Chorus: When Synthetic Audiences Should Decide or Abstain

Berat Tuna Karli [1]

Abstract

Teams often use LLM-based simulations to compare variants before real outcome data is available. However, a forced synthetic winner can be misleading when the available evidence is weak. Chorus treats synthetic audience evaluation as selective decision support. Persona-conditioned LLM agents compare a control and challenger in counterbalanced order and return signed judgments with rationales. The system then recommends a direction, or abstains while reporting disagreement diagnostics and a validation route. We evaluate Chorus with Gemini 2.5 Flash-Lite on 96 outcome-labeled cases, mostly public A/B case studies plus a small set of human-panel labels. With 64 personas, forced directional prediction matched the reference outcome in 75/96 cases (78.1%). A synthetic-vote diagnostic made recommendations on 67/96 cases (69.8%), was correct on 56/67 covered cases (83.6%), and reduced wrong decisive recommendations from 21/96 (21.9%) to 11/96 (11.5%) while abstaining on 29 cases (30.2%). These results describe selective decision behavior on this benchmark. They do not show calibrated population inference, causal lift, or distribution-free risk control. Instead, they show that selective metrics can reveal failures that forced winner accuracy hides, and that Chorus can be used as an empirical workflow for abstention, disagreement inspection, and validation routing. Chorus should therefore be read as pre-test decision support, not as a replacement for live experiments or human research.

1. Introduction

Before outcome data exists, teams still have to decide what is worth testing: which variant to prioritize, which wording to try, or which concept to show users. In these settings, the most useful synthetic-audience output is not always a forced winner. A decision-support system should also ask whether the evidence is strong enough to act, where audience perspectives conflict, and whether the next step should be revision, human review, or real-world validation.

Chorus is designed for early comparison tasks where a forced choice may be premature. The workflow can accept text, screenshots, images, documents, wording, survey context, product descriptions, or mixed inputs supported by the underlying model. The evaluation here does not include audio or video. For interpretability, Chorus still encourages simple control/challenger comparisons with one primary change; broad redesigns can be processed but are harder to attribute.

Chorus builds or selects personas from the audience, objective, decision mode, and context. Each persona compares neutral-labeled control and challenger variants in counterbalanced order and returns a structured signed vote, rationale, and self-reported confidence. The persona count should be interpreted as an evidence budget rather than a population sample. The 64 persona-conditioned judgments are used to expose agreement, disagreement, and weak-signal cases, not to estimate population preference shares. With multiple challengers, Chorus compares each challenger to the control and then ranks or shortlists challengers based on covered selective decisions; native multi-way comparison is not evaluated here.

Chorus is closest to synthetic A/B evaluation systems such as SimAB, but its emphasis is selective decision behavior rather than only winner prediction. It recommends a winner only when the synthetic evidence is strong enough, abstains otherwise, reports disagreement across synthetic audience segments, and routes weak cases to revision, live A/B testing, human-panel feedback, or more context.

The interface follows this same distinction. Appendix C shows setup and report surfaces that collect the decision context and expose synthetic votes, persona disagreement, evidence clusters, criterion scores, and revision guidance.

This paper makes three main contributions:

1. A selective decision-support formulation for synthetic audience evaluation, where outputs can be control, challenger, or abstention with an explicit next-step route.

[1]Meshona.ai. Correspondence to: Berat Tuna Karli <berat@meshona.ai>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

2. A persona-panel workflow that records signed votes, rationales, self-reported confidence, and disagreement diagnostics without interpreting personas as representative or independent people.
3. An empirical evaluation on 96 outcome-labeled cases, reporting forced accuracy, coverage, covered-decision accuracy, wrong decisive recommendations, tier and suite behavior, persona-budget trade-offs, and model-family sensitivity.

Evaluation terms. For clarity, we define the evaluation terms used throughout. The reference winner is the public A/B or human-panel outcome used for validation. "Forced directional accuracy" checks whether Chorus's forced direction matches that reference winner before applying the abstention rule. "Covered-decision accuracy" is accuracy only when Chorus makes a decision. "Coverage" is the proportion of cases where Chorus makes a recommendation; the rest are abstained. A "wrong decisive recommendation" is a non-abstaining recommendation for the wrong winner. "Tier A" cases are cleaner, direct outcome-linked cases with quantitative outcomes and close input fidelity. "Tier B" cases are outcome-linked but lower-fidelity cases, such as multi-screen tests, text-only versions, before/after sources, weaker pair-specific metrics, or descriptions combining multiple changes.

2. Related Work

A/B testing and simulated experimentation. When there are enough visitors to a site and outcomes can be properly tracked, the usual method for estimating a treatment effect is A/B testing or a similar experiment (Kohavi et al., 2020). Chorus is not a causal estimator. It is pre-test decision support for stepping back, shortening a list of options, deciding what to validate next, or making low-stakes reversible decisions. The most similar prior research is SimAB, which uses personas to have agents evaluate webpage versions and then checks results against historical A/B tests (Rieder et al., 2026). Chorus's goals are similar, but instead of forcing a winner, it focuses on deciding, abstaining, showing disagreement, improving something, or routing an idea to real validation. AgentA/B and SimGym are also about using agents or traffic-grounded simulation for websites and retail settings (Lu et al., 2025a; Castelo et al., 2026).

UX agents, synthetic users, and persona validity. UX agents, synthetic users, and persona validity are central to this setting. UXAgent examines LLM-agent usability testing and provides qualitative and quantitative results with video recordings (Lu et al., 2025c). Customer-simulation studies ask whether a digital twin can represent a customer's shopping journey, while large-scale work shows that prompt-only agents can be weak without stronger grounding or training (Sun et al., 2025; Lu et al., 2025b). PersonaCite and persona-evaluation work point to risks around evidence, abstention, perspective diversity, consistency, and reliability (Truss, 2026; Paglieri et al., 2026; Kim et al., 2026; Amin et al., 2025; Morocho et al., 2026). These findings motivate Chorus's conservative interpretation of personas as probes of plausible decision-relevant reasoning patterns, rather than samples from a population or replacements for real users.

LLM-as-judge, calibration, and abstention. LLM-as-judge reliability, calibration, and abstention under uncertainty are also central to Chorus. LLMs are often used to evaluate open-ended outputs (Zheng et al., 2023; Liu et al., 2023). Rephrasing, self-consistency, and multi-agent debate can reduce some inconsistency, but repetition alone does not make an answer accurate (Wang et al., 2023; Du et al., 2023). Calibration asks whether model confidence matches empirical correctness, and LLM confidence should be treated carefully (Dawid, 1982; Guo et al., 2017; Kadavath et al., 2022; Lin et al., 2022). Selective prediction studies systems that can reject or abstain under uncertainty (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017), and Chorus uses this principle directly.

3. Task Formulation

Let A denote the audience or community description, g the decision objective, m the decision mode, and X the task context. X can include structured information, free-form text, images, documents, or any mixture of inputs supported by the underlying model and application. A decision involves a control, v_0, and one or more challengers, $\{v_1, \ldots, v_k\}$. Currently, each pairing of v_0 and v_j is assessed separately.

This is expressed as:

$$d_j = F(A, g, m, X, v_0, v_j). \tag{1}$$

The paired comparisons can be used to select an alternative, shortlist candidates, or flag cases requiring review. This method of breaking the process down makes it clear what each prediction in the research refers to: it is a binary control/challenger comparison.

Chorus provides a "decision object" as its output, rather than simply stating which option is the best. This object is

$$D_j = (\hat{y}_j, a_j, u_j, g_j, E_j). \tag{2}$$

Within this, $\hat{y}_j$ is the system's recommendation of control, challenger, or abstention, a_j is a suggested next step, u_j summarizes uncertainty, g_j summarizes audience disagreement, and E_j is an evidence package that includes reason clusters, strengths, weaknesses, and trace metadata. The a_j step could be to choose the control, choose a challenger, refine the challenger, run a live A/B test, collect human-panel

feedback, or gather more context. Framing the outcome as a "decision object" is the primary distinction between selective decision support and forced synthetic A/B prediction.

Appendix C explains how the reporting interface presents these areas of information as a complete decision package, instead of just the winning option.

A group of personas of size t is created by a dynamic generator or selector. This is represented as

$$P_t = \{p_i\}_{i=1}^{t} = G(A, g, m, X, s, t), \tag{3}$$

where s is a seed. This notation is procedural, not probabilistic: personas are not independent draws from a true population. Each persona then makes a structured judgment, expressed as

$$z_i = (w_i, r_i, q_i, a_i, b_i, o_i). \tag{4}$$

Here, w_i is a vote for the control (-1), no preference (0), or the challenger $(+1)$, r_i is the reasoning behind the vote, q_i is self-reported confidence, a_i and b_i are the strengths and weaknesses, and o_i records the presentation order. The signed-vote mean is:

$$\bar{s}_t = \frac{1}{t}\sum_{i=1}^{t} w_i. \tag{5}$$

4. Workflow

Chorus first reads the audience, objective, mode, situational background, variants, persona budget, batch size, and seed; it then creates or chooses personas relevant to the task. The variants are presented to each persona in balanced order with neutral labels. Each persona makes a structured comparative choice, and labels are mapped back to control and challenger. Chorus returns a decision record with recommendation, uncertainty, disagreement summary, evidence clusters, and a next-step route: choose challenger, retain control, modify the challenger, run a live A/B test, collect human-panel feedback, or gather more context.

Dynamic persona creation. Chorus dynamically builds personas. In general applications, it uses a large language model to construct personas, taking into account the audience, goal, mode, context, and existing personas. For gaming and product scenarios, it uses axes of coverage including motivation, attention window, spend sensitivity, novelty tolerance, aesthetic preference, and friction tolerance. The aim is to cover plausible decision-relevant perspectives, not to invent demographic details.

Batched generation and judging. The benchmark uses 64 personas for each case, split into eight planned batches of eight judgments. The full group of 64 personas must be assessed before stopping, so early stopping is disabled. This makes evidence accumulation visible and allows later checks of whether fewer personas would have given the same result.

Implementation model. In the benchmark, Gemini 2.5 Flash-Lite is used through Google Application Default Credentials for persona creation, comparative judgments, and result summaries. The system records request attempts, successful responses, prompt tokens, output tokens, total tokens, retry/error counts, and usage metadata.

Chorus is intended as a flexible decision-support process, not only as a final evaluation tool. Appendix C shows how the system handles modes, multimodal content, audiences, locations, persona counts, and additional challenger options. Chorus is designed for early, reversible product, marketing, and design decisions, and for non-binding communication stress tests where the next step is human review or real-world validation. It still encourages simple control/challenger comparisons to help attribution.

5. Decision Uncertainty and Selective Abstention

5.1. Panel Statistics

After collecting judgments from t personas, $n_-(t)$, $n_0(t)$, and $n_+(t)$ count votes for control, no preference, and challenger. These numbers describe the synthetic panel as it is; we are not treating them as samples from a larger real population.

The first summary is the neutral rate:

$$r_0(t) = \frac{n_0(t)}{t}. \tag{6}$$

A high neutral rate suggests many personas did not see much difference between the two options. The second summary is vote entropy. We calculate $\pi_c(t) = n_c(t)/t$ for each choice: control, neutral, and challenger. The normalized entropy is defined below; the $\log 3$ denominator sets the scale.

$$H_t = -\frac{1}{\log 3}\sum_{\substack{c\in\{-,0,+\}\\ \pi_c(t)>0}} \pi_c(t)\log \pi_c(t). \tag{7}$$

When the panel is equally divided between the control, neutral, and challenger, the entropy is at its highest, $\log 3$. Dividing by $\log 3$ means the entropy H_t will be between 0 and 1. Close to 0 means most personas voted the same way, and close to 1 means opinions were widely spread. The third summary is a basic ranking score.

$$c_t^{raw} = \frac{1}{2} + \frac{1}{2}|\bar{s}_t|\left(1 - r_0(t)\right). \tag{8}$$

This score increases when the directional margin, $\bar{s}_t$, is large and the neutral rate is low. We only use this score to rank cases by apparent decisiveness in risk-coverage analysis; it is not a probability of correctness.

5.2. Wrong Decisive Recommendations and Disagreement Diagnostics

The key failure mode is not just being wrong; it is being wrong while still making a decision. For each case j with a reference outcome, $\hat{y}_j$ is the system's recommendation and $y_j^\star$ is the reference winner. We use an indicator for a "wrong decisive recommendation".

$$f_j = \begin{cases} 1, & \hat{y}_j \in \{\text{control}, \text{challenger}\} \text{ and } \hat{y}_j \neq y_j^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The wrong-decisive rate is the average of these indicators over all tested cases. It is the number of times the system picked the wrong winner without abstaining. This is central for Chorus, because it is meant to avoid making a judgment when the synthetic evidence is not strong enough.

Abstentions should also be interpretable. Chorus explains why a case is inconclusive rather than returning no explanation. Common reasons include a small vote margin, high neutral rate, high vote entropy, disagreeing persona segments, weak or conflicting rationales, criticism of both options, or too little input context.

5.3. Synthetic Vote Diagnostic and Abstention

The synthetic-vote diagnostic shows dispersion in the signed vote mean. Importantly, it is not an estimate of product lift, legal correctness, community-wide causal response, or the probability that the recommendation is correct. Persona judgments are created by an LLM workflow and could be correlated, prompt-sensitive, or contaminated by public examples. Each vote is converted to a number between 0 and 1, a Wilson-score calculation is then applied (Wilson, 1927), and the outcome is scaled back to the range $[-1, 1]$. This yields $[L_t, U_t]$ as the range for the signed vote mean. A selective decision is made based on this range:

$$y_t = \begin{cases} \text{challenger}, & L_t > 0, \\ \text{control}, & U_t < 0, \\ \perp, & \text{otherwise.} \end{cases} \tag{10}$$

If the whole range is positive, the synthetic panel favors the challenger. If the whole range is negative, the inclination is toward the control. When the range includes zero, Chorus abstains. Section 6.2 assesses whether this abstention method decreases wrong decisive recommendations, though it does not imply that the Wilson-style calculation is uniquely justified.

6. Benchmark

6.1. Cases and Protocol

The system was tested on 108 binary scenarios; 96 of these, where reference outcomes are known, are used to measure headline performance, and the remaining 12 are for regression testing. Runtime, token use, and API-call counts are measured across all 108 cases, including the 12 not included in the main accuracy measurement.

Of the 96 validation cases, 48 are Tier A and 48 are Tier B. Of these, 30 are expected to favor control and 66 the challenger. The cases include commerce, gaming stores, gaming concepts, Wikimedia, gaming monetization, gaming ad creative, and gaming onboarding; six gaming-concept cases have human-panel outcome labels. Each validation case employs 64 personas, divided into eight groups of eight, and runs to completion, resulting in 6,144 persona-level judgments. This 64-persona run finished without case errors and produced 11.5 million tokens, 7,108 API attempts, and 7,106 successful API responses.

These scenarios were taken from publicly viewable A/B testing results and author-held human-panel artifacts. They are used as a non-public offline directional validation set, not as a public benchmark release or deployment calibration set. Because the cases are partly public and because benchmark design choices were not fully fixed before system development, we treat the results as an initial characterization of system behavior rather than a guarantee of future performance. Table 1 summarizes the suites, and Appendix A describes missing provenance fields.

Table 2 provides a performance breakdown by evaluation tier.

While the tiers have comparable coverage, they differ after a decision is made. Tier B has lower forced and covered-decision accuracy, and its wrong-decisive rate is more than four times higher. The 95% case-bootstrap ranges are 84.8% to 100% versus 58.1% to 87.9% for covered-decision accuracy, and 0.0% to 10.4% versus 8.3% to 31.3% for wrong-decisive rate (Tier A versus Tier B).

Table 3 reports the same suite-level diagnostics in the main paper so that suite behavior is visible next to the tier discussion. The small suites remain descriptive smoke checks and should not be read as stable domain-level estimates. Appendix B gives the corresponding suite-level plot and additional persona-count diagnostics.

6.2. Selective Abstention Results

Table 4 contrasts forced prediction, simple always-pick baselines, and selective abstention.

The always-challenger baseline is correct in 66 of the 96

Table 1. Benchmark card for the 96 outcome-labeled validation cases. C/Ch is the expected-control versus expected-challenger split. Tier A/B summarizes input fidelity and outcome linkage. The table reports aggregate suite metadata only; case-level source assets are not released.

Suite	Cases	C/Ch	Tier A/B	Outcome source	Known limitations
Public commerce	35	9/26	27/8	public A/B	Source dates, effect sizes, and leakage status incomplete.
Public gaming store	27	7/20	7/20	public A/B	App-store asset reconstructions; pair fidelity varies.
Public gaming concept	10	5/5	9/1	public A/B + human panel	Mixed public/human-panel labels; small concept-level suite.
Public Wikimedia	9	5/4	3/6	public A/B	Public fundraising tests; benchmark-contamination risk.
Gaming monetization	7	4/3	0/7	public A/B	Small suite; reconstructed or combined treatment descriptions.
Gaming ad creative	4	0/4	0/4	public A/B	Very small suite; smoke check only.
Gaming onboarding	4	0/4	2/2	public A/B	Very small suite; split fidelity labels; no stable domain estimate.

Table 2. Tier summary for the 64-persona run on the 96-case validation set. C/Ch is the expected-control versus expected-challenger split. Covered-decision accuracy is computed only on non-abstaining decisions made by the synthetic-vote diagnostic.

Tier	Cases	C/Ch	Directional acc.	Covered acc.	Coverage	Inconcl.	Wrong decis.
Tier A	48	7/41	41/48 (85.4%)	31/33 (93.9%)	33/48 (68.8%)	15/48 (31.3%)	2/48 (4.2%)
Tier B	48	23/25	34/48 (70.8%)	25/34 (73.5%)	34/48 (70.8%)	14/48 (29.2%)	9/48 (18.8%)

cases because 66 favor the challenger; this imbalance is why forced accuracy alone is not sufficient. Forced prediction covers every case and achieves 78.1% (75/96) accuracy, with 21 wrong forced calls. The synthetic-vote diagnostic abstains from 29 cases, covers 67, and achieves 83.6% (56/67) covered-decision accuracy, leaving 11 wrong decisive calls out of 96. Raw-score and margin selections perform similarly at comparable coverage, so the result supports selective abstention and ranking more than the Wilson-style calculation as uniquely justified. The 95% case-bootstrap intervals for forced directional accuracy are 69.8% to 86.5%, for diagnostic coverage 60.4% to 79.2%, for diagnostic covered-decision accuracy 73.9% to 92.1%, and for the wrong-decisive rate 5.2% to 17.7%. A reduction of 10/96, or 10.4 percentage points, in wrong decisive calls was observed in paired testing, with a 95% case-bootstrap interval of 5.2 to 16.7 percentage points.

As a sensitivity check against threshold overfitting, we also ran a retrospective 5-fold cross-fitted threshold analysis. Thresholds were selected on four folds and evaluated on the held-out fold. This separates threshold selection from evaluation within the dataset, although it remains retrospective and should not be interpreted as prospective risk control or distribution-free validation. In target-coverage mode, with thresholds aiming for roughly 70% coverage, the persona-panel raw-score and margin thresholds covered 67 and 68 of the 96 cases, with 56/67 and 56/68 correct covered decisions, resulting in 11 and 12 wrong decisive calls. Using repeated thresholds without personas, 68 and 69 cases were covered, with 55/68 and 56/69 correct covered decisions and 13 wrong decisive calls in both instances.

Figure 1 shows the corresponding risk-coverage curves for the 64-persona validation run.

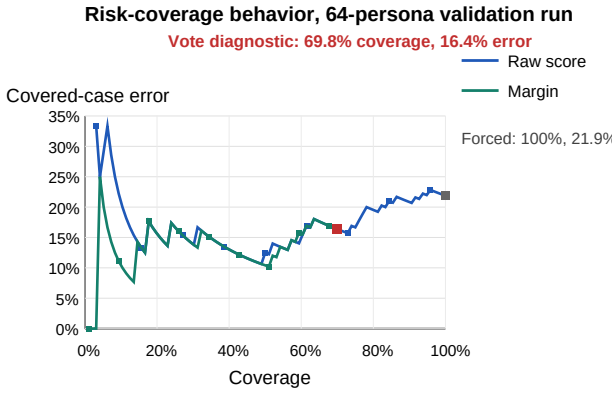

Figure 1. Risk-coverage behavior for the 64-persona Gemini 2.5 Flash-Lite run on the 96-case validation set. Score-ranked curves are uncalibrated ranking diagnostics. The synthetic-vote diagnostic covers 67/96 cases (69.8%) with 11/67 covered-case errors (16.4%).

6.3. Persona Budget Frontier

Tables 5 and 6 detail how the number of personas relates to performance. Gemini 2.5 Flash-Lite was used to repeat the process with 8, 16, 32, and 64 personas, and the 64-persona panel is contrasted with two no-persona baseline comparisons: a single direct judge and a 64-call repeated-

Table 3. Suite-level descriptive metrics for the 64-persona run on the 96-case validation set. Covered-decision accuracy is computed only on non-abstaining synthetic-vote diagnostic decisions.

Suite	Cases	Directional acc.	Covered acc.	Coverage	Inconcl.	Wrong decis.
Commerce	35	77.14%	81.48%	77.14%	22.86%	14.29%
Gaming store	27	77.78%	86.67%	55.56%	44.44%	7.41%
Gaming concept	10	80.00%	100.00%	20.00%	80.00%	0.00%
Wikimedia	9	77.78%	77.78%	100.00%	0.00%	22.22%
Gaming monetization	7	57.14%	66.67%	85.71%	14.29%	28.57%
Gaming ad creative	4	100.00%	100.00%	100.00%	0.00%	0.00%
Gaming onboarding	4	100.00%	100.00%	100.00%	0.00%	0.00%

Table 4. Headline selective prediction results on the 96-case validation set. Coverage is the fraction of cases with a non-abstaining decision.

Method	Covered	Correct/covered	Abstain	Wrong decis./all
Always control	96/96 (100.0%)	30/96 (31.3%)	0/96 (0.0%)	66/96 (68.8%)
Always challenger	96/96 (100.0%)	66/96 (68.8%)	0/96 (0.0%)	30/96 (31.3%)
Directional forced	96/96 (100.0%)	75/96 (78.1%)	0/96 (0.0%)	21/96 (21.9%)
Synthetic-vote diagnostic	67/96 (69.8%)	56/67 (83.6%)	29/96 (30.2%)	11/96 (11.5%)
Raw score selective	64/96 (66.7%)	53/64 (82.8%)	32/96 (33.3%)	11/96 (11.5%)
Margin selective	65/96 (67.7%)	54/65 (83.1%)	31/96 (32.3%)	11/96 (11.5%)

sampling baseline.

The single direct judge is a baseline based solely on forced responses; no selective diagnostic was used, so diagnostic figures are not relevant. Accuracy, excluding ties in direction, is almost unchanged across persona budgets, which means that looking at accuracy alone would make 8 or 16 personas appear competitive; the significant decrease is in diagnostic coverage. The 32-persona run roughly halves token usage compared to 64 personas, with only a 0.23 percentage-point shift in non-tie directional accuracy. The no-persona baseline is not independent sampling. It is more decisive under the interval diagnostic, but has lower forced accuracy, lower covered-decision accuracy, and a higher wrong-decisive rate than the 64-persona panel. This is consistent with personas adding signal here, but not with personas being a representative population sample or independent human-like judgments.

6.4. Model Family Sensitivity

Table 7 reports a 16-persona sensitivity check across three Gemini Flash-family conditions. In this run, Gemini 2.5 Flash-Lite had the highest non-tie accuracy, fastest execution time, and lowest wrong-decisive rate. The comparison is a sensitivity check rather than a general model recommendation, because it uses one persona budget, one benchmark, one seed, and one provider family.

7. Discussion

How Chorus differs from a typical synthetic A/B test. The difference is not about comparing options from a persona's perspective; SimAB already did something similar. Instead, Chorus treats synthetic groups as selective decision support. We ask whether they lower wrong decisive recommendations, point out disagreement, and flag difficult situations for another review or real-world validation.

What we can and cannot say. On this benchmark, abstention reduced wrong decisive recommendations while preserving substantial coverage. The remaining 11 wrong decisive recommendations are important: Chorus should route weak cases better, not be treated as a substitute for validation. Tier B exposes this weak-evidence regime especially clearly. It has coverage similar to Tier A, but lower covered-decision accuracy and a higher wrong-decisive rate. The benchmark does not establish causal lift, representative public opinion, legal correctness, demographic representativeness, native multi-way ranking, or family-wise risk control.

8. Limitations

Chorus is related to synthetic A/B testing, particularly SimAB, but this work does not include a direct empirical SimAB comparison. The contribution here is the selective decision behavior around personas: covered-case error, wrong-decisive rate, disagreement reporting, and routing unclear cases to validation.

The synthetic personas are not real respondents or representative population samples. They reflect model priors, possible stereotypes, and culture-specific assumptions, and persona wording can make weak evidence appear more persuasive than it is. The validation runs use English-language,

Table 5. Persona-count frontier on the 96-case validation set. Token totals are run-wide for the 108-case runner. Forced accuracy is computed among non-tie forced calls.

Judgments per case	Forced acc. (non-tie)	Forced wrong/all	Diagnostic coverage	Covered acc.	Wrong decisive / all	Tokens per run
64	75/96 (78.1%)	21/96 (21.9%)	67/96 (69.8%)	56/67 (83.6%)	11/96 (11.5%)	11.52M
32	74/95 (77.9%)	21/96 (21.9%)	63/96 (65.6%)	53/63 (84.1%)	10/96 (10.4%)	5.75M
16	71/90 (78.9%)	19/96 (19.8%)	46/96 (47.9%)	40/46 (87.0%)	6/96 (6.3%)	2.84M
8	69/89 (77.5%)	20/96 (20.8%)	36/96 (37.5%)	31/36 (86.1%)	5/96 (5.2%)	1.41M

Table 6. Persona-free ablations on the 96-case validation set. Token totals are from completed 96-case ablation runs. Forced accuracy is computed among non-tie forced calls.

Condition	Judgments per case	Forced acc. (non-tie)	Forced wrong/all	Diagnostic coverage	Covered acc.	Wrong decisive / all	Tokens per run
No-persona repeated	64	69/96 (71.9%)	27/96 (28.1%)	85/96 (88.5%)	68/85 (80.0%)	17/96 (17.7%)	9.56M
Single direct judge	1	66/93 (71.0%)	27/96 (28.1%)	n/a	n/a	n/a	0.15M

U.S.-market-oriented persona and audience framing; personas adapted for other countries have not been evaluated. The completed ablations compare against a single direct judge and a 64-call persona-free repeated baseline, but they do not show that disagreement among synthetic personas mirrors disagreement among humans.

The synthetic-vote diagnostic is a dispersion heuristic over synthetic votes, not a calibrated estimate of real-world accuracy. It does not estimate causal lift or provide an anytime-valid confidence sequence under dependent LLM outputs. The retrospective 5-fold cross-fitted threshold analysis does not provide prospective risk control, and 96 cases are not enough for fine-grained calibration. Public case-study outcomes may also carry leakage risk because the original cases might have appeared in model training data.

Prompt and run sensitivity. The order of the options is balanced, but the benchmark does not test paraphrased prompts, changed judging rules, a wider range of models, or different persona seeds. The model-family check only includes three 16-persona Gemini Flash-family conditions, and the persona-budget frontier does not estimate how much results change from run to run.

9. Ethics and Broader Impact

The main ethical risk is over-reliance. Because Chorus produces fast, fluent reports, users may mistake synthetic persona outputs for evidence about actual populations. Chorus should be used for reversible early-stage decisions such as revision, shortlisting, critique, or validation planning. It should not replace live A/B testing, user research, accessibility review, legal review, public consultation, or domain expertise. For consequential decisions, Chorus outputs should trigger expert review and real data collection rather than serve as proof of what people believe or want.

10. Conclusion

Chorus treats persona simulation as selective decision support rather than forced synthetic A/B prediction. In the 96-case validation set, forced winner accuracy hid a meaningful failure mode: 21/96 forced calls picked the wrong winner, while the synthetic-vote diagnostic reduced wrong decisive recommendations to 11/96 by abstaining on weak-signal cases. The result does not calibrate deployment risk, but it shows why systems should report coverage, abstention, disagreement, and wrong decisive decisions alongside winner accuracy. Chorus is most useful when strong synthetic signals guide early choices and weak signals trigger revision, human panels, or real experiments.

Table 7. Sixteen-persona Gemini Flash-family sensitivity check. Calls, tokens, and wall time are run-wide for the 108-case runner; accuracy, coverage, and wrong-decisive metrics use the 96-case validation subset. This is a sensitivity check, not a general model recommendation.

Model	Calls	Wall	Tokens	Non-tie acc.	Diag. cov.	Covered acc.	Wrong decis.	Raw cov.
2.5 Flash-Lite	1,728	1:37	2.84M	78.89%	47.92%	86.96%	6.25%	66.67%
2.5 Flash	1,728	2:47	2.87M	75.53%	50.00%	85.42%	7.29%	64.58%
3.1 Lite Preview	1,728	2:10	2.66M	75.56%	57.29%	85.45%	8.33%	68.75%

References

Amin, D., Salminen, J., Ahmed, F., Tervola, S. M. H., Sethi, S., and Jansen, B. J. Creating and evaluating personas using generative AI: A scoping review of 81 articles, 2025. URL https://arxiv.org/abs/2504.04927.

Castelo, A., Zanjani Foumani, Z., Fan, A., Koay, K. Y., Malik, V., Zhu, Y., Li, H., Feghhi, M., Uliana, R., Xie, S., Zhang, Z., Martins, A. O., Zhao, M., Pelland, F., Faerman, J., LeBlanc, N., Glazer, A., McNamara, A., Wang, L., and Wu, Z. SimGym: Traffic-grounded browser agents for offline A/B testing in e-commerce, 2026. URL https://arxiv.org/abs/2602.01443.

Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970.

Dawid, A. P. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate, 2023.

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

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know, 2022.

Kim, M., Im, S., Choi, J., Lee, J., Shim, C., Hong, H., and Choi, E. PICon: A multi-turn interrogation framework for evaluating persona agent consistency, 2026. URL https://arxiv.org/abs/2603.25620.

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

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words, 2022.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Lu, Y., Hsu, T.-Y., Gu, H., Cui, L., Xie, Y., Headden, W., Yao, B., Veeragouni, A., Liu, J., Nag, S., Wang, J., and Wang, D. AgentA/B: Automated and scalable web A/B testing with interactive LLM agents, 2025a. URL https://arxiv.org/abs/2504.09723.

Lu, Y., Huang, J., Han, Y., Yao, B., Bei, S., Gesi, J., Xie, Y., Sang, Y., Wang, Z., He, Q., and Wang, D. Can LLM agents simulate multi-turn human behavior? evidence from real online customer behavior data, 2025b. URL https://arxiv.org/abs/2503.20749.

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., and Wang, D. UXAgent: An LLM agent-based usability testing framework for web design, 2025c. URL https://arxiv.org/abs/2502.12561.

Morocho, E. E. T., Cima, L., Fagni, T., Avvenuti, M., and Cresci, S. Assessing the reliability of persona-conditioned LLMs as synthetic survey respondents, 2026. URL https://arxiv.org/abs/2602.18462.

Paglieri, D., Cross, L., Cunningham, W. A., Leibo, J. Z., and Vezhnevets, A. S. Persona generators: Generating diverse synthetic personas for arbitrary contexts, 2026. URL https://arxiv.org/abs/2602.03545.

Rieder, T., Schneider, M., Truss, M., Tsaplin, V., Rublea, A., Dere, S., Chicharro Sanz, F., Reiss, T., and Dogan, M. D. SimAB: Simulating A/B tests with persona-conditioned AI agents for rapid design evaluation, 2026. URL https://arxiv.org/abs/2603.01024.

Sun, L., Fu, S., Yao, B., Lu, Y., Li, W., Gu, H., Gesi, J., Huang, J., Luo, C., and Wang, D. LLM agent meets agentic AI: Can LLM agents simulate customers to evaluate agentic-AI-based shopping assistants?, 2025. URL https://arxiv.org/abs/2509.21501.

Truss, M. PersonaCite: VoC-grounded interviewable agentic synthetic AI personas for verifiable user and design research, 2026. URL https://arxiv.org/abs/2601.22288.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A. Benchmark Card and Tier Criteria Details

Appendix A provides additional details on tier interpretation and missing provenance fields for the benchmark card in Table 1. The 96-case validation set is a non-public offline validation set, not a public benchmark release or deployment calibration set. Most cases are reconstructed from publicly viewable A/B case studies, and six gaming-concept cases use human-panel artifacts available to the authors. In Table 1, "public" means that the reference outcome was publicly viewable; it does not mean that the original pages, screenshots, assets, or panel records are freely redistributable. For that reason, this submission reports aggregate metrics and suite-level metadata but does not release a case-level manifest or source assets. A public release would require source-by-source permission and redistribution review. The machine-readable artifacts record case identifiers, titles, suites, expected winners, outcome-source labels, tier labels, vote counts, abstention status, and aggregate diagnostics. They do not always systematically track source URLs, collection dates, human-panel protocols, effect sizes, or label-reliability estimates. Because of these missing details, we treat those fields as missing and do not infer them from the results.

The tier labels indicate how closely the benchmark cases reflect the outcome being measured. They should be read as descriptive subgroups, not as something decided in advance for calibration. Tier A represents the clearest and most direct connections between input and outcome, with quantitative outcomes and a close match to the original input. Tier B covers outcome-linked cases with lower fidelity, including multi-screen tests, text-only reconstructions, before/after sources, less specific pairwise metrics, or combined treatment descriptions. The tier split is useful because it shows where Chorus is more likely to be wrong, but it should not be used to say the abstention rule is calibrated within each tier.

Ideally, each case would include the source URL, collection date, how the outcome was found, the human-panel size and protocol when relevant, expected winner, effect size or vote margin when available, inclusion and exclusion criteria, tier rationale, and whether the materials were original or reconstructed. The artifacts available right now cover only part of that information, and this paper treats the missing provenance as a weakness. The validation set should therefore be read as an author-held stress-test set for Chorus, not as a reusable benchmark artifact.

B. Persona Count Benchmark Details

The main part of the paper uses the 64-persona run as the headline setting. Figure B1 shows how the decisions were made, Figure B2 shows how tiers behaved, Table B1 adds persona-count run details, and Figure B3 shows suite-level behavior.

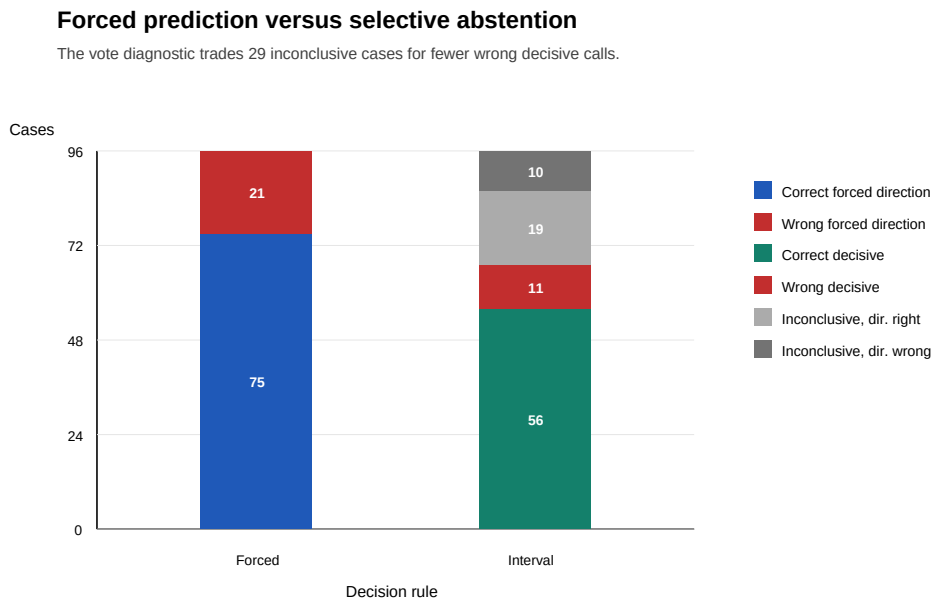

Figure B1. Breakdown of decisions for forced directional prediction and for the synthetic-vote diagnostic on the 96-case validation set. The abstention rule turns part of the area where the system would have been wrong into cases where it does not give an answer.

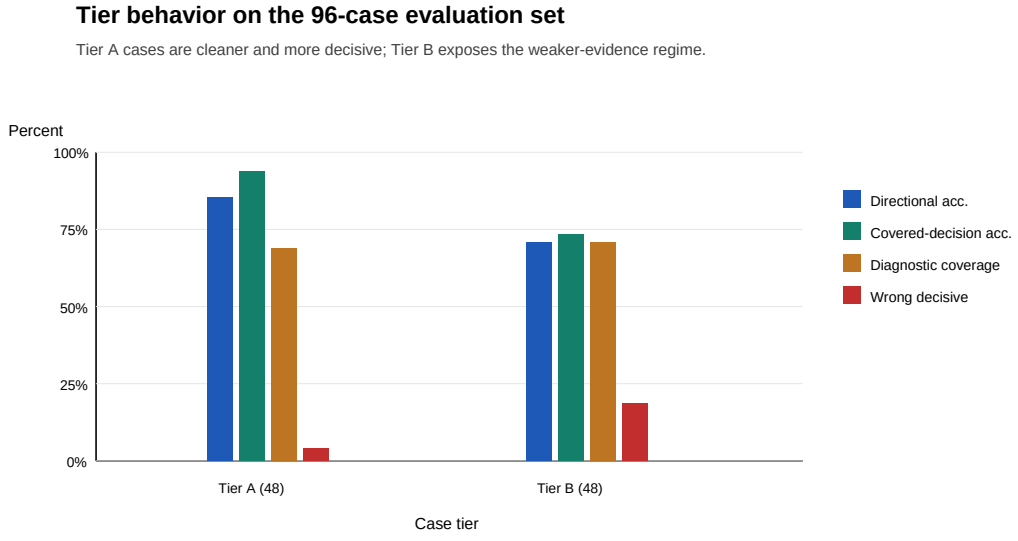

Figure B2. Tier-level benchmark information for the 64-persona run. Tier B has similar coverage, but lower accuracy when it does decide and a higher rate of wrong decisive recommendations.

Table B1. Persona-count frontier for Gemini 2.5 Flash-Lite. Calls, tokens, and wall time are run-wide for the 108-case runner; accuracy, coverage, and wrong-decisive metrics use the 96-case validation subset.

Metric	64p	32p	16p	8p
Run-wide judgment calls	6,912	3,456	1,728	864
Wall time	6:37:03	3:39:15	1:36:44	0:48:19
Total tokens	11.52M	5.75M	2.84M	1.41M
Non-tie directional coverage	100.00%	98.96%	93.75%	92.71%
Non-tie directional acc.	78.12%	77.89%	78.89%	77.53%
Diagnostic coverage	69.79%	65.62%	47.92%	37.50%
Covered-decision acc.	83.58%	84.13%	86.96%	86.11%
Wrong-decisive rate	11.46%	10.42%	6.25%	5.21%
Margin coverage	67.71%	61.46%	38.54%	0.00%
Raw-score coverage	66.67%	67.71%	66.67%	73.96%

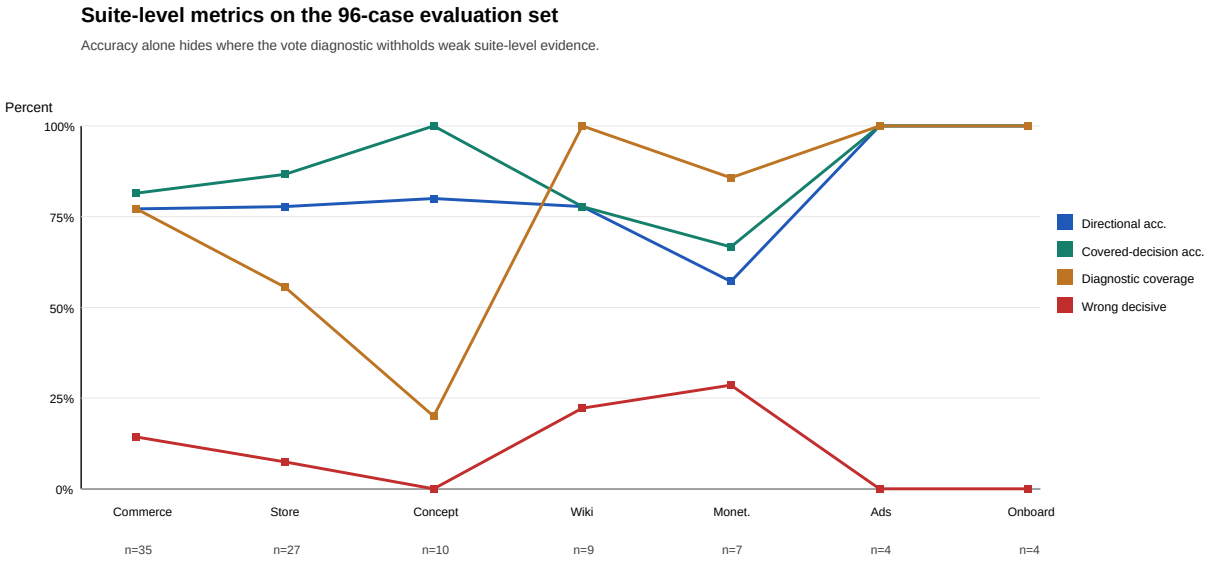

Figure B3. Suite-level benchmark information on the 96-case validation set. Small suites should be read as descriptive smoke checks, not as stable domain-level estimates.

The suite-level descriptive table is reported in the main paper as Table 3; it is not repeated here to avoid duplicating the same evidence.

C. Interface Screenshots

Figures C1, C2, C3, and C4 show the current Chorus product interface used to set up simulations and inspect reports. These are interface examples, not additional benchmark evidence. Terms such as "respondent," "confidence," or "feedback," where visible, refer to synthetic persona outputs and self-reported model labels, not real people or calibrated probabilities. Report-summary screenshots that display legacy formal-inference labels are omitted because the benchmark does not treat those labels as calibrated population inference.

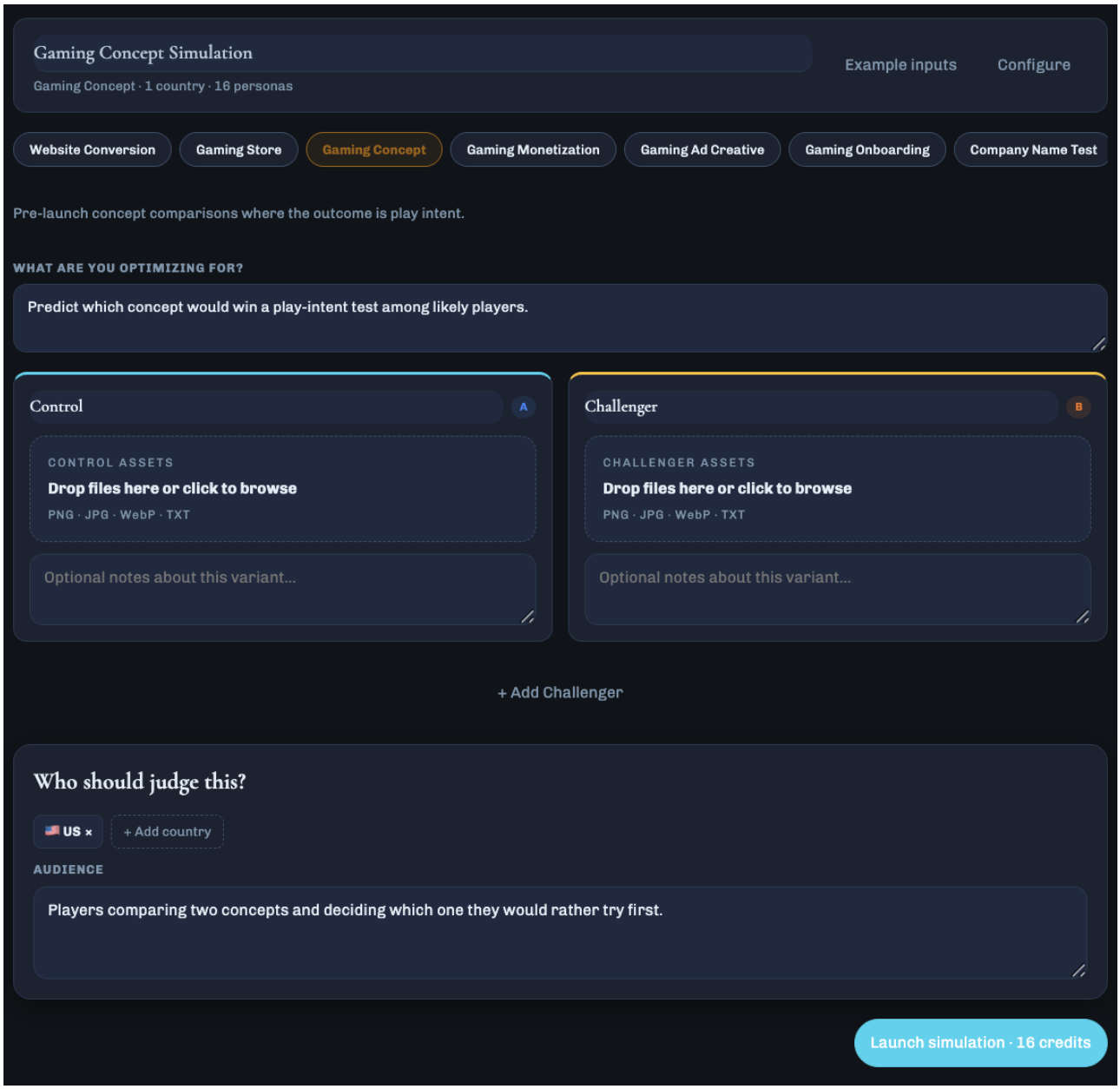

Figure C1. Example Chorus simulation setup screen. Before starting a batched synthetic panel, the interface asks for the decision type, optimization goal, control and challenger assets, optional additional challengers, audience definition, countries, and persona count.

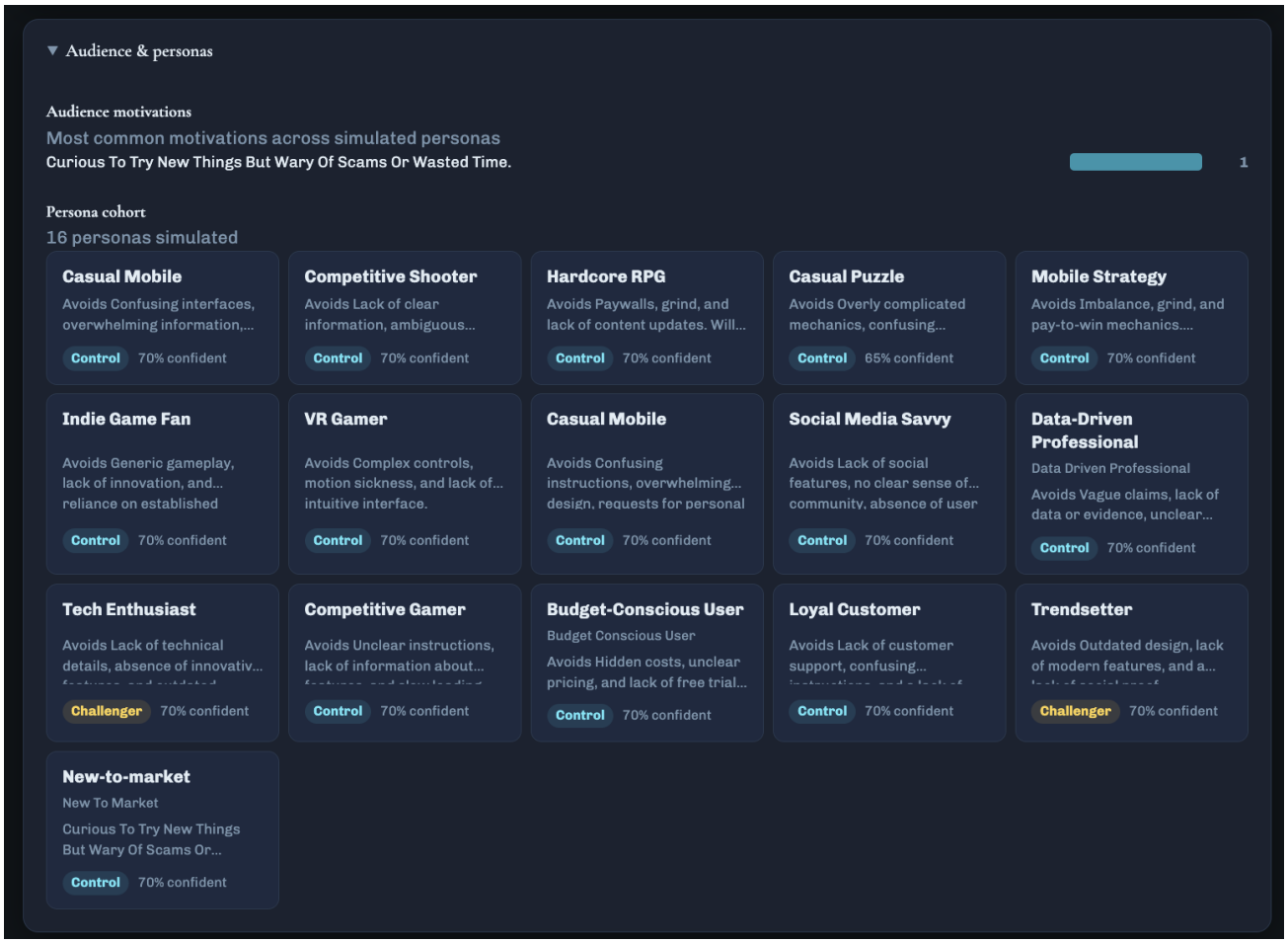

Figure C2. Audience and persona section showing the simulated persona cohort, segment labels, persona-level decisions, and persona self-reported signal labels. These labels are not calibrated probabilities.

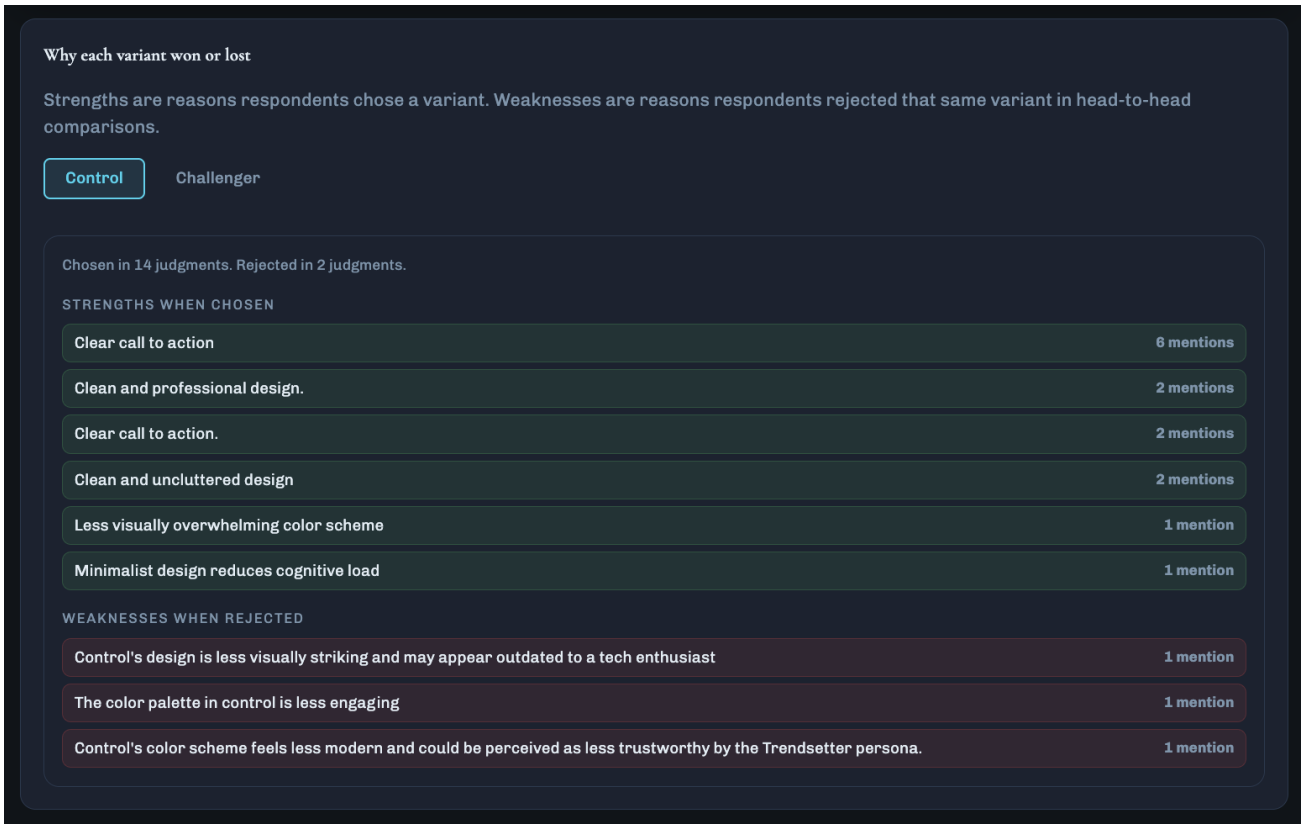

Figure C3. Variant evidence section clustering reasons why synthetic personas chose or rejected each variant in persona-level head-to-head comparisons.

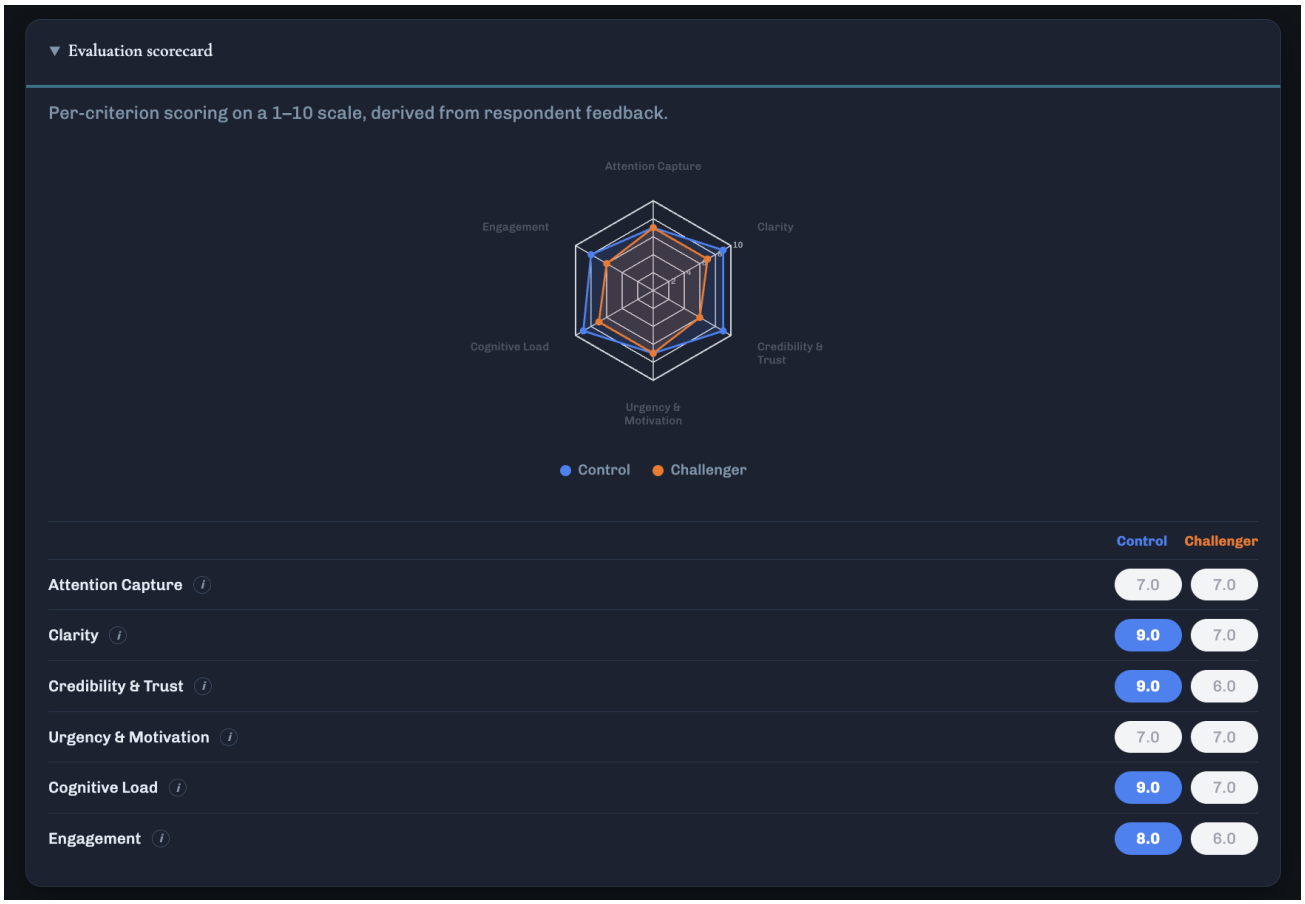

Figure C4. Evaluation scorecard section summarizing criterion-level scores derived from synthetic persona feedback.