to artifact w, independent of which one-shot method submitted it, so the same lower bound applies to every such representation at horizon $T = 1$.

The T-round statement is the same triangle-inequality argument applied to full transcript laws: for any artifact-only panel-change transcript law $Q^{(T)}$,

$$\eta_T = d_{\mathrm{TV}}(R_0^{(T)}, R_1^{(T)}) \le d_{\mathrm{TV}}(R_0^{(T)}, Q^{(T)}) + d_{\mathrm{TV}}(Q^{(T)}, R_1^{(T)}),$$

so one of the two approximation errors is at least $\eta_T/2$. For the repeated extremal example, $R_1^{(T)} = \delta_0^{\otimes T}$, while $R_0^{(T)}$ assigns probability $(1-\eta)^T$ to the all-zero transcript. Hence

$$d_{\mathrm{TV}}(R_0^{(T)}, R_1^{(T)}) = 1 - (1-\eta)^T.$$

When $T\eta \le 1$, $1 - (1-\eta)^T \ge 1 - e^{-T\eta} \ge T\eta/2$, giving a lower bound of at least $T\eta/4$. □

Lemma B.3 (Transcript factorization under (AO)+(MB))**.** *Fix a benchmark B with artifact space $\mathcal{W}$ and feedback space $\mathcal{O}$. Suppose (AO) and (MB) hold for B, i.e., there exists a Markov kernel $Q_B : \mathcal{W} \to \Delta(\mathcal{O})$ such that for every method $\mathcal{A}$, every round t, and every measurable $A \subseteq \mathcal{O}$,*

$$\mathbb{P}^{\mathcal{A}}_B(o_t \in A \mid \mathcal{I}_{t-1}, w_t) = Q_B(A \mid w_t) \quad \text{a.s.} \tag{10}$$

Then for every method $\mathcal{A}$ (with submission kernels $\pi_t(\cdot \mid \mathcal{I}_{t-1})$ under (AO)) and every horizon T,

$$\mathbb{P}^{\mathcal{A}}_B(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, Q_B(do_t \mid w_t), \tag{11}$$

Proof of Lemma B.3. Fix a method $\mathcal{A}$. Under (AO), at each round t the method chooses

$$w_t \sim \pi_t(\cdot \mid \mathcal{I}_{t-1}).$$

By (MB) in (10), conditional on $(\mathcal{I}_{t-1}, w_t)$ the benchmark draw satisfies

$$o_t \sim Q_B(\cdot \mid w_t).$$

Therefore, conditional on $\mathcal{I}_{t-1}$, the pair (w_t, o_t) is generated by

$$w_t \sim \pi_t(\cdot \mid \mathcal{I}_{t-1}), \qquad o_t \sim Q_B(\cdot \mid w_t).$$

Iterating over $t = 1, \dots, T$ and applying the standard chain rule / sequential composition for Markov kernels yields Equation (11). □

Lemma B.4 (Factorization implies (MB))**.** *Fix a benchmark B. Suppose that on the aggregate-only interface, there exists a Markov kernel $Q_B : \mathcal{W} \to \Delta(\mathcal{O})$ such that for every method $\mathcal{A}$ and every horizon T,*

$$\mathbb{P}^{\mathcal{A}}_B(dw_{1:T}, do_{1:T}) = \prod_{t=1}^{T} \pi_t(dw_t \mid \mathcal{I}_{t-1})\, Q_B(do_t \mid w_t).$$

Then B satisfies (MB) in the sense of Condition 2; namely, for every method $\mathcal{A}$, every t, and measurable $A \subseteq \mathcal{O}$,

$$\mathbb{P}^{\mathcal{A}}_B(o_t \in A \mid \mathcal{I}_{t-1}, w_t) = Q_B(A \mid w_t) \quad \text{a.s.}$$

Proof of Lemma B.4. Fix $\mathcal{A}, t$ and a measurable $A \subseteq \mathcal{O}$. Let

$$U_{t-1} := (S, \widetilde{H}_{t-1}),$$

so that $\mathcal{I}_{t-1} = \sigma(U_{t-1})$ on the aggregate-only interface. Apply the displayed factorization with horizon $T = t$, and let R_{t-1} denote the marginal law of U_{t-1}. Let φ be any bounded nonnegative measurable function of (U_{t-1}, w_t). Writing a realization of U_{t-1} as u_{t-1}, we have

$$\begin{aligned}
\mathbb{E}_B^{\mathcal{A}}\big[\varphi(U_{t-1}, w_t)\, \mathbf{1}_{\{o_t \in A\}}\big] &= \int \varphi(u, w)\, \mathbf{1}_A(o)\, R_{t-1}(du)\, \pi_t(dw \mid u)\, Q_B(do \mid w) \\
&= \int \varphi(u, w)\, Q_B(A \mid w)\, R_{t-1}(du)\, \pi_t(dw \mid u) \\
&= \mathbb{E}_B^{\mathcal{A}}[\varphi(U_{t-1}, w_t)\, Q_B(A \mid w_t)]\,.
\end{aligned}$$

Therefore $Q_B(A \mid w_t)$ is a version of the conditional expectation

$$\mathbb{E}_B^{\mathcal{A}}\big[\mathbf{1}_{\{o_t \in A\}} \mid U_{t-1}, w_t\big]\,.$$

Equivalently,

$$\mathbb{P}_B^{\mathcal{A}}(o_t \in A \mid \mathcal{I}_{t-1}, w_t) = Q_B(A \mid w_t) \quad \text{a.s.}$$

This is exactly (MB). □

B.3. Proofs in Section 4

Proof of Lemma 4.1. Let $v := \sigma^2/N_{\text{eval}}$, and let p_x denote the density of $\mathcal{N}(x, v)$:

$$p_x(o) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(o-x)^2}{2v}\right).$$

By definition,

$$d(x, y) = \mathbb{E}_{O \sim \mathcal{N}(x,v)}\left[\log \frac{p_x(O)}{p_y(O)}\right].$$

Compute the log-likelihood ratio:

$$\begin{aligned}
\log \frac{p_x(O)}{p_y(O)} &= -\frac{(O-x)^2}{2v} + \frac{(O-y)^2}{2v} \\
&= \frac{(O-y)^2 - (O-x)^2}{2v} \\
&= \frac{(x-y)\big(2O - x - y\big)}{2v}.
\end{aligned}$$

Taking expectation under $O \sim \mathcal{N}(x, v)$ yields

$$\begin{aligned}
d(x, y) &= \frac{x-y}{2v}\, \mathbb{E}\big[2O - x - y\big] \\
&= \frac{x-y}{2v}\, (2x - x - y) = \frac{(x-y)^2}{2v} = \frac{(x-y)^2}{2\sigma^2/N_{\text{eval}}},
\end{aligned}$$

□

Proof of Lemma 4.2. Fix any $(w, w') \in \mathcal{S}_r$ and abbreviate $\Delta := \Delta(w, w') = \mu_{\text{pers}}(w) - \mu_{\text{pers}}(w')$. Under the Gaussian reduced-form model in Lemma 4.1, an aggregate evaluation using N_{eval} independent evaluation units returns

$$\widehat{\mu}_{\text{pers}}(w) \sim \mathcal{N}\left(\mu_{\text{pers}}(w), \frac{\sigma^2}{N_{\text{eval}}}\right), \qquad \widehat{\mu}_{\text{pers}}(w') \sim \mathcal{N}\left(\mu_{\text{pers}}(w'), \frac{\sigma^2}{N_{\text{eval}}}\right),$$

independently (given w, w'). Hence

$$\widehat{\Delta} \;:=\; \widehat{\mu}_{\mathrm{pers}}(w) - \widehat{\mu}_{\mathrm{pers}}(w') \sim \mathcal{N}\!\left(\Delta, \frac{2\sigma^2}{N_{\mathrm{eval}}}\right).$$

If $\Delta \le 0$, then $\mathbb{P}(\widehat{\mu}_{\mathrm{pers}}(w) \le \widehat{\mu}_{\mathrm{pers}}(w'),\, \Delta > 0 \mid w, w') = 0$. If $\Delta > 0$, then the event $\{\widehat{\mu}_{\mathrm{pers}}(w) \le \widehat{\mu}_{\mathrm{pers}}(w')\}$ is $\{\widehat{\Delta} \le 0\}$. For any $t > 0$, Markov's inequality gives

$$\mathbb{P}(\widehat{\Delta} \le 0) = \mathbb{P}\Big(e^{-t\widehat{\Delta}} \ge 1\Big) \le \mathbb{E}\Big[e^{-t\widehat{\Delta}}\Big] = \exp\!\left(-t\Delta + \frac{t^2}{2}\cdot\frac{2\sigma^2}{N_{\mathrm{eval}}}\right) = \exp\!\left(-t\Delta + \frac{t^2\sigma^2}{N_{\mathrm{eval}}}\right).$$

Optimizing over t yields $t^\star = \frac{N_{\mathrm{eval}}\Delta}{2\sigma^2}$, hence

$$\mathbb{P}(\widehat{\Delta} \le 0) \le \exp\!\left(-\frac{N_{\mathrm{eval}}\Delta^2}{4\sigma^2}\right) = \exp\!\left(-\frac{N_{\mathrm{eval}}}{2}\cdot\frac{\Delta^2}{2\sigma^2}\right).$$

Recalling $\mathrm{SNR}(w, w') := \frac{\Delta(w,w')^2}{2\sigma^2}$, we obtain the conditional bound

$$\mathbb{P}(\widehat{\mu}_{\mathrm{pers}}(w) \le \widehat{\mu}_{\mathrm{pers}}(w'),\, \Delta(w, w') > 0 \,|\, w, w') \le \exp\!\left(-\frac{N_{\mathrm{eval}}}{2}\,\mathrm{SNR}(w, w')\right).$$

Now draw $(W, W') \sim \nu$ and let $V := \mathrm{SNR}(W, W')$. Taking expectations,

$$\mathbb{P}(\widehat{\mu}_{\mathrm{pers}}(W) \le \widehat{\mu}_{\mathrm{pers}}(W'),\, \Delta(W, W') > 0) \le \mathbb{E}\left[\exp\!\left(-\frac{N_{\mathrm{eval}}}{2}V\right)\right].$$

Let $\mathcal{B} := \{V < \kappa_Q(q)\}$ and $\mathcal{G} := \{V \ge \kappa_Q(q)\}$. By the definition of $\kappa_Q(q)$ as the q-quantile / q-robust discriminability of V (cf. (9) and its quantile-equivalent form), we have $\nu(\mathcal{B}) \le q$. Therefore,

$$\begin{aligned}\mathbb{E}\Big[e^{-\frac{N_{\mathrm{eval}}}{2}V}\Big] &= \mathbb{E}\Big[e^{-\frac{N_{\mathrm{eval}}}{2}V}\mathbf{1}_{\mathcal{B}}\Big] + \mathbb{E}\Big[e^{-\frac{N_{\mathrm{eval}}}{2}V}\mathbf{1}_{\mathcal{G}}\Big] \\ &\le \nu(\mathcal{B})\cdot 1 + e^{-\frac{N_{\mathrm{eval}}}{2}\kappa_Q(q)}\nu(\mathcal{G}) \;\le\; q + e^{-\frac{N_{\mathrm{eval}}}{2}\kappa_Q(q)}.\end{aligned}$$

This proves the claimed bound.

Finally, if $N_{\mathrm{eval}} \ge \frac{2}{\kappa_Q(q)}\log\frac{1}{\delta}$, then $\exp(-\frac{N_{\mathrm{eval}}}{2}\kappa_Q(q)) \le \delta$, so

$$\mathbb{P}(\widehat{\mu}_{\mathrm{pers}}(W) \le \widehat{\mu}_{\mathrm{pers}}(W'),\, \Delta(W, W') > 0) \le q + \delta.$$

□

C. Extended discussions

C.1. AO and panel change – a counterexample

The equivalence in Lemma 3.2 highlights that (AO) is not merely technical: if micro-level information leaks (raw votes, rater identities, ordering, etc.), then two evaluation pipelines can induce the *same* aggregate kernel $Q(\cdot \mid w)$ yet still be distinguishable (and exploitable) by an adaptive method. The following minimal construction makes this necessity direction concrete.

Proposition C.1 (Violating (AO) can break panel change on the actual method interface even when aggregate kernels match)**.** *There exist two benchmarks B and B' that induce the* same *reduced-form kernel $Q(\cdot \mid w)$ on the aggregate feedback space $\mathcal{O}$, but such that if the benchmark reveals the raw vote vector $(Z_1, \ldots, Z_m)$ to the method (violating (AO)), then there is an adaptive method $\mathcal{A}$ whose aggregate transcript law $\mathbb{P}^{\mathcal{A}}(w_{1:T}, o_{1:T})$ differs between B and B'. In the construction below, the total-variation distance between the two aggregate transcript laws is at least 0.24. Consequently, the swap cannot be panel change on the actual method interface in the sense of Definition 4.*

Proof. We specify two benchmarks and an adaptive method.

Spaces and aggregation. Let $\mathcal{W} = \{0,1\}$, $\mathcal{Z} = \{0,1\}$, $\mathcal{O} = \{0,1\}$, and take micro-panel size $m = 2$. Let the aggregation map be the XOR (disagreement) statistic

$$\Gamma(z_1, z_2) := z_1 \oplus z_2 \in \{0,1\}.$$

Thus the aggregate output is $o = 1$ iff the two individual votes disagree.

Two micro-instruments with identical aggregate kernels. Let $p_0 := 0.1$ and $p_1 := 0.4$. Define benchmark B so that, for each submitted artifact $w \in \{0,1\}$, the two micro-votes are independent Bernoulli draws

$$Z_1, Z_2 \overset{i.i.d.}{\sim} \text{Bernoulli}(p_w).$$

Define benchmark B' identically except that each vote is *flipped* in distribution:

$$Z'_1, Z'_2 \overset{i.i.d.}{\sim} \text{Bernoulli}(1 - p_w).$$

(Panel sampling is irrelevant here; one can take a degenerate panel distribution and absorb everything into the micro-instrument.)

Now compute the induced reduced-form kernels on $\mathcal{O}$. Under B,

$$\mathbb{P}_B(o = 1 \mid w) = \mathbb{P}_B(Z_1 \neq Z_2 \mid w) = 2p_w(1 - p_w).$$

Under B',

$$\mathbb{P}_{B'}(o = 1 \mid w) = \mathbb{P}_{B'}(Z'_1 \neq Z'_2 \mid w) = 2(1 - p_w)p_w = 2p_w(1 - p_w).$$

Hence the aggregate channels coincide exactly:

$$Q_B(\cdot \mid w) = Q_{B'}(\cdot \mid w) \qquad \forall\, w \in \mathcal{W}.$$

Concretely, $Q(o = 1 \mid 0) = 2(0.1)(0.9) = 0.18$ and $Q(o = 1 \mid 1) = 2(0.4)(0.6) = 0.48$ for both benchmarks.

AO violation (raw-vote leakage) and an adaptive distinguisher. Now suppose (AO) is violated and the benchmark releases the raw vote vector (Z_1, Z_2) (or (Z'_1, Z'_2)) to the method in addition to the aggregate o. Consider the following horizon-$T = 2$ adaptive method $\mathcal{A}$:

- Round 1: submit $w_1 = 0$.
- Observe the *first* raw vote and set $w_2 := Z_{1,1}$ (i.e., the round-1 vote of evaluator 1).

This is a valid adaptive strategy under the leaked interface, but it is *not* measurable with respect to the aggregate-only history (w_1, o_1).

Under benchmark B, $\mathbb{P}(w_2 = 1) = \mathbb{P}(Z_{1,1} = 1 \mid w_1 = 0) = p_0 = 0.1$. Under benchmark B', $\mathbb{P}(w_2 = 1) = \mathbb{P}(Z'_{1,1} = 1 \mid w_1 = 0) = 1 - p_0 = 0.9$. Therefore the distribution of the second-round submission w_2 differs between the two benchmarks even though the aggregate kernel $Q(\cdot \mid w)$ is identical.

Because $o_2 \sim Q(\cdot \mid w_2)$ in both benchmarks, this also induces a difference in the aggregate outcome at round 2:

$$\begin{aligned}\mathbb{P}_B(o_2 = 1) &= \mathbb{P}_B(w_2 = 1)\, Q(o = 1 \mid 1) + \mathbb{P}_B(w_2 = 0)\, Q(o = 1 \mid 0) \\ &= 0.1 \cdot 0.48 + 0.9 \cdot 0.18 = 0.21, \\ \mathbb{P}_{B'}(o_2 = 1) &= 0.9 \cdot 0.48 + 0.1 \cdot 0.18 = 0.45.\end{aligned}$$

Thus the aggregate transcript laws $\mathbb{P}_B^{\mathcal{A}}(w_{1:2}, o_{1:2})$ and $\mathbb{P}_{B'}^{\mathcal{A}}(w_{1:2}, o_{1:2})$ differ. Indeed, the event $\{o_2 = 1\}$ has probabilities 0.21 and 0.45, so

$$d_{\mathrm{TV}}\big(\mathbb{P}_B^{\mathcal{A}}(w_{1:2}, o_{1:2}), \mathbb{P}_{B'}^{\mathcal{A}}(w_{1:2}, o_{1:2})\big) \geq 0.45 - 0.21 = 0.24.$$

Conclusion. Panel change on the actual method interface, in the sense of Definition 4, requires aggregate-only observation. The present construction violates (AO) and, moreover, produces different aggregate transcript laws under an adaptive method. Therefore the swap cannot be panel change on the actual method interface. □

The same phenomenon occurs if the benchmark releases panel identities (or stable rater IDs): even if the aggregate score distribution $Q(\cdot \mid w)$ is unchanged, the extra identifier acts as a side channel that an adaptive method can condition on, producing different submission sequences and hence different aggregate transcripts. This is exactly why (AO) is a *benchmark-hygiene* requirement: it rules out side channels through which the method can distinguish evaluator implementations that are otherwise identical at the aggregate level.

C.2. Practical realism of method-blind evaluation (MB)

Condition 2 is a *protocol* requirement: conditional on the submitted artifact w, the distribution of the returned aggregate feedback must not depend on who submitted w or how it was produced. This is intentionally stronger than what one gets "by default" in many human-judgment settings, because there is extensive evidence that evaluations can shift when raters are exposed to provenance cues (e.g., explicit labels such as "AI-generated" vs. "human-written") or other contextual metadata beyond the content being judged (Bogert et al., 2022; Osborne & Bailey, 2025; Levitt & List, 2011; de Quidt et al., 2018a). In our framework, such provenance dependence is precisely an (MB) violation: two identical artifacts w can induce different score distributions if the evaluator observes additional information correlated with the producing method.[2]

How realistic is (MB) in practice? (MB) is best viewed as a *design target* that is often approximately achievable, but not automatic. It is most realistic in settings where the outcome is behavioral and passively recorded (classic online A/B tests), since users are typically not told which method produced what they see and outcomes like click-through or conversion are not direct subjective judgments (Kohavi et al., 2020). By contrast, in explicit rating / preference-judgment pipelines (crowd or expert), (MB) is fragile: even minimal provenance cues (labels, branding, prior beliefs about model quality, expectations about "humans vs. AI") can create systematic shifts in ratings for the *same* artifact, violating (MB). This fragility is also consistent with the motivation for anonymous pairwise evaluation interfaces used in large-scale leaderboards (Chiang et al., 2024) and with documented ordering/labeling effects in LLM-judge-style protocols (Dominguez-Olmedo et al., 2024).

The key takeaway is that (MB) should not be read as an empirical claim about human invariance; it is a *benchmark hygiene constraint* that must be enforced (or at least audited) by the benchmark implementation.

C.3. Extension to heteroscedastic Gaussian reduced-form kernels

Lemma 4.1 uses a homoscedastic Gaussian reduced-form kernel, i.e., $\mathrm{Var}(o \mid w)$ is constant across artifacts. In many benchmarks, however, the aggregate score variance depends on the submitted artifact: some artifacts elicit highly consistent

[2] If provenance cues are literally part of the submitted artifact (e.g., the text contains "as an AI language model"), then any resulting penalty is *not* an (MB) violation, because it is a function of w itself. The problematic case for (MB) is when the evaluator is shown extra metadata (model name, submitter identity, method label, timestamp-based context, etc.) that is *not* a function of w.

micro-responses (low noise), while others are ambiguous or polarizing (high noise). This motivates the more general *heteroscedastic* Gaussian model

$$Q_{\text{pers}}(\cdot \mid w) = \mathcal{N}\big(\mu_{\text{pers}}(w),\, \sigma^2_{\text{pers}}(w)\big), \qquad \sigma_{\text{pers}}(w) > 0. \tag{12}$$

(Here $\sigma^2_{\text{pers}}(w)$ denotes the variance of the *aggregate* feedback draw $o \sim Q_{\text{pers}}(\cdot \mid w)$. Any micro-level panel size has already been absorbed into this reduced-form variance.)

Lemma C.2 (KL divergence under heteroscedastic Gaussians)**.** *Suppose $\mathcal{O} = \mathbb{R}$ and (12). Then for any $w, w' \in \mathcal{W}$,*

$$\begin{aligned} &D_{\text{KL}}\big(Q_{\text{pers}}(\cdot \mid w) \,\big\|\, Q_{\text{pers}}(\cdot \mid w')\big) \\ &= \frac{1}{2}\left[\log \frac{\sigma^2_{\text{pers}}(w')}{\sigma^2_{\text{pers}}(w)} + \frac{\sigma^2_{\text{pers}}(w)}{\sigma^2_{\text{pers}}(w')} - 1\right] + \frac{\big(\mu_{\text{pers}}(w) - \mu_{\text{pers}}(w')\big)^2}{2\,\sigma^2_{\text{pers}}(w')}. \end{aligned} \tag{13}$$

Equation (13) highlights two separable sources of distributional distinguishability: (i) a *variance-mismatch* term (the bracketed expression), which is zero iff $\sigma^2_{\text{pers}}(w) = \sigma^2_{\text{pers}}(w')$, and (ii) a *mean-separation* term, which scales the squared mean gap by $1/\sigma^2_{\text{pers}}(w')$. In contrast to the homoscedastic case, the KL divergence is generally *asymmetric*: $D_{\text{KL}}(Q(\cdot \mid w)\|Q(\cdot \mid w')) \neq D_{\text{KL}}(Q(\cdot \mid w')\|Q(\cdot \mid w))$ when $\sigma^2_{\text{pers}}(w) \neq \sigma^2_{\text{pers}}(w')$.

Proof of Lemma C.2. Fix $w, w' \in \mathcal{W}$ and abbreviate

$$\mu := \mu_{\text{pers}}(w), \quad \mu' := \mu_{\text{pers}}(w'), \quad \sigma^2 := \sigma^2_{\text{pers}}(w), \quad \tau^2 := \sigma^2_{\text{pers}}(w').$$

Let $P := \mathcal{N}(\mu, \sigma^2)$ and $Q := \mathcal{N}(\mu', \tau^2)$ with Lebesgue densities

$$p(o) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(o-\mu)^2}{2\sigma^2}\right), \qquad q(o) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(o-\mu')^2}{2\tau^2}\right).$$

By definition,

$$D_{\text{KL}}(P\|Q) = \mathbb{E}_{O\sim P}\left[\log \frac{p(O)}{q(O)}\right].$$

Compute the log-likelihood ratio:

$$\begin{aligned} \log \frac{p(O)}{q(O)} &= \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(O-\mu)^2}{2\sigma^2}\right) - \left(-\frac{1}{2}\log(2\pi\tau^2) - \frac{(O-\mu')^2}{2\tau^2}\right) \\ &= \frac{1}{2}\log\frac{\tau^2}{\sigma^2} + \frac{(O-\mu')^2}{2\tau^2} - \frac{(O-\mu)^2}{2\sigma^2}. \end{aligned}$$

Taking expectations under $O \sim \mathcal{N}(\mu, \sigma^2)$, we use the identities

$$\mathbb{E}[(O-\mu)^2] = \sigma^2, \qquad \mathbb{E}[(O-\mu')^2] = \text{Var}(O) + (\mathbb{E}[O] - \mu')^2 = \sigma^2 + (\mu - \mu')^2.$$

Hence

$$\begin{aligned} D_{\text{KL}}(P\|Q) &= \frac{1}{2}\log\frac{\tau^2}{\sigma^2} + \frac{1}{2\tau^2}\Big(\sigma^2 + (\mu-\mu')^2\Big) - \frac{1}{2\sigma^2}\sigma^2 \\ &= \frac{1}{2}\left[\log\frac{\tau^2}{\sigma^2} + \frac{\sigma^2}{\tau^2} - 1\right] + \frac{(\mu-\mu')^2}{2\tau^2}. \end{aligned}$$

Substituting back $\sigma^2 = \sigma^2_{\text{pers}}(w)$ and $\tau^2 = \sigma^2_{\text{pers}}(w')$ gives (13).

□

D. Experiment: Discriminability Calibration for the Persona-Based Ad Benchmark

This experiment provides a discriminability audit for the persona-simulation experiment used in demonstrating the effectiveness of TEXTBO (Kang & Yoganarasimhan, 2025), a prompt optimization based self-improving AI method. Specifically, we focus on the ad optimization experimentation proposed in this paper.

In the ad optimization experiment, TEXTBO iteratively improves the *prompt for ad image generation* that is fed to an image model to produce the ad creative, using evaluation feedback to decide which prompt edits to keep. Concretely, it runs an automated prompt improvement loop: *write a prompt* $\rightarrow$ *generate an ad image* $\rightarrow$ *evaluate it on a target audience* $\rightarrow$ *reflection* $\rightarrow$ *edit the prompt and repeat*.

As in Kang & Yoganarasimhan (2025), we consider eight synthetic ad campaign scenarios that cover a diverse range of products across distinct categories, each defined by a creative brief that outlines the strategic and creative direction (see Web Appendix C.1 of Kang & Yoganarasimhan (2025) for the full creative briefs):

- Scenario 1: "GreenBite," a new plant-based burger patty.
- Scenario 2: "AuraSonics X1," high-end, noise-canceling wireless earbuds.
- Scenario 3: "Odyssey E-SUV," a new all-electric family SUV.
- Scenario 4: "Oasis Eco-Lodge," a secluded, luxury resort with beautiful natural surroundings.
- Scenario 5: "Momentum," a mobile-first banking app for freelancers and the gig economy.
- Scenario 6: "MindGarden," a subscription-based meditation and mindfulness app.
- Scenario 7: "Aeterno," a classic, automatic Swiss-made wristwatch with a heritage design.
- Scenario 8: "SyncFlow," a project management and collaboration software platform for remote teams.

For evaluation, TEXTBO utilizes Twin-2k-500 persona dataset (Toubia et al., 2025). For a fixed scenario and prompt, TextBO evaluates an ad image by: (i) sampling 200 personas from the Twin-2k-500 *training* split, (ii) conditioning a multimodal LLM judge on each persona's survey answers, requesting an effectiveness rating on the 1–5 scale, (iii) converting the judge's log-probabilities over {1,2,3,4,5} into a real-valued expected score per persona, and (iv) averaging over the 200 personas to produce a single scalar score. This single scalar is the only feedback used by the optimization method. We use Gemini 2.5 Flash with the same meta-prompt used in Kang & Yoganarasimhan (2025), which is provided in Figure 1.

Benchmark hygiene. This setup is designed to satisfy the two protocol conditions in the main theorem. Aggregate-only observation (AO) is enforced because the method-facing benchmark output is only the averaged scalar score; persona identities, individual ratings, log-probabilities, raw judge completions, and sample ordering are not returned to the prompt optimizer. Method-blind evaluation (MB) is enforced by holding fixed the persona sampling rule, judge model, meta-prompt, score conversion, and aggregation rule across submitted artifacts. The judge prompt contains the persona profile and the ad artifact, but not the optimizer identity, iteration number, or whether the artifact was produced by TEXTBO or a baseline method.

Here, in this paper's language, an artifact w corresponds to [scenario + prompt + its generated ad image]. We denote an evaluation operation of an artifact w, which returns a single scalar $o \in \mathbb{R}$, as `Eval(w)`. In the original TEXTBO setup, one call uses a micro-panel of $m = 200$ personas. Repeating `Eval(w)` with fresh independent persona ratings yields i.i.d. draws $o \sim Q_{\text{pers}}(\cdot \mid w)$. As TEXTBO iteratively improves the prompt for ad image generation, we would like to answer the following question:

How many independent persona ratings are needed to be confident that the improved prompt is better than the original prompt?

In this problem, $d_W(w, w')$ naturally corresponds to the number of clause-level instruction edits needed to transform prompt π into π'. Therefore, we set the resolution $r = 1$, i.e., we require the benchmark to reliably distinguish prompts that differ by at least one step of prompt improvement.

There are many ways of choosing the sampling distribution of (w, w') pairs. Kang & Yoganarasimhan (2025) reported results for ten steps of prompt improvement from the initial prompt; we tested ten improvement variants for each step. This

Persona prompt for simulating ad effectiveness.

SYSTEM: { You are an AI assistant. Your task is to answer the TASK as if you are the individual described in the 'Persona Profile' (which contains their past survey responses). Remain consistent with the persona's past survey responses and stated characteristics. Carefully follow any instructions provided for the new question, including formatting requirements. }
PERSONA DATA: {
Which part of the United States do you currently live in?
Question Type: Single Choice
Options:
1 - Northeast (PA, NY, NJ, RI, CT, MA, VT, NH, ME)
2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)
3 - South (TX, OK, AR, LA, KY, TN, MS, AL, WV, DC, MD, DE, VA, NC, SC, GA, FL)
4 - West (WA, OR, ID, MT, WY, CA, NV, UT, CO, AZ, NM)
5 - Pacific (HI, AK)
Answer: 2 - Midwest (ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH)
What is the highest level of schooling or degree that you have completed?
Question Type: Single Choice
Options:
1 - Less than high school
2 - High school graduate
3 - Some college, no degree
4 - Associate's degree
5 - College graduate/some postgrad
6 - Postgraduate
Answer: 3 - Some college, no degree
. . . (Many other survey questions and answers) . . .
Suppose you were given $5 and had to offer to another (anonymous) person a way to split the money. The other person can either accept or reject your offer. If the other person accepts your offer, you would each receive the amount you proposed. If the other person rejects your offer, you would both receive $0. How much would you offer to the other person?
Question Type: Single Choice
Options:
1 - $0
2 - $1
3 - $2
4 - $3
5 - $4
6 - $5
Answer: 3 - $2
. . . (Many other survey questions and answers) . . .
}

AD IMAGE: [image]

TASK:
Return only one item from ["1","2","3","4","5"] for ad effectiveness.
Effective Score Scale Definition:
1: Extremely Unlikely. The persona would actively ignore or be annoyed by this ad.
2: Unlikely. The persona would likely scroll past without a second thought.
3: Mediocre. It is hard to decide whether the personal would click or don't click.
4: Likely. The persona is intrigued and has a good chance of clicking to learn more.
5: Extremely Likely. The persona is the ideal target; a click is almost certain.
No explanation. Just the score.

Figure 1. Meta-prompt for simulating the effectiveness of a given ad-persona combination.

totals 100 TEXTBO prompt improvement cases per scenario. We then estimate $\widehat{\kappa}_Q(q)$ per scenario. By Lemma 4.2, the predicted number of independent persona ratings per prompt required to decide between two r-separated artifacts is

$$N_{\text{eval}}^{\text{req}} \;=\; \left\lceil \frac{2}{\widehat{\kappa}_Q(q)} \log \frac{1}{\delta} \right\rceil .$$

Table 1 provides $\widehat{\kappa}_Q(q)$ and $N_{\text{eval}}^{\text{req}}$ for each scenario. Across the scenarios, $N_{\text{eval}}^{\text{req}}$ varies from 46 to 1303, with an average of 469.25. This shows that the original micro-panel size $m = 200$ is not too bad, but increasing the independent persona-rating budget to about 500 could have been a more conservative choice.

Table 1. $\widehat{\kappa}_Q(q)$ and required independent evaluation units for $q = 0.05$ and $\delta = 0.05$

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Average
$\widehat{\kappa}_Q(q)$	0.00508	0.02320	0.13178	0.00942	0.02955	0.13069	0.0046	0.07548	0.051225
$N_{\text{eval}}^{\text{req}}$	1180	259	46	637	203	46	1303	80	469.25

E. Experiment: Persona-Based Sales Call-Conversation Stopping Benchmark

This experiment provides an illustrative extension of our framework to an interactive, sequential benchmark. We adapt the sales-conversation stopping task of Manzoor et al. (2025) to a persona-simulated setting while preserving the method-facing structure of the source task: the submitted artifact is still a silent stopping policy that acts on transcript prefixes, development still relies on full trajectories with ex post optimal labels, and evaluation still occurs on held-out calls. The goal is therefore a *structural analogue* of the original benchmark, not a literal reconstruction of the proprietary telecommunications dataset.

Source task. Manzoor et al. (2025) studies when a seller should stop an outbound sales call. For call j, at decision epoch $t = 1, \dots, T_j$, the state $s_{j,t}$ is the transcript prefix observed up to time t, and a stopping policy $\pi(a \mid s_{j,t})$ chooses $a_{j,t} \in \{\texttt{wait}, \texttt{quit}\}$. Let

$$\tau_j^\pi := \inf\{t : a_{j,t} = \texttt{quit}\},$$

with $\tau_j^\pi = T_j$ if the policy never quits before the conversation ends. The objective is to solve

$$\max_\pi \ \mathbb{E}\left[\sum_{t<\tau_j^\pi} c_t + q_j(\tau_j^\pi)\right],$$

where c_t is the opportunity cost of continuing the call and $q_j(\tau_j^\pi)$ is the terminal payoff if the sale is still alive when the policy stops. A useful structural feature is that the stopping policy does not control the dialogue itself: `wait` only reveals more of an exogenously generated conversation, while `quit` ends observation. This makes the task an optimal-stopping problem over transcript prefixes rather than a general dialogue-control problem.

Design goal and scope. The source paper studies real outbound sales calls from a telecommunications campaign and evaluates policies on held-out historical conversations. Our simulated benchmark aims to preserve three design elements of that setup:

1. the artifact is a *silent* `wait`/`quit` policy over transcript prefixes;
2. offline development uses full call trajectories to infer pseudo-expert stopping labels under a fixed reward model; and
3. final evaluation uses calls that are disjoint from development data.

Twin-2K-500 is used to parameterize both sides of the latent call, but in an asymmetric way. On the buyer side, the benchmark aims to cover the entire held-out buyer panel rather than repeatedly sampling a small subset of buyers. On the seller side, we introduce a fixed roster of synthetic salespeople to mirror the finite salesperson pool in the source study, while keeping that roster frozen across methods so that seller variation remains part of the environment rather than a developer-controlled degree of freedom. We also replace literal wall-clock checkpointing with transcript-stage checkpointing: instead of using 30/60/90 seconds, the stopping policy acts after the first, second, and third buyer utterances. Given that failed calls in the source study last about 169 seconds on average, those original checkpoints are best understood as early qualification stages rather than as precise timing targets; the buyer-stage schedule is our transcript analogue of that structure.

E.1. Benchmark construction.

We use the Twin-2K-500 persona dataset (Toubia et al., 2025) to define both a buyer pool and a seller roster. Let $\mathcal{S}$ denote a fixed roster of 80 seller personas, chosen to be disjoint from the buyer side; the count 80 is used to match the source study's roughly 79 observed salespeople at the level of benchmark design. Let $\mathcal{U}$ denote the remaining buyer-persona pool, and partition it into two disjoint sets

$$\mathcal{U}_{\text{off}} \cap \mathcal{U}_{\text{field}} = \varnothing, \qquad \mathcal{U}_{\text{off}} \cup \mathcal{U}_{\text{field}} = \mathcal{U},$$

with approximately half of the buyer personas in each split. The offline split $\mathcal{U}_{\text{off}}$ is used to build a synthetic logged dataset for policy development, while the field split $\mathcal{U}_{\text{field}}$ is reserved for prospective evaluation on unseen buyers. In the realized half-panel instantiation, $\mathcal{U}_{\text{field}}$ contains 1,029 buyers, and a six-scenario benchmark sweep evaluates every held-out buyer once per scenario while independently assigning a seller drawn from $\mathcal{S}$ for each buyer-scenario call.

Frozen latent call generator. Let G be a frozen simulator for one outbound cross-selling call. Given a seller persona $s \in \mathcal{S}$, a buyer persona $p \in \mathcal{U}$, and simulator randomness ω, it generates a *complete latent conversation*

$$\tilde{\xi} = (\tilde{x}_{1:\tilde{N}}, \tilde{y}, \tilde{T}^{\text{nat}}),$$

where $\tilde{x}_{1:\tilde{N}}$ is the full transcript, $\tilde{y} \in \{0, 1\}$ is the natural sale outcome, and $\tilde{T}^{\text{nat}}$ is the natural call duration in seconds. In the realized sweep reported below, seller and buyer utterances were generated with Gemini 2.5 Flash, using the API model identifier `gemini-2.5-flash`, the prompts in Figures 2 and 3, temperature 0.7, top-p 0.95, no external tools or retrieval, and per-turn output caps of 96 tokens for seller utterances and 80 tokens for buyer utterances. The call alternates seller and buyer turns, begins with the seller, and is capped at 16 total utterances. Seller assignments and scenario sweeps use pseudorandom seed 20260430; model-sampling randomness is represented by the realized transcript panel. Because hosted model snapshots can change, policy evaluation is run only after that transcript panel and the post-processed outcome and timing fields have been frozen.

The phase state supplied to the seller prompt is updated deterministically from the transcript prefix before each seller turn. The precedence order is `identity` when the buyer asks who is calling or why, `busy` when the buyer raises time pressure, `brush_off` when the buyer says they are not interested, asks for information later, or requests a callback, `price_question` when the buyer asks about cost or billing, `protection_question` when the buyer asks about coverage or product mechanics, `follow_up` after the buyer has asked multiple concrete questions or after the seller has made a closing attempt, `opening` at the start of the call, and `generic` otherwise. The simulator stops naturally at the first explicit same-call acceptance, explicit final refusal, deferral after a closing attempt, seller close, or the 16-utterance cap. At the cap, the call is treated as no-sale unless the buyer has explicitly agreed to add the product during the call.

The outcome and duration fields are assigned by fixed post-processing rules rather than by a policy-dependent judge. The natural outcome is $\tilde{y} = 1$ only when the generated buyer transcript contains an unambiguous agreement to add the product now, such as "yes, add it," "go ahead and add it," or "I'll take it today"; requests for email, a callback, time to think, or additional information are coded as $\tilde{y} = 0$. The acceptance, deferral, and refusal phrase classes are fixed before policy evaluation and applied identically across all submitted methods. The natural duration is computed from the generated transcript by a deterministic timing model,

$$\tilde{T}^{\text{nat}} = 20 + \sum_{i=1}^{\tilde{N}} \left(4 + \frac{60}{130}\, n_i\right),$$

where n_i is the word count of utterance i. Thus durations are measured in seconds, include a 20-second setup term, a 4-second turn-transition term, and spoken time at 130 words per minute. The checkpoint time $\tilde{T}_k^{\text{buyer}}$ is computed by applying the same timing formula through the end of the k-th buyer utterance. Crucially, this latent conversation and its outcome and timing fields are generated *before* any submitted stopping policy is applied.

This design is the simulator analogue of the source paper's historical-call dataset. Because the submitted artifact never speaks and only decides whether to keep observing, `wait` does not alter the underlying continuation of the call. Future dialogue is therefore treated as exogenous, exactly as in the replay-style counterfactual logic of Manzoor et al. (2025). Seller heterogeneity enters only through the frozen roster $\mathcal{S}$ and the random assignment of a seller to a buyer before the call starts; the submitted policy neither observes seller identity as a reusable external label nor controls which sellers appear.

In our implementation, G is not modeled as an unconstrained free-form chatbot. Pilot generations with open-source chat models indicated that, without additional prompt discipline, the generated calls became too polished, too verbose, and too willing to invent concrete coverage or pricing details that the seller should not know. We therefore make the latent-call generator *phase-structured* through the prompts themselves.

Scenario bank. To avoid concentrating all of the benchmark variation in a single product whose uptake is driven mostly by security anxiety, we instantiate the seller side through a bank of six outbound cross-selling campaigns. The campaigns are chosen so that different construct families already measured in Twin-2K-500 become behaviorally relevant: anxiety and risk preferences, green values, intertemporal and budgeting preferences, communal and family orientation, cognitive closure and comparison-shopping style, and uniqueness and status orientation. The benchmark protocol is identical across scenarios; only the seller-side product block, seller-style conditioning, the most likely objection patterns, and the subset of persona features emphasized in the buyer prompt change. All six offers are intended to be plausible but not obviously compelling

cold-call purchases, so that same-call conversion remains in the low single digits or low tens rather than becoming trivially high.

Table 2 summarizes the six scenarios. In a full-scale instantiation, one can either report scenario-specific results or pool them by sampling a scenario together with the seller-buyer pair before each latent call is generated. With one stochastic draw per buyer-scenario pair and an independently sampled seller from the 80-person roster, a 1,000-buyer panel yields 6,000 calls, putting the simulated log in the same order of magnitude as the source study's training split. Figures 2 and 3 show the exact seller and buyer prompt templates for Scenario S1; the other five scenarios reuse the same phase-structured scaffold with scenario-specific product information substituted into the seller-side block.

Table 2. Six outbound sales-call scenarios used to diversify the latent-call generator. The scenarios are chosen to activate different subsets of Twin-2K-500 measures while keeping the task structure fixed.

ID	Campaign	Product information given to the seller	Twin-2K-500 constructs most likely to matter
S1	Mobile Secure Plus	Existing mobile subscribers are offered a device-protection add-on with a free 30-day trial and then USD 7.99/month; the seller may mention scam and spam filtering, phishing alerts, cloud backup up to 256 GB, remote lock/wipe help, and approved replacement support with a USD 89 service fee.	Anxiety, neuroticism, risk aversion, conscientiousness, household size, financial literacy.
S2	Green Power Match	Existing household electricity customers are offered an optional renewable-energy rider with the first month free and then USD 6.99/month; the rider matches household usage with renewable supply, provides a monthly carbon-impact report, and sends peak-time conservation alerts, with no home visit or equipment change.	Green values, political views, conscientiousness, numeracy, financial literacy, household size.
S3	Rainy Day Auto-Save	Existing checking-account customers are offered a paid automated savings service with the first 60 days free and then USD 3.99/month; the service rounds purchases into savings, allows an optional weekly transfer, pauses automatically when balances run low, and provides emergency-fund goal nudges.	Mental accounting, discounting, present bias, financial literacy, numeracy, income.
S4	Family Locator Plus	Existing wireless customers are offered a family-safety add-on with a free 30-day trial and then USD 6.99/month per account; the add-on covers up to four lines and includes live location, arrival alerts, low-battery alerts, and SOS check-ins.	Household size, communal values, empathy, conscientiousness, employment status and time pressure.
S5	Price Lock Promise	Existing home-internet customers are offered a paid rate-lock add-on with the first month free and then USD 5.99/month; the offer locks the current internet price for 12 months, includes outage text alerts and priority troubleshooting, and requires no equipment change.	Need for closure, maximization, risk aversion, financial literacy, household budget sensitivity.
S6	Upgrade Pass	Existing mobile customers are offered a premium upgrade membership with the first month free and then USD 9.99/month; the offer includes early handset-upgrade eligibility after 12 paid months, a USD 75 accessory credit, and launch-day ordering priority, with benefits ending if the membership is cancelled.	Need for uniqueness, self-monitoring, agency, tightwad/spendthrift tendency, consumer minimalism.

Seller prompt used in Scenario S1 of the sales-call generation run.

You are a commissioned outbound telesales agent in a high-volume cross-selling campaign at a U.S. telecommunications firm. You are calling an existing mobile subscriber and trying to interest them in a paid mobile-protection add-on for their phone plan.

Product information you are allowed to use:
- Product name: Mobile Secure Plus
- Price: first 30 days free, then USD 7.99 per month, added to the customer's existing mobile bill
- Main benefits: scam and spam call filtering; suspicious-link and phishing alerts; cloud backup for photos, contacts, and messages up to 256 GB; priority tech support and remote lock/wipe help for a lost phone; replacement support for theft or accidental damage, subject to approval
- Limits and conditions: one approved replacement claim per 12 months; USD 89 service fee for an approved replacement claim; no coverage for pre-existing damage or cosmetic wear only; after the free 30-day trial, the service continues monthly unless the customer cancels; customer can cancel any time after the free 30-day trial

Seller profile: {`full Twin-2K-500 seller summary text`}
Likely selling style: {`seller-derived behavior notes`}

Conversation so far: {`transcript history or [start of call]`}
Current call phase: {`opening, identity, busy, brush_off, price_question, protection_question, follow_up, generic`}

Follow a realistic call-center script:
1. Verify you are speaking with the right person.
2. Briefly explain why you are calling.
3. Ask at most one simple qualifying or needs question when it fits the flow of the call.
4. Offer a concrete value proposition such as extra phone security, device protection, backup support, scam filtering, or peace of mind.
5. Handle objections briefly and honestly.
6. If the prospect seems like a plausible fit, make a light closing attempt.
7. If the prospect is clearly not interested, close the call politely.

Constraints:

- Stay consistent with the seller profile.
- Sound like a human call-center salesperson.
- Use plain spoken phone language, not polished marketing copy.
- Keep each turn to 1–3 short spoken sentences.
- Aim for roughly 12–40 words; going a little longer is acceptable when answering a concrete practical question.
- Do not narrate or use stage directions.
- Do not mention being an AI.
- Do not use placeholders such as [Customer Name] or [Address].
- Use the product information above consistently.
- Do not invent any prices, coverage details, exclusions, fees, or legal terms beyond the product information above and anything already stated in the conversation.
- If you do not know a detail beyond the product information above, say you would need to check or send it later.
- Keep pushing the call forward without sounding aggressive.
- Most calls will not end in a sale, but a small minority do.
- Do not treat 'I will read the email later' or 'call me back later' as a completed sale.
- A completed sale happens only if the prospect clearly agrees during this phone call to add the product now.
- A realistic non-trivial call often lasts several exchanges, commonly around 8–16 total turns, rather than ending after the first mild objection.
- Do not offer to email details as the default next move; use email or a callback only after the prospect asks for it or after you have already made a real on-call value case.
- If the fit seems strong, it is realistic to mention the free 30-day trial and ask directly whether they want it added today.
- Do not invent extra concessions such as automatic cancellation at the end of the trial, online-only cancellation, reminder services, waived fees, or any other billing or legal terms not explicitly listed above.
- Let the seller's pacing and tone vary across personas, but do not let seller heterogeneity change the underlying product facts.

Write the salesperson's next utterance only. Keep it realistic, phone-natural, and non-aggressive. A moderately detailed answer is fine if the prospect asked a concrete question.

Figure 2. Seller prompt used in Scenario S1. The remaining five scenarios in Table 2 reuse the same phase-structured prompt scaffold, but replace the seller-side product block and the corresponding scenario-specific seller constraints. At runtime, placeholders are instantiated with the current transcript prefix, the seller's phase, the seller and buyer persona summaries, and seller- and buyer-derived behavior notes.

Buyer prompt used in Scenario S1 of the sales-call generation run.

You are the prospect in an outbound telesales call. Respond as the person described in the persona profile below.

Call context:
- The caller works for a telecommunications company.
- They are cross-selling "Mobile Secure Plus," a paid mobile-protection add-on, to an existing mobile subscriber.
- This is a first-contact outbound call.
- You did not request this call.

Instructions:
- Stay consistent with the persona profile.
- Speak naturally, like a real person on the phone.
- Use casual spoken language rather than polished written prose.
- Keep each response to 1–3 short spoken sentences.
- Default to a guarded cold-call reaction. Aim for roughly 6–30 words, and go past that only when the persona would genuinely engage.
- Show curiosity, skepticism, budget concerns, time pressure, or indifference only when they fit the persona.
- Do not use placeholders such as [Customer Name].
- Do not sound like a lawyer, compliance officer, or policy analyst unless the persona strongly suggests that style.
- Do not narrate or use stage directions.
- Do not mention being an AI.
- In a cold call, it is realistic to ask who is calling, ask for the short version, say you are busy, ask what it costs, ask what it covers, say you already have something similar, or ask them to send the details later.
- Most first-contact calls like this do not end in an immediate purchase. Only agree to buy if the product clearly fits the persona and the seller addresses the main hesitation.
- If you ask the seller to email details or call back later, that is not a sale; it means you are deferring the decision.
- A realistic rejection can come after a few back-and-forth turns rather than immediately.
- If you are undecided but still listening, it is realistic to ask two or three concrete practical questions before rejecting or deferring.
- A sale is more plausible when the fit is good and the seller frames the offer as a low-risk trial rather than a permanent commitment.
- If the seller sounds too vague about billing, renewal, or cancellation, caution is realistic.

Persona profile:
{`full Twin-2K-500 persona summary text`}

Likely phone-call behavior:
{`persona-derived behavior notes`}

Conversation so far:
{`transcript history`}

Cold-call default: guarded, somewhat impatient, but still human. Common realistic reactions are: 'who is this?', 'what's this about?', 'I'm busy', 'how much is it?', 'what does it cover?', 'I already have something like that', or 'send it to me later'. If you are still listening, it is realistic to ask a couple of concrete practical questions before deciding. A low-risk trial can make a same-call purchase more believable when the product otherwise fits.

Write the buyer's next utterance only. Keep it realistic, phone-natural, and consistent with the persona. A practical question or objection is better than a generic refusal.

Figure 3. Buyer prompt used in Scenario S1. The remaining five scenarios in Table 2 reuse the same phase-structured prompt scaffold, but replace the call context and corresponding scenario-specific buyer objections. At runtime, placeholders are instantiated with the current transcript prefix, the buyer persona summary, and buyer-derived behavior notes.

Offline development data. Using only $\mathcal{U}_{\text{off}}$, we sweep over the buyer personas in the offline panel crossed with the six scenarios. For each buyer $p \in \mathcal{U}_{\text{off}}$ and scenario $a \in \{S1, \ldots, S6\}$, we independently draw a seller $s \sim \text{Unif}(\mathcal{S})$, draw simulator randomness ω, run $G(a, s, p, \omega)$, and record the resulting latent call. Writing one logged episode as

$$\tilde{\xi}_m = (\tilde{x}_{m,1:\tilde{N}_m}, \tilde{y}_m, \tilde{T}_m^{\text{nat}}),$$

we obtain an offline logged dataset

$$\mathcal{D}_{\text{off}} = \{\tilde{\xi}_m\}_{m=1}^{M_{\text{off}}}.$$

We randomly split $\mathcal{D}_{\text{off}}$ into development subsets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$. For each latent call, let $h_{m,k}$ denote the transcript prefix ending at the k-th buyer utterance, and let B_k denote that decision opportunity. We consider the buyer-stage schedules $\{B_1, B_2\}$ and $\{B_1, B_2, B_3\}$, which play the role of the source study's early stopping opportunities. This synthetic log plays the same role as the logged call dataset in Manzoor et al. (2025): it supports classifier baselines, pseudo-expert labels for imitation learning, and offline threshold selection before any prospective evaluation is run.

Pseudo-expert labels and imitation learning. For each latent call in $\mathcal{D}_{\text{train}}$ and each schedule, we compute the reward that would have been obtained by quitting at each feasible checkpoint together with the natural terminal time. Let τ_m^* denote the checkpoint that maximizes this reward. We then construct the imitation dataset exactly in the spirit of the source paper: prefixes before τ_m^* receive the label `wait`, the prefix at τ_m^* receives `quit`, and post-τ_m^* prefixes are augmented with `quit` labels to teach recovery from delayed stopping. This construction can be used to fit a stochastic stopping policy $\pi_{\hat{\theta}}(a \mid h_k)$ on $\mathcal{D}_{\text{train}}$, with deterministic action thresholds tuned on $\mathcal{D}_{\text{val}}$ by backward induction, mirroring the original offline pipeline.

Prospective field evaluation by hidden replay. After the simulator, calibration targets, and all model-selection choices are frozen using only $\mathcal{U}_{\text{off}}$ and $\mathcal{D}_{\text{off}}$, we evaluate policies on the disjoint field split $\mathcal{U}_{\text{field}}$. In the realized field-style instantiation, one benchmark sweep covers the entire held-out buyer panel once per scenario. For each buyer $p \in \mathcal{U}_{\text{field}}$ and scenario a, we independently draw a seller $s \sim \text{Unif}(\mathcal{S})$, draw simulator randomness ω, generate a latent full call $\tilde{\xi} = G(a, s, p, \omega)$, and *then* apply the submitted stopping policy only as an observation-censoring rule on that call.
The submitted artifact w is a *silent* stopping policy: it never speaks to the prospect and only decides whether the call should continue. Decision opportunities occur after the first, second, and optionally third buyer utterance, depending on the schedule. At each checkpoint B_k, the policy observes the current transcript prefix h_k and outputs either `wait` or `quit`. If the policy quits, the benchmark censors the latent call at that stage; otherwise the call continues along the *pre-generated* latent trajectory until the next buyer-stage checkpoint or a natural terminal sale/no-sale outcome. No future utterances are regenerated as a function of the submitted policy. This is the key design choice that keeps the simulator aligned with the original replay-style counterfactual benchmark.

Reward, benchmark output, and interface interpretation. For episode ℓ, let $\tilde{\xi}_\ell = (\tilde{x}_{\ell,1:\tilde{N}_\ell}, \tilde{y}_\ell, \tilde{T}_\ell^{\text{nat}})$ denote the latent full call. Let $\tilde{T}_{\ell,k}^{\text{buyer}}$ denote the elapsed natural time when the k-th buyer utterance ends, and let τ_ℓ^w be the first buyer-stage checkpoint at which policy w outputs `quit`, with $\tau_\ell^w = \infty$ if it never quits before the natural end. The realized duration under policy w is

$$T_\ell^w := \begin{cases} \tilde{T}_{\ell,\tau_\ell^w}^{\text{buyer}}, & \text{if } \tau_\ell^w < \infty, \\ \tilde{T}_\ell^{\text{nat}}, & \text{if } \tau_\ell^w = \infty. \end{cases}$$

We use the same economic structure as the source problem, with unit sale payoff and linear time cost:

$$R_\ell(w) = \mathbf{1}\{\tilde{y}_\ell = 1 \text{ and } \tilde{T}_\ell^{\text{nat}} \le T_\ell^w\} - c_{\text{time}} T_\ell^w.$$

This is the simulator counterpart of the source paper's waiting-cost-plus-terminal-payoff formulation: time is costly at rate c_{time}, and a unit sale reward is realized only if the latent call ends in a sale before the policy quits. In the reported tables, T is measured in seconds and $c_{\text{time}} = 5.595 \times 10^{-4}$, equivalent to 0.0336 sale units per minute. This choice is a reporting normalization rather than a fitted parameter; changing it changes the reward tradeoff but not the benchmark-interface construction. This normalization sets the no-stop baseline to zero on the completed replay panel; equivalently, the reported mean rewards are incremental utilities relative to no stopping and are not otherwise rescaled. The micro-response is $z_\ell := R_\ell(w)$.

In the language of this paper, an artifact w is a complete stopping policy. One benchmark call `Eval(w)` evaluates w on a held-out panel of L persona-conditioned latent calls, such as the $6|\mathcal{U}_{\text{field}}| = 6{,}174$ buyer-scenario calls in the realized sweep, assigns each call an independently sampled seller from $\mathcal{S}$, scores the latent calls under policy w, and returns the aggregate

$$o = \frac{1}{L}\sum_{\ell=1}^{L} R_\ell(w).$$

Repeating `Eval(w)` with fresh seller assignments and simulator randomness yields i.i.d. draws

$$o \sim Q_{\text{pers}}(\cdot \mid w).$$

For interpretability, the benchmark may also report aggregate sale rate, average duration, and fixed-budget sales obtained by reallocating saved time to fresh calls, exactly as in the original paper's expected-sales metric. The key point is that the sequential interaction occurs *inside* the evaluation channel. From the outer benchmark-interface perspective, evaluating a policy still maps w to a random aggregate output o, so the theory in the main text applies without change once call-level reward is taken as the micro-response.

Prompt realism. Realism is imposed primarily through the prompt design in Figures 2 and 3. On the seller side, realism comes from combining script discipline with a fixed seller roster: the seller must verify identity, state the reason for the call briefly, use a product-appropriate pitch, answer objections without over-selling, and avoid unsupported specifics, while tone and pacing may vary across seller personas. On the buyer side, realism comes from modeling a cold-called consumer rather than a cooperative chatbot: responses are kept short, attention is limited, and common early reactions include identity questions, brief skepticism, time-pressure objections, status-quo objections, and requests to send information later. These constraints give the generator a sales-call interface, but they do not make it a calibrated reconstruction of the source environment.

Benchmark hygiene. This setup is designed to satisfy the two protocol conditions in the main theorem. Aggregate-only observation (AO) is enforced at the benchmark boundary: the developer receives only policy-level aggregates and not persona identities, hidden simulator state, random seeds, or reusable per-call logs from the prospective evaluation split. The policy itself still observes the transcript prefix required to act within an episode, but that within-episode observation is part of the task definition rather than an external side channel. Method-blind evaluation (MB) is enforced by holding fixed the seller prompt, buyer prompt, outcome parser, decoding parameters, timing rules, calibration targets, and persona formatting across all submitted methods.

Policies under comparison. This benchmark naturally supports several policy families: fixed deadline rules, sale-probability classifiers with stopping thresholds, sequential ensembles of checkpoint-specific classifiers, imitation-learning policies trained on pseudo-expert stopping labels derived from $\mathcal{D}_{\text{off}}$, and RLVR policies trained directly in the simulator using the verifiable reward $R_\ell(w)$ (Lambert et al., 2024). The offline/field split ensures that model development and deployment-style evaluation remain separated even though both stages are simulated.

Distance, resolution, and discriminability. To apply Lemma 4.2, we need a task-appropriate distance on stopping policies. We generate a probe set $\mathcal{H}$ of transcript prefixes from held-out development trajectories under a frozen exploration policy and define the behavioral distance

$$d_{\mathcal{W}}(w, w') := \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} \mathbf{1}\{w(h) \neq w'(h)\}.$$

We set $r = 0.05$, so two policies are regarded as meaningfully different if they disagree on at least 5% of probe states. This is the sequential analogue of the "one meaningful prompt edit" choice in Appendix D: here the relevant unit is not surface form but policy behavior over representative transcript prefixes.

For the calculations reported in Table 6, we define ν to be uniform over the r-separated policy pairs induced by the three replay baselines in Table 5: no stopping, the last-checkpoint deadline rule, and the prefix-screen heuristic. The reported $\widehat{\kappa}_Q(q)$ is therefore the estimated lower-tail discriminability for this stated comparison distribution. The same calculation can also be applied under alternative choices of ν, such as distributions over threshold perturbations, neighboring deadline rules, prompt edits to the stopping agent, or nearby fine-tuning checkpoints. For back-end auditing, paired hidden episode panels

compare nearby policies under common random numbers even though only aggregates are exposed externally. The same sample-complexity calculation then applies:

$$N_{\text{eval}}^{\text{req}} = \left\lceil \frac{2}{\widehat{\kappa}_Q(q)} \log \frac{1}{\delta} \right\rceil .$$

Operationally, $N_{\text{eval}}^{\text{req}}$ is the number of independently sampled persona-conditioned latent-call draws per policy needed to distinguish two r-separated stopping policies with error probability at most $q + \delta$. It is not necessarily the number of unique buyer personas. When $N_{\text{eval}}^{\text{req}}$ exceeds the 1,029-buyer held-out panel size, as it does for several rows below, the calculation implies that one would need repeated simulator draws over the fixed panel with fresh seller, scenario, and generation randomness, a larger held-out persona panel, or both.

What this experiment adds. Relative to Appendix D, this benchmark changes both the task structure and the evaluation regime. The ad benchmark is a one-shot preference judgment, whereas the present benchmark is interactive and sequential. Relative to the source sales-call paper, the present appendix replaces real historical calls with a fixed latent-call generator and evaluates policies by *hidden replay* on unseen personas rather than by live dialogue control. The point is to show how the benchmark-interface framework and the discriminability analysis extend to sequential policy evaluation in a controlled LLM-persona environment while preserving the original counterfactual logic at the interface level.

The benchmark is summarized by four tables. Table 3 summarizes the generated six-scenario dataset itself, including the held-out buyer coverage, the realized seller support from the fixed 80-person roster, and the scenario-level call statistics. In the realized generation sweep, the held-out field split contributes 1,029 buyers and therefore 6,174 buyer-scenario calls overall, with a natural same-call sale rate of 9.26%, an average duration of 165.5 seconds, and an average of 9.8 turns per call. Table 4 reports the corresponding six-scenario calibration diagnostics relative to the source study, including a held-out TF–IDF logistic predictor of eventual failure from the B_1 prefix. These diagnostics also show where the simulator is not a calibrated replication of the source environment: successful calls are much shorter and early failure is much less predictable than in the source data. We therefore interpret the experiment as an illustrative benchmark-interface stress test rather than as a faithful reconstruction of the original sales-call environment. Table 5 then evaluates three simple replay baselines on that same completed six-scenario sweep under each buyer-stage schedule, namely a no-stop baseline, a last-checkpoint deadline rule, and a prefix-screen heuristic. Finally, Table 6 reports the resulting $\widehat{\kappa}_Q(q)$ and $N_{\text{eval}}^{\text{req}}$ values for several schedule-specific evaluation slices, including a short-call audit defined by the realized median natural duration.

Table 3. Summary of the generated six-scenario sales-call dataset. The run covers 1,029 held-out buyers, a fixed 80-person disjoint seller roster, and 6,174 generated calls.

Slice	Calls	Buyers	Sellers	Sale rate	Avg. duration (s)	Avg. turns
Overall	6,174	1,029	80	9.26%	165.5	9.8
S1: Mobile Secure Plus	1,029	1,029	80	10.88%	192.0	11.3
S2: Green Power Match	1,029	1,029	80	12.83%	168.3	10.0
S3: Rainy Day Auto-Save	1,029	1,029	80	3.11%	134.6	7.9
S4: Family Locator Plus	1,029	1,029	80	14.38%	168.3	10.1
S5: Price Lock Promise	1,029	1,029	80	8.65%	141.9	8.3
S6: Upgrade Pass	1,029	1,029	80	5.73%	187.9	11.2

Table 4. Calibration diagnostics for the persona-based sales-conversation stopping benchmark. The simulator column uses the completed six-scenario generation sweep over the 1,029-buyer held-out half-panel with the fixed 80-seller disjoint roster. The AUC row reports a held-out TF–IDF logistic predictor of eventual failure using transcript prefixes through B_1.

Diagnostic	Source study	Simulator
Sale rate	5.5%	9.26%
Average duration (s)	195	165.5
Failed-call duration (s)	169	165.8
Successful-call duration (s)	630	162.7
Share of call time on failed calls	82%	90.9%
Held-out AUC of early failure predictor	0.94	0.60

Table 5. Illustrative replay-policy comparison on the completed six-scenario sweep over the 1,029-buyer held-out half-panel. Mean reward uses T in seconds, $c_{\text{time}} = 5.595 \times 10^{-4}$, and the no-stop normalization described in the text. The prefix-screen heuristic quits when the current buyer-stage prefix already contains a strong early disinterest, defer, or status-quo marker. "Fixed-budget sales" is the expected sales rate obtained by reallocating saved time to fresh calls drawn from the same distribution, following the source paper's time-reallocation calculation.

Policy	Schedule	Mean reward	Sale rate	Avg. duration	Fixed-budget sales
No-stop baseline	$\{B_1, B_2\}$	0.0000	9.26%	165.5	9.26%
Last-checkpoint deadline	$\{B_1, B_2\}$	−0.0277	0.66%	61.4	6.49%
Prefix-screen heuristic	$\{B_1, B_2\}$	−0.0073	6.92%	136.6	8.54%
No-stop baseline	$\{B_1, B_2, B_3\}$	0.0000	9.26%	165.5	9.26%
Last-checkpoint deadline	$\{B_1, B_2, B_3\}$	−0.0286	2.30%	92.2	6.40%
Prefix-screen heuristic	$\{B_1, B_2, B_3\}$	−0.0100	6.09%	126.6	8.27%

Table 6. Estimated discriminability and required latent-call draws for the sales-conversation stopping benchmark at $q = 0.05$ and $\delta = 0.05$, using the three six-scenario replay baselines in Table 5. Short-call audit rows restrict to calls below the median natural duration of 164.5 seconds. Within each slice, $r = 0.05$ and ν is uniform over the three displayed baseline policy pairs with $d_{\mathcal{W}}(w, w') \geq r$. The last column is descriptive only and is not used to define the lower-tail $\widehat{\kappa}_Q(q)$.

Evaluation slice	$\widehat{\kappa}_Q(q)$	$N_{\text{eval}}^{\text{req}}$	Largest observed pair
Overall, $\{B_1, B_2\}$	0.000351	17,063	No-stop vs. Deadline
Short calls, $\{B_1, B_2\}$	0.001375	4,359	No-stop vs. Deadline
Overall, $\{B_1, B_2, B_3\}$	0.000691	8,676	No-stop vs. Deadline
Short calls, $\{B_1, B_2, B_3\}$	0.001891	3,169	No-stop vs. Deadline

F. Further discussions

F.1. Further discussions on contribution

What does the robustness result add? We now present the exact equivalence as the zero-error case of a quantitative robustness statement, not as the main mathematical endpoint. The exact result gives precise meaning to the phrase "personas are just another panel": the actual protocols admit a literal-panel-change representation on every adaptive transcript law. The substantive value is the surrounding stability analysis. Theorem 3.3 shows that approximate method-blindness gives $T\varepsilon$-approximate panel-change representation, while Proposition 3.4 shows that provenance sensitivity of size η creates an $\eta/2$ lower bound against any artifact-only panel-change representation. The AO counterexample is also quantitative: raw-vote leakage creates an $O(1)$ method-interface failure, with total-variation distance at least 0.24, even though the aggregate kernels match exactly (Proposition C.1).

F.2. Further discussions on extensions

Would more empirical validation strengthen the arguments in this paper? The central claim in this paper is an identification theorem about benchmark interfaces, not an empirical claim that a particular persona pipeline tracks humans. The paper includes a worked κ_Q calibration (Appendix D) to demonstrate the intended workflow: κ_Q is designed to be measured and to turn "persona quality" into a budget/design question. Requiring extensive multi-domain experimentation is therefore more about strengthening the paper's applied guidance than about validating the correctness of the main theorem.

Can sample complexity easily be extended beyond pairwise? Pairwise comparison is the primitive building block, and this is easily extended beyond pairwise comparison. For example, if the workflow is "compare K fixed artifacts," pairwise error control can be achieved with a union bound by allocating per-comparison failure probability $\delta' = \delta/C(K, 2)$ (or δ/K for tournament-style brackets), which yields only a log-factor change in required samples.