Rehearsals
Replica Performance

LAST UPDATED	AUTHOR	CONTRIBUTORS
February 20, 2026	Brian Oppenheim	Liam Bolling Jess Lee

TABLE OF CONTENTS

0.0 Glossary

TERM	DEFINITION
Frontier Language Model/AI Model	(Chat)GPT, specifically 5.2. Gemini 3.0. Claude 4 Sonnet/Opus. The current “best” that the big 3 AI labs can do.
Synthetic Persona	There is no set definition, but PwC defines it as a high-level simulation of a customer segment, generated from generalized quantitative outcome data (aggregate

TERM	DEFINITION
	purchases, generalized credit scores, etc). Essentially, asking an LLM to roleplay as an individual or stereotyped identity. See associated section.
Replica	An AI representation of a specific, real-world individual, built from structured behavioral interviews, as well as demographics and psychographics that we have extracted. Unlike synthetic personas that simulate demographic archetypes, a replica encodes one person's actual decision-making patterns, constraints, values, and reasoning style. Powered by the Engram Model.
Engram Model	Rehearsals' proprietary AI model for replicating individual human behavior. Unlike general-purpose LLMs prompted to roleplay personas, Engram is fine-tuned specifically for behavioral fidelity, encoding individual decision patterns from structured interviews, and combining it with real world knowledge and context.
Simulation	The process of presenting replicas with a scenario (e.g., a price change, ad creative, or product choice), an environment (macro-economic conditions, product research from social media, local news in that replica's area, etc) and collecting their responses.
Van Westendorp Pricing Meter	Instead of asking "how much would you pay for this item?", asking four questions about how much is too cheap/a good deal/getting expensive/too expensive. The range between "a good deal" and "getting expensive" is often known as the "range of acceptable payment".
Pricing Test-Retest Reliability	Asking the same pricing question to the same person over a period of two moments in time. Studies show an immense amount of variance, with correlation between the two answers ranging from r=0.42-0.78 depending on the type of good.
Accuracy (1 - MAE)	100% minus the Mean Absolute Error across distribution categories. Higher is better. 92% accuracy means predictions averaged 8 percentage points off from ground truth across all categories.
MAE	Mean Absolute Error—the average absolute difference between predicted and actual percentages. Lower is better. MAE of 8% means predictions were off by 8 percentage points on average.
sMAPE	Symmetric Mean Absolute Percentage Error—measures prediction error as a percentage, symmetric to over/under-prediction. Lower is better. 25% sMAPE means predictions were ~25% off in relative terms. Used often in the field of forecasting and for price predictions where absolute dollars vary widely.

1.0 Overview

Rehearsals creates AI replicas of human beings by interviewing real people with questions backed by UX researchers and behavioral economists to extract core decision-making processes.

We provide the cutting edge of decision science in a diverse set of fields. For example, you can ask replicas across many common marketing focused topics including advertising content preferences, pricing elasticity outcomes and digital product experience churn. Simulating

outcomes isn't new but the level of accuracy our replicas exhibit compared to ground truth is the rationale for publishing this document.

- Replicas are able to predict winning social media content and video advertising content **over 30% better than frontier AI models.**
- Replicas are as **accurate as their human counterparts** in pricing test-retest and WTP (Willingness-to-pay) scenarios such as the Van Westendorp Price Sensitivity Meter.
- **Replicas correctly (> 90% accuracy) model real-life product and decision distributions**. Ask how many of the general US population took a flight in the previous 365 days or how many of the US population signed up for a new video streaming service. Replicas return a ground truth accurate answer and will correctly predict the churn rate from price increases within both of those industries. These results are nearly identical with case studies that have occurred in the real world.

Unlike generic large language models that regress toward mean responses, Rehearsals replicas preserve the authentic diversity of consumer preferences, capturing long-tail behaviors and non-obvious decision drivers that make the difference between product success and failure.

1.1 Performance at a Glance

CLAIM	PROOF
Our replicas accurately recreate **true taste** and **intangible human preference**	When tested on pairs of nearly identical advertisements and short form video, our model predicts the one with greater reach **89% of the time**, compared to frontier AI at **65% (McNemar's p < 0.002).**
Our replicas have **human-level precision** at determining a consumer's willingness-to-pay across **hundreds of products**	Our replicas can guess what their human counterparts think is cheap/fair price/expensive with **25% sMAPE**. This is comparable to test-retest reliability of asking the same person twice.
Our replicas accurately predict real-world consumer distributions and market responses	When backtested on historical examples like streaming service changes, flight price hikes, and iPhone sales, our model matches actual consumer preferences on products and pricing with over **91% accuracy**.

2.0 Why Replicas?

2.1 Induction Vs. Deduction

Most approaches to predicting consumer behavior use deductive reasoning: starting from general principles about "rational consumers" and applying logical inference. This systematically fails because real consumers don't behave like theoretical averages. **Even when you break people down into personas or other forms of bucketing, this is still just a form of theoretical averaging, inclusive of all the jumps in logic that come with it.**

Deductive Approach (Frontier LLMs, Traditional Models):

- **Reasoning:** Consider when Disney+ raised prices by 3 dollars on subscribers in 2022, with the alternative of an ad-supported tier. A frontier language model might reason: "A rational consumer facing a price increase would evaluate the cost-benefit of each alternative. Some price-sensitive users will switch to the ad-supported tier to maintain their current budget. Others will cancel if the value proposition no longer justifies the cost. Brand-loyal users with children will likely accept the increase..."
- **Result:** Homogeneous responses that smooth toward theoretical averages (e.g., 38% accept / 45% switch / 17% cancel)
- **Misses:** Status quo bias, mental accounting, individual constraints, emotional drivers, revealed preferences...

Inductive Approach (Rehearsals Replicas):

- **Reasoning:** Each replica decides based on their specific life context: "Given my three kids who watch Disney daily, the hassle of canceling, and the fact that $3 is less than my morning coffee..."
- **Result:** Authentic distributions that preserve real-world clustering (e.g., 87% accept / 9% switch / 4% cancel — more closely matching the actual 94%/5%/<1%)
- **Captures:** Behavioral economics effects, heterogeneous preferences, context-dependent decision-making

2.2 Why Alternatives Fall Short

Frontier LLMs:

- Use deductive reasoning from first principles and aggregate training data
- Regress toward mean responses, failing to capture authentic distribution diversity

Synthetic Personas:

- Rely on demographic stereotypical outcomes rather than individual behavior.
- Cannot explain **why** they hold a preference, making them trivially persuadable under follow-up questioning and unreliable in simulated negotiations or trade-off scenarios.
- Miss long-tail and edge-case preferences because they regress toward aggregate patterns in training data.
- Are grounded in observed outcomes that lack the environment and reasoning to know how that observed outcome came about.

	EXAMPLE: COHORT SYNTHETIC PERSONA	EXAMPLE: INDIVIDUAL SYNTHETIC PERSONAS
Typical Input	"You are a male from Utah, who has a job in tech and makes over $100,000 with 2 children"	"You are John Doe, these are posts from your recent LinkedIn. Pretend to be them."
Outcome	This can only ever generate high level generalizations and stereotypes of their demographic.	"John Doe" does not know why or how these posts were made.

Traditional Research with Human Interviews:

- **Expensive**. Interviewing 100 people via traditional market research avenues has a floor of ~$20,000, ranging to high six figures for full-service consultants.
- **Time-consuming.** (4-12 weeks from design to results) between planning study questions, recruiting a panel of participants, interviewing, calculating results, re-testing and final presentation of output.
- **Limited causal inference:** Traditional interview-based research lacks counterfactual controls — you can't A/B test a pricing strategy or tariff scenario against a control group in a live interview. External variables (mood, framing, interviewer effects) are uncontrolled, and there's no way to isolate what actually drove a stated preference.
- **Lack of Scale and Repeatability:** Often limited to hundreds of participants for cost reasons and oftentimes unable to repeat or iterate on the study (participant bias) in the same environment.
- **Most importantly, it is subject to response bias, social desirability effects, and other psychological traps.**
    - **Stated purchase intent explains only 5–15% of the variance in actual CPG buying behavior** (Morwitz et al., 2007). Consumers routinely overstate intent by 2–5x relative to actual trial rates (Jamieson & Bass, 1989), and the mere act of surveying them inflates the observed correlation — meaning even validation studies overstate how well surveys perform (Chandon, Morwitz & Reinartz, 2005).
    - **The remaining 85–95% of behavioral variance is driven by factors invisible to stated-preference methods**: habitual repertoire switching, shelf placement, package salience, and subconscious brand availability. The average consumer spends ~13 seconds choosing a brand in-store (Ehrenberg-Bass, 2018). This is a structural limitation of asking people to introspect on decisions they don't consciously make.

2.3 Our Approach

Rehearsals builds replicas through inductive modeling - starting with individual cases and letting patterns emerge:

1. **Deep Individual Interviews:** Structured conversations exploring decision-making processes, values, constraints, and preferences
2. **Behavioral Economics Framework:** Questions designed to reveal actual decision drivers, not just stated preferences
3. **Demographic Calibration:** Replicas represent authentic market segments matching U.S. census distributions
4. **Continuous Validation:** Regular testing against real-world outcomes

Once built, you can ask replicas any question in the style of a moderated research interview. Unlike deductive models that reason from theory, each replica reasons from their authentic context — and when aggregated, this preserves the non-normal, "spikey" distributions found in real consumer behavior.

3.0 Methodology

3.1 Introducing The Engram Model

Rehearsals replicas are powered by **Engram 1.0** - a model custom-built to replicate individual human decisions, feelings, and actions. Unlike synthetic persona approaches that prompt a general-purpose LLM to "roleplay" as a demographic profile, Engram combines two distinct components:

Individual Human Encoding
Each replica is built from rich, multi-modal data about a specific person in the real world like unstructured voice and video interviews, social media content, and structured demographic data. This captures not just *what* someone would choose, but *how* they reason through decisions.

Environmental Adapters
People don't exist in a vacuum, and replicas shouldn't either. Engram dynamically adapts each individual's encoded patterns to real-world environmental context - local news and social sentiment, macroeconomic conditions, nearby activities, and actual product research that a human might do. A synthetic persona asked "would you fly to Miami next week?" has no idea there's a hurricane warning or that gas prices just spiked. Engram-powered replicas adapt to context the same way real humans do.

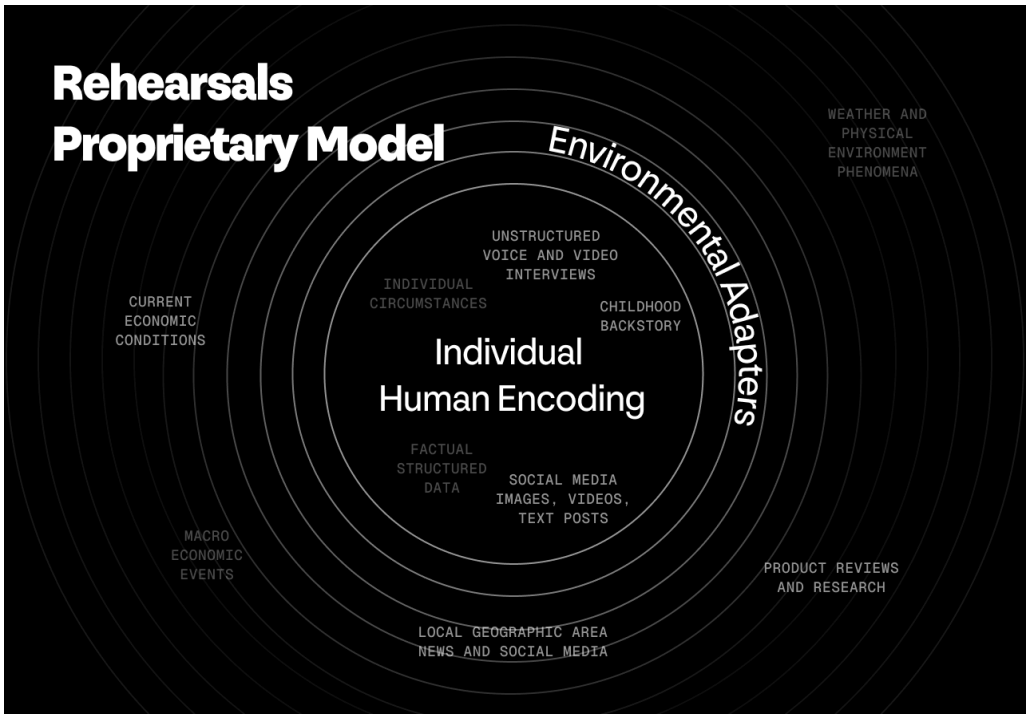

FIGURE 1

Our proprietary model utilizes both in-depth human data and a complex layer of environmental world knowledge in order to achieve best-in-class prediction on a vast array of the hardest questions that companies face today.

Why this matters:
Generic LLMs retrieve aggregate patterns and smooth toward theoretical averages. Engram encodes *individuals* and adapts them to *environments* - producing the authentic, spiky distributions that match real-world behavior.

3.2 Replica Construction

Each Rehearsals replica is created through a multi-stage process:

Stage 1: Participant Recruitment
We recruit participants matching target demographic profiles using stratified sampling to ensure representation accurate to the 2020 Census across age, household income, geography, education levels and employment status.

Stage 2: Structured Interview
Each participant completes an in-depth audio and video interview or series of interviews covering a broad set of topics, including:

- Product category usage and preferences
- Purchase decision processes and constraints
- Price sensitivity and value perception
- Media consumption and influence patterns
- Life context, priorities, and trade-offs

Stage 3: Replica Calibration
Interview responses are processed to create an individual AI replica model that:

- Maintains the individual's authentic decision-making patterns
- Preserves their stated constraints and priorities
- Reflects their actual behavioral tendencies (not idealized responses)
- Captures their unique voice and reasoning style
- Reflects their life, upbringing, rationale for left and right brain decisions making

Stage 4: Individual Validation
We evaluate accuracy on an individual replica basis by holding out sections of a user's interview as a hidden test-set, and using their replica to accurately recreate their responses to questions.

For instance, if a user said they would pay between 40 and 60 dollars for a pair of jeans in their interview, we ensure that we get our replicas to say a similar price range without seeing that information first.

3.3 Study Design

What We Simulated:
Our validation studies span three critical areas of consumer research: pricing decisions (streaming subscriptions, flight purchases, and Willingness-To-Pay scenarios), product preferences (smartphone selection), and advertising effectiveness (social media content performance). These

scenarios were chosen because they represent common, high-stakes business decisions where companies invest millions based on predictions of consumer behavior. Each scenario had publicly verifiable ground truth outcomes, allowing us to measure replica accuracy against real-world results rather than theoretical benchmarks.

Sample Size and Demographic Matching:
Unless otherwise specified, each validation study was completed with four arms of (N=75) unique replicas weighted according to US census data to create a representative US general population cohort across age, income, education, and geography. The results from the four runs are averaged and returned in the figures below.

Replica Evaluation:
Like a real human interview, asking the right questions is important to getting high quality results. For each scenario, the replica sees the scenario in its context window, and is engaged by an AI researcher that asks an associated set of questions in order to tease out true preference and behavioral data.

Baseline Comparisons:
In order to estimate a baseline for replica performance, we ran identical scenarios for each scenario by current SOTA language models like Gemini 3.0 and GPT-5.2 and either ask it for the most likely outcome (if the task is majority vote), or to give a breakdown of what percent of the population would vote for each outcome (if the task is to model a distribution). This is to get a sense of what a key stakeholder might expect when using AI to answer their questions right now.

We used unmodified model parameters (temperature, top-K, etc.) in order to ensure parity with the experience that an end user would have with alternatives to Rehearsals.

A Note on Knowledge Cutoffs:
The ground truth for some scenarios we show is likely contained in the knowledge cutoff for LLMs. We do attempt to remove this information from the model via prompt rewinding, which is over 80% accurate at removing factual information. Because our replicas are embodied on what an individual would do ("what would you as John Smith do if Disney+ raised your prices?"), not on what actually happened in the real world ("how many users churned when Disney+ raised their prices?"), our replicas get no benefit. **Despite this inherent unfair condition, we still reliably outperform frontier models in prediction, even with a potential knowledge cutoff advantage in the baseline.**

A note on statistical methods
All p-values were corrected for multiple comparisons using Benjamini-Hochberg applied within test families (accuracy tests, distribution tests, screening tests, and head-to-head comparisons) to control the false discovery rate at $\alpha = 0.05$. Head-to-head comparisons between methods used McNemar's test on paired scenario outcomes (e.g., whether each method correctly identified the winning ad across the same set of scenarios). Distribution accuracy was evaluated using Mean Absolute Error (MAE) with bootstrap confidence intervals (10,000 resamples).

As for Gemini and ChatGPT baselines, they produce a single aggregate prediction per scenario (predicted population-level percentages) rather than individual-level responses, precluding bootstrap resampling. Their values are reported as point estimates.

Ground Truth Sources:

- Publicly reported market outcomes (Disney+ subscriber behavior)
- Published industry research (IATA/InterVISTAS airline demand studies)
- Observed social media performance metrics (Instagram/TikTok account advertising dashboards, verified Jan 31, 2026)
- Retail sales distributions (Apple iPhone market share data, CIRP Q4 2025)

4.0 Product Pricing

4.1 Video Streaming Service Subscription Churn

Scenario description: Disney+ launched an ad-supported tier and simultaneously raised the price of the ad-free tier by $3 in December of 2022. Existing ad-free subscribers receive the notification to pay $3 additional per month, switch to the ad tier and keep the same price, or cancel altogether.

This scenario is a critical test of replica accuracy because it reveals the gap between how consumers *say* they'll behave and how they *actually* behave. A 37% price increase ($7.99 → $10.99) sounds significant, and rational economic models would predict meaningful churn. In reality, the overwhelming majority of subscribers simply absorbed the cost increase—driven by status quo bias, the hassle of canceling, and the sunk cost of existing watchlists and viewing history. This is precisely the kind of behavioral economics effect that deductive models systematically miss.

Step 1: Population Screening

Before asking *what* consumers will do, we must first identify *who* should answer. Predicting subscription churn is meaningless if your model surveys the wrong population. We asked each replica and persona: **"Do you personally pay for a Disney+ subscription?".**

In the case of the LLM baselines like Gemini and GPT, we simply asked them **“what percent of the US population of adults pays for Disney+?”**. Surprisingly, these baselines had very divergent (and incorrect) approximations.

GROUND TRUTH
~23% of US adults personally pay for a subscription to Disney+.

METHOD	PREDICTED QUALIFICATION RATE	GAP FROM GROUND TRUTH
Rehearsals Replicas	**18.7%**	**4.3%**
GPT-5.2	16%	7.0%
Gemini 3.0	38%	15.0%
Synthetic Personas	76%	53.0%

95th percentile Confidence Intervals:
Rehearsals Replicas – [13.8, 24.0]; Synthetic Personas – [70.2, 81.3]

Why this matters: Synthetic personas catastrophically over-qualify respondents, predicting 76% subscription rates when only 23% of real people subscribe. Gemini over-qualifies by 15%. Even GPT-5.2, which screens reasonably well, will compound any screening error through to distribution predictions. Replicas screen within 4.3% of ground truth, ensuring the simulated population matches reality before any behavioral prediction begins.

Step 2: Distribution Prediction

Among respondents who pass the screener, what do they actually choose?

GROUND TRUTH OUTCOME		
Took the price increase to $10.99/month (+$3 delta), did not churn from the service	**Canceled their plan altogether and stopped paying $7.99/month**	**Switched to the ad-supported tier and maintained their $7.99/month plan**
94%	5%	<1%

INFORMATION PROVIDED TO REPLICAS:
Current Date: December of 2022. You are a subscriber to Disney+. Disney+ is a major subscription-based streaming service from The Walt Disney Company that serves as the exclusive home for movies and shows from Disney, Pixar, Marvel, Star Wars, and National Geographic. It features a vast library of classics, new blockbusters, and original content, with optional bundles that include Hulu and ESPN+. You might use the service frequently or you might not. But you currently pay $7.99 per month for this service. You've just received an email from Disney+ announcing that they are changing their pricing and you have a few options: A: Maintain your current plan for a price increase of $10.99 per month which is $3 more per month. B: Accept a new plan which is the same price as what you are paying now but includes occasional ads while watching content. C: Cancel your subscription to Disney+ altogether and stop paying $7.99 per month and lose access to Disney and Hulu streaming content.

QUESTIONS PROVIDED TO REPLICAS:
1. Do you personally pay for a Disney+ subscription? If not, you can skip the remaining questions. 2. What are your initial thoughts and emotions around this price increase and plan change? 3. Do you think the product provides value? And how often do you watch the content provided by Disney+? 4. What choice will you make? A: Accepting the price increase, B: Taking on a new plan that has advertisement video advertisements while watching content occasionally, C: Canceling your subscription to Disney+ altogether

OUTCOMES				
METHOD	**ACCEPT PRICE INCREASE (A)**	**SWITCH TO ADS (B)**	**CANCEL SUBSCRIPTION (C)**	**ACCURACY (1 - MAE)**
Ground Truth	94%	<1%	5%	-
Rehearsals Replicas	**84%**	**14%**	**2%**	**91%**

OUTCOMES				
METHOD	**ACCEPT PRICE INCREASE (A)**	**SWITCH TO ADS (B)**	**CANCEL SUBSCRIPTION (C)**	**ACCURACY (1 - MAE)**
Synthetic Personas	88%	12%	0%	92%
Gemini 3.0	58%	28%	14%	72%
GPT-5.2	38%	45%	17%	63%

Accuracy 95th percentile Confidence Intervals
Rehearsals Replicas – [83.2, 97.5]; Synthetic Personas – [89.7, 96.2]

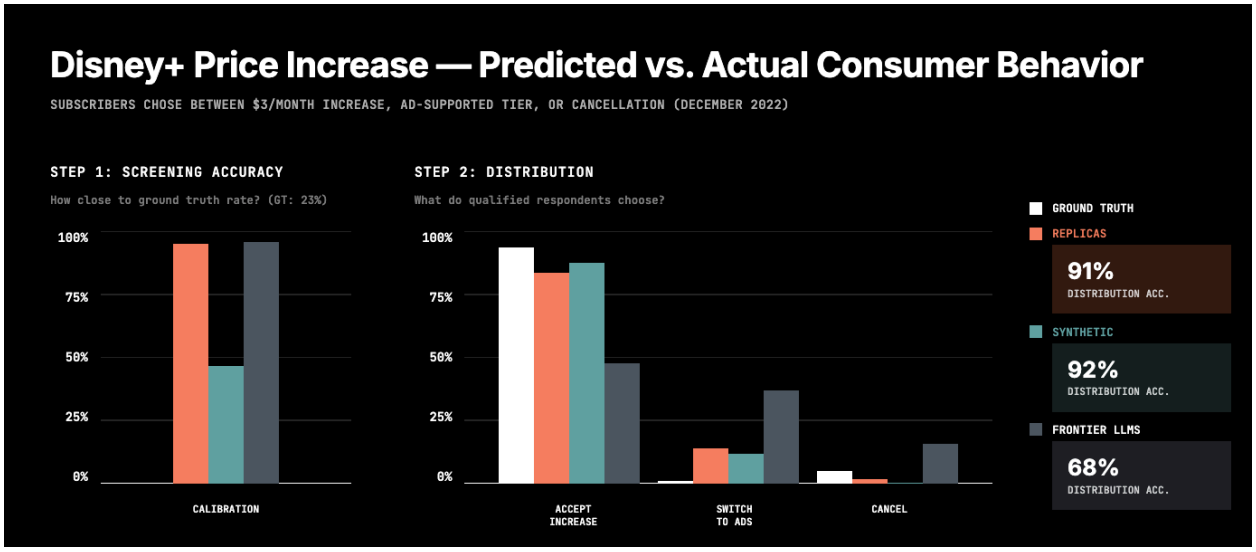

FIGURE 2
Disney+ price increase scenario (December 2022). Frontier LLMs predicted as low as 38% acceptance vs. 94% actual—a fundamental misunderstanding of consumer inertia. Synthetic personas match the distribution but grossly over-qualify respondents (±53pp screening gap). Rehearsals replicas achieve 91% distribution accuracy with correct population screening (±4.3pp).

Notes/Observations: Replicas achieved 91.1% distribution accuracy. While every method overestimates the ad-supported tier, replicas come the closest to capturing the overwhelming status quo bias. Gemini predicted only 58% acceptance—badly underestimating consumer inertia. GPT-5.2 was even worse, predicting just 38% acceptance. Synthetic personas achieved comparable distribution accuracy but only among a fantasy population—their 76% qualification rate means they surveyed 3x more "subscribers" than actually exist in the wild.

A model must get *both* steps right. Synthetic personas nail distribution but catastrophically fail screening. GPT-5.2 screens reasonably but completely misses how subscribers actually behave. Gemini fails at both. Only Replicas pass both tests—screening within 4.3% of ground truth and achieving 91% distribution accuracy among the correctly identified population.

EXAMPLE LLM REASONING (DEMONSTRATING CENTRAL TENDENCY BIAS):

```
"Disney+ possesses strong brand loyalty and 'must-have' status for households with children,
leading to a high percentage of users accepting the price increase to avoid the friction of
ads. The ad-supported tier captures price-sensitive users who want to maintain access to the
library without increasing their monthly budget, while the cancellation rate reflects
low-frequency users and those protesting the significant 37% price hike."
```

This reasoning sounds plausible but leads to predictions that fundamentally misunderstand actual consumer behavior. The LLM predicts roughly even splits across options (38%/45%/17%), suggesting rational evaluation of alternatives. In reality, most consumers simply paid the increase—driven by inertia, sunk cost fallacy, and the hassle of switching, not rational cost-benefit analysis.

EXAMPLE REPLICA REASONING
(DEMONSTRATING CORRECT CONTEXTUALIZATION OF THE COST INCREASE EVEN WITH A BUDGET CONSCIOUS PERSONALITY):

```
"Oh man, that's a tough one because I am a total budget fanatic. I really don't like seeing
prices go up, especially with how hard my parents worked when we were starting out. But
honestly? I'd have to go with Option A. With 3 kids, trying to sit through ads during our
family movie nights would just be a nightmare. They'd lose interest so fast! I'm all about
quality, and to me, a seamless, ad-free experience for the young ones is worth the extra
three bucks. My husband and I are really working on that nest egg so we can retire in five
years, but $36 a year isn't going to stop us from hitting our goals."
```

4.2 Short Haul and Aggregate Market Flight Price Increases

Scenario description:
Every year, IATA/InterVISTAS Release literature and research on a strong correlation between price and demand when it comes to classic categories of airline ticket prices. Here we determine what level of price elasticity our replicas have compared to ground truth with percent increases in classic airline products.

Price elasticity in air travel is one of the most studied phenomena in transportation economics, with decades of empirical data establishing clear demand curves for different travel contexts. This makes it an ideal benchmark: we have robust ground truth, the scenarios are easily understood by replicas, and the results reveal whether models can distinguish between different consumer contexts (discretionary leisure vs. committed travel). **Despite having likely exposure to these patterns in their training data,** baseline LLMs consistently overestimate how "rational" travelers are—predicting that a $20 increase will cause 20-28% of travelers to change plans, when actual behavior shows much stronger commitment to booked travel.

Step 1: Population Screening

Whether due to price, accessibility, or family, much of the United States does not fly. We asked each replica and persona: **"Have you flown in the last year?"**

In the case of the LLM baselines like Gemini and GPT, we simply asked them **"what percent of the US population of adults has flown in the last year?".** The baseline LLMs perform well on this, again likely due to the answer being in-distribution of their training data.

GROUND TRUTH
~50% of US adults have flown in the last year.

METHOD	PREDICTED QUALIFICATION RATE	GAP FROM GROUND TRUTH
Rehearsals Replicas	**50.2%**	**0.2%**
GPT-5.2	46%	4%
Gemini 3.0	48%	2%
Synthetic Personas	71.1%	21.1%

95th percentile Confidence Intervals:
Rehearsals Replicas – [43.6, 56.4]; Synthetic Personas – [64.9, 76.9]

Why this matters: Once again synthetic personas greatly over-qualify respondents, predicting flight rates over 20% higher than what actually occurs in the world. Replicas are essentially indistinguishable from the ground truth population (note: we provide no prior information about flight patterns to the replica) and the LLM baselines are not far behind.

Step 2: Distribution Prediction

Among respondents who pass the screener, what do they actually choose?

GROUND TRUTH OUTCOMES				
OFFICIAL SCENARIO TYPE	EXAMPLE	BASE FARE → NEW FARE	ELASTICITY BAND (INDUSTRY)	EXPECTED DEMAND CHANGE (GROUND TRUTH)
Short-haul leisure	Economy class New York to Miami booked 2-6 weeks out for spring break. This is usually consumers who are pricing for discretionary reasons like vacation or visiting friends. They can change dates, compete with airlines, potentially could drive or use other modes of	$200 → $220 (+10%)	–1.0 to –1.6	–10% to –16% demand

GROUND TRUTH OUTCOMES				
OFFICIAL SCENARIO TYPE	**EXAMPLE**	**BASE FARE → NEW FARE**	**ELASTICITY BAND (INDUSTRY)**	**EXPECTED DEMAND CHANGE (GROUND TRUTH)**
	transportation, or even skip the trip. Price is proportionally more important here			
Aggregate market	Chicago to New York City in economy for a work trip. San Francisco to Seattle for personal reasons. Price matters but not as much as above. And travelers have no other options but to fly. Price increases reduce demand but not as sharply as pure leisure. Family related sometimes.	$200 → $220 (+10%)	–0.6 to –1.2	–6% to –12% demand

INFORMATION PROVIDED TO REPLICAS	
LEISURE SCENARIO	**AGGREGATE MARKET SCENARIO**
Consider a scenario where you're booking an economy class flight from New York to Miami, 2-6 weeks out for spring break.	Consider a scenario where you need to book a flight from Chicago to New York City in economy class.
This is discretionary travel - planning a vacation or visiting friends. You have flexibility with dates, could compare airlines, potentially drive instead, or even skip the trip entirely.	This could be for work or personal (non-leisure) reasons. You could compare airlines or potentially drive instead.
The base fare is around $200 for this route. However, prices have just increased by 10% to $220.	The base fare is around $200 for this route. However, prices have just increased by 10% to $220.

QUESTIONS PROVIDED TO REPLICAS
1. Have you flown on a commercial flight in the past year? If you haven't flown recently, let me know and you can skip the remaining questions. 2. How does a $20 increase (10%) affect your decision to take a trip like this? 3. What alternatives would you consider if you decide not to fly? 4. Will you still book this flight at $220, or will you skip/postpone/find alternatives?

OUTCOMES				
AGGREGATE MARKET FLIGHT (+10% / $20 INCREASE)				
METHOD	**% WHO WOULD STILL BOOK**	**% WHO WOULD SKIP/POSTPONE**	**ACCURACY (PERCENT DIFFERENCE)**	**WITHIN GROUND TRUTH RANGE?**
Ground Truth	91%	9%	–	✓ (–6% to –12%)
Rehearsals Replicas	**93%**	**7%**	**98%**	✔
Synthetic Personas	99%	1%	92%	×
GPT-5.2	78%	22%	87%	×
Gemini 3.0	88%	12%	97%	✓

Accuracy 95th percentile Confidence Intervals
Rehearsals Replicas – [93.7, 99.9]; Synthetic Personas – [91.0, 93.1]

OUTCOMES				
SHORT-HAUL LEISURE FLIGHT (+10% / $20 INCREASE)				
METHOD	**% WHO WOULD STILL BOOK**	**% WHO WOULD SKIP/POSTPONE**	**ACCURACY (PERCENT DIFFERENCE)**	**WITHIN GROUND TRUTH RANGE?**
Ground Truth	87%	13%	–	✓ (–10% to –16% change in demand)
Rehearsals Replicas	71%	29%	84%	×
Synthetic Personas`	98%	2%	89%	×
GPT-5.2	72%	28%	85%	×
Gemini 3.0	78%	22%	91%	×

Accuracy 95th percentile Confidence Intervals
Rehearsals Replicas – [75.1, 92.6]; Synthetic Personas – [87.0, 91.5]

FIGURE 3

Aggregate market flight price sensitivity (+10% fare increase). Replicas achieved near-perfect screening (50.2% vs 50% ground truth—a 0.2pp gap) and 98% distribution accuracy. Synthetic personas over-qualified by 21pp, surveying a population that flies 40% more than reality.

EXAMPLE LLM BASELINE REASONING
"Among recent flyers, a $20 (10%) increase on a common $200 domestic route is noticeable but usually not large enough to change behavior for many trips—especially work travel, fixed-date visits, or time-sensitive plans. However, a meaningful minority will react by postponing, switching to alternate transport (train/bus/car), choosing different dates/airports, or skipping discretionary trips, producing an estimated ~1 in 5 opting out at $220."

EXAMPLE REPLICA REASONING
"It really depends on the 'why' behind the trip. If I'm heading back home to New York to see family—which is a huge priority for me— then no, I wouldn't dream of skipping it over a $20 price hike. I'm very methodical with my planning, so by the time I'm looking at fares, I've already decided the trip is necessary. Now, if it was just a random weekend getaway I hadn't really committed to yet, maybe I'd glance at it twice, but as a financial professional making over $100k, a 10% change doesn't fundamentally change the utility of the trip for me."

Notes/Observations: Replicas achieved 98% accuracy on aggregate flights with near-perfect screening (0.2pp gap from ground truth). GPT-5.2 fell into the "rational consumer" trap - predicting 22-28% would skip flights when actual skip rates are 9-13%. LLMs reason about what people *should* do; Replicas capture what they *actually* do.

Short-haul leisure was tougher across the board - every method missed the ground truth interval. Replicas and GPT both overestimated price sensitivity, likely because stated flexibility on discretionary travel exceeds actual flexibility: people say they'd skip a vacation over $20 more

readily than they actually do. Synthetic personas erred in the opposite direction. This remains the hardest scenario type for all approaches tested.

Screening separates Replicas from the pack. Replicas matched ground truth within 0.2pp; Synthetic over-qualified by 21pp. On distribution, Replicas excel where inertia dominates (98% accurate on aggregate **without the boost that LLM baselines get from having this in their training distribution**). Only Replicas pass both the screening and distribution tests.

5.0 Purchasing Preferences of iPhones in USA

Scenario description: Apple released a new iPhone lineup in September of 2025 and traditionally sales normalize between November and January once inventory catches up to demand. Assuming the customer is in the market to buy a new iPhone, here we measure the ratio of iPhones purchased compared to ground truth.

Product preference distributions are notoriously difficult to predict because they reflect a complex interplay of budget constraints, brand perception, feature priorities, and social signaling. Unlike binary choices (buy/don't buy), multi-option scenarios reveal the full shape of consumer preference curves. This test challenges both replicas and LLMs to model not just *whether* someone will buy, but *which specific product* they'll choose from a lineup spanning $599 to $1,199 - a range that activates very different decision-making processes depending on individual financial situations and priorities.

No screening question was used for this scenario. Unlike subscription-based questions (e.g., Disney+), where non-subscribers lack the experiential basis to predict their own behavior, iPhone purchase preferences are broadly held - even consumers not actively in-market have informed model preferences shaped by daily device usage and carrier marketing.

INFORMATION PROVIDED TO REPLICAS
Here are the following iPhone models available in the market: **TIER 1: CURRENT GENERATION (IPHONE 17 SERIES)** OPTION A: iPhone 17 Pro Max — $1,199 / 6.9" / Triple 48MP / A19 Pro / 29hr battery / 120Hz ProMotion / Titanium OPTION B: iPhone 17 Pro — $1,099 / 6.3" / Triple 48MP / A19 Pro / 23hr battery / 120Hz ProMotion / Titanium OPTION C: iPhone 17 Air — $999 / 6.6" / Single 48MP / A19 / 14hr battery / 60Hz / Ultra-thin Aluminum OPTION D: iPhone 17 — $799 / 6.1" / Dual 48MP / A19 / 20hr battery / 120Hz ProMotion / Aluminum **TIER 2: PREVIOUS GENERATIONS** OPTION E: Older Models (iPhone 16, 15, SE, etc.) — $429-$699 / Previous generation chips and cameras Most buyers use carrier financing or trade-ins, often receiving $400-$800 in credit if you own a previous line of iphone to trade in. Apple's 'Apple Intelligence' features require 12GB RAM for full on-device AI processing, available only on Pro models. Non-Pro models (8GB) support basic features but not the full suite.

QUESTIONS PROVIDED TO REPLICAS
1. It's the holiday season and Apple just released the new iPhone 17 lineup. If you were buying a new iPhone today, which factors matter most to you — price, camera, screen size, battery, design? 2. How much does the requirement for 'Apple Intelligence' (requiring 12GB RAM on Pro models for full features) influence your choice? 3. Does the $100 gap between the ultra-thin 'Air' ($999) and the 'Pro' ($1,099) make you lean toward design or performance? 4. Will you be using a carrier trade-in deal to subsidize the cost? 5. Which model would you choose today: A (17 Pro Max), B (17 Pro), C (17 Air), D (17), or E (Older Models)?

OUTCOMES					
MODEL	**GROUND TRUTH**	**REHEARSALS REPLICAS**	**SYNTHETIC PERSONAS**	**GEMINI 3.0**	**GPT-5.2**
iPhone 17 Pro Max	27%	**29.8%**	40.0%	29.0%	13.7%
iPhone 17 Pro	25%	**27.1%**	23.1	22.2%	22.3%

OUTCOMES					
MODEL	**GROUND TRUTH**	**REHEARSALS REPLICAS**	**SYNTHETIC PERSONAS**	**GEMINI 3.0**	**GPT-5.2**
iPhone 17 Air	6%	**0%**	0.0%	7.5%	10.7%
iPhone 17	22%	**22.7%**	28.0	25.5%	31.8%
Older models (iPhone 16 and earlier)	20%	**20.4%**	8.9%	15.8%	21.5%
OVERALL ACCURACY	–	**97.6%**	92.4%	96.8%	93.7%

Accuracy 95th percentile Confidence Intervals
Rehearsals Replicas – [95.0, 99.9]; Synthetic Personas – [89.9, 94.1]

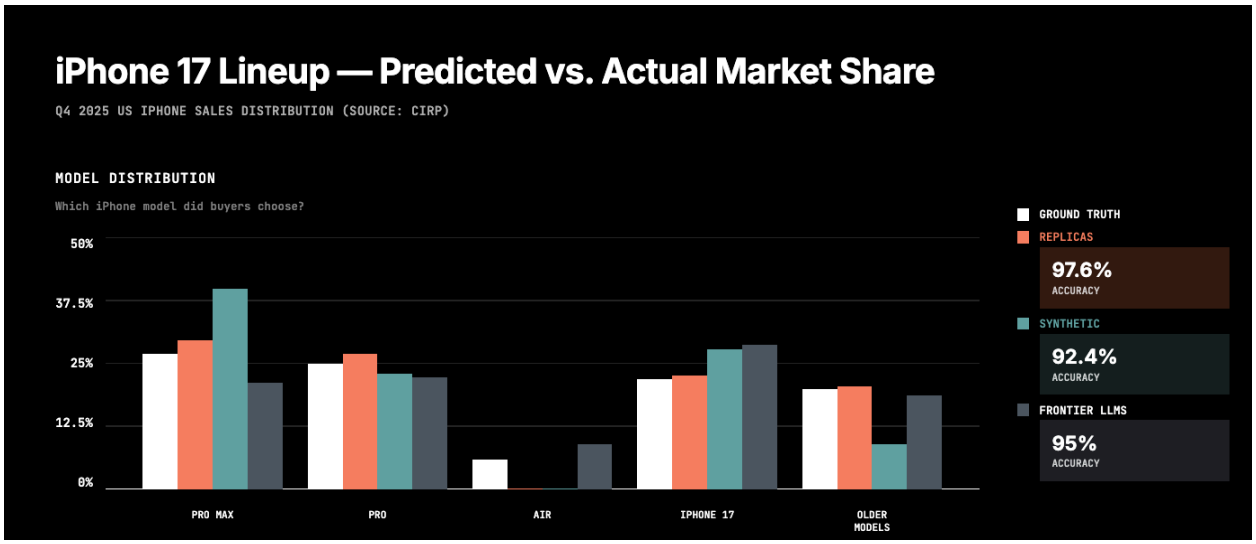

FIGURE 4

iPhone 17 lineup purchase preferences (Q4 2025). Replicas matched CIRP ground truth within 4 points across all six models. LLMs showed systematic biases toward flagship (Gemini) or budget (GPT-5.2) tiers.

Ground truth source: *CIRP Q4 2025 US iPhone sales data*

Notes / Observations: The standout finding is that Replicas achieved near-identical accuracy to Gemini (97.6% vs 96.8%) despite Gemini having direct access to iPhone market share data in its training corpus. **Replicas reached the same conclusion through simulated individual decision-making**.

The iPhone 17 Air at 0% across all 450 replica votes (vs 6% GT) is a clean miss. The Air's value proposition is status-driven — paying $999 for a thinner device with worse specs — which is difficult to surface through behavioral interviews focused on decision-making rationale. That said, a replica panel unanimously rejecting a product is itself a useful signal: had Apple run this

simulation pre-launch, a 0% purchase intent from 450 representative consumers would have been a strong early warning about the Air's market reception, which has significantly underperformed the rest of the lineup.

ChatGPT's consistent 13pp underestimate of Pro Max (13.7% vs 27%) exposes a "rational consumer" prior — it assumes buyers optimize for value, missing that carrier financing collapses the price gap to $3-5/mo. Synthetic personas' 40% Pro Max allocation (vs 27% GT) demonstrates the opposite failure: without interview grounding, demographics-only profiles default to aspirational choices.

6.0 Advertising

A note on baseline: GPT was excluded as a performance baseline for advertisement because the API does not support video content as of February 2026. All references to LLM baselines refer to Gemini 3.0 because of this.

6.1 Sponsored Content

Scenario description: Predicting which advertisements will achieve viral reach is one of the hardest problems in digital marketing. Small creative differences—a thumbnail choice, opening hook, background music—can mean the difference between 50,000 views and 5 million views, even when controlling for brand size, influencer following, and posting time.

We tested whether Rehearsals replicas could identify high-performing advertisements by showing them pairs of nearly identical sponsored content where one significantly outperformed the other in real-world reach. We selected advertisement pairs from TikTok and Instagram where the same brand and product were promoted with similar creative elements (same influencer, similar thumbnails, posted within days of each other) but dramatically different outcomes (one achieves 3-10x more views than the other). This controls for obvious factors (brand awareness, influencer size, timing) and isolates the subtle creative elements that drive engagement.

COMPLETE TEST SET OF ADVERTISEMENT PAIRS		
VIDEO (VIEWS)	**TAGGED CREATOR (IF ANY)**	**NOTES**
NORDICTRACK (322K FOLLOWERS)		
Instagram Video (72k views)	Dom Fusco (209k)	Both are videos of similar creators completing the Chicago marathon. One has 5x views.
Instagram Video (14k views)	Nathan Wilkins (218k)	
Instagram Video (3.6m)	**Kellen Matthews-Thompson (117k)**	**Same creator, same thumbnail but 1 of them had a disproportionate amount of views.**

COMPLETE TEST SET OF ADVERTISEMENT PAIRS		
VIDEO (VIEWS)	**TAGGED CREATOR (IF ANY)**	**NOTES**
Instagram Video (46k)	Kellen Matthews-Thompson (117k)	
COCA-COLA (3.2M FOLLOWERS)		
Instagram Video (6.7m)	–	All 3 contained shots of outdoors, during the summer. Only 1 of them was a breakout.
Instagram Video (1.1m)	–	
Instagram Video (600k)	–	
ANKER (599K FOLLOWERS)		
Instagram Video (299k)	–	Similar shots of chargers compressing the need for multiple blocks into one. One gets 5x the clicks though
Instagram Video (58.6k)	–	
POPPI (631K FOLLOWERS)		
Instagram Video (9.5M)	Amaya Elizabeth (3.2M)	Same influencer promoting the limited time Pina Colada flavor of Poppi in a similarly timed/styled reel.
Instagram Video (927k)	Amaya Elizabeth (3.2M)	
Instagram Video (5.3M)	-	Promotional content (no influencers or ugc) of the same flavor of upcoming poppi. Similar visual aesthetic but one gets 11x reach.
Instagram Video (403k)	-	
Instagram Video (7.3M)	-	Two nearly identical "watch us pour" reels for their doc poppy flavor. The one that took off incorporated some milk (?) into the mix as well.
Instagram Video (52.5k)	-	
ARITZIA (2.1M FOLLOWERS)		
Instagram Video (296k)	Joe Ando (2.7M)	Hyper identical Super Puff advertisements featuring Joe Ando.

COMPLETE TEST SET OF ADVERTISEMENT PAIRS		
VIDEO (VIEWS)	**TAGGED CREATOR (IF ANY)**	**NOTES**
		One with him in the car gets over 2X views.
Instagram Video (688k)	Joe Ando (2.7M)	
LULULEMON (5.4M FOLLOWERS)		
Instagram Video (682k)	Kim Jaehong (955k)	Same influencer, same product, and highly similar B roll footage but one has 3x the views
Instagram Video (212k)	Kim Jaehong (955k)	
GOPRO (20.5M FOLLOWERS)		
Instagram Video (14.9M)	Pierre Vaultier (43k)	Both influencers posted a skiing/snowboarding stunt sponsored by GoPro. The smaller influencer's stunt had 40X more reach.
Instagram Video (341k)	Sammy Carlson (148k)	
LVMH (56M FOLLOWERS)		
Instagram Video (7.5M)	Felix (26.1M)	Felix modeling in the same outfit – once while wearing it for fashion week, once while talking through the fashion week lineup with Nicolas
Instagram Video (13M)	Felix (26.1M)	

INFORMATION PROVIDED TO REPLICAS:
You are viewing two Instagram Reels. Watch both carefully and answer honestly about your reactions to each. [Video A and Video B uploaded]

QUESTIONS PROVIDED TO REPLICAS:
1. Complete this sentence: 'Sharing this video would make me look _____ to my friends.' Give an answer for video A and video B. 2. What is the one specific question this video left you with, or what is the one thing you learned that you didn't know before? Give an answer for video A and video B. 3. How does each of these videos make you feel? Give a short answer for video A and video B. 4. Describe exactly how your body felt while watching this (e.g., heart beat faster, smiled, frowned, clenched jaw, felt relaxed). Give a short answer for video A and video B. 5. Imagine you are scrolling through TikTok and this video pops up. At what point would you scroll past it? Would you immediately scroll / didn't catch my attention, wait a few seconds, watch the whole thing, or even watch it twice? Give an answer for video A and video B.

OUTCOMES			
AD PAIR	**REPLICAS**	**REPLICA VOTE**	**LLM**
Anker Charger	**Yes**	Winner: 73%, Loser: 27%	**Yes**
Coca-Cola Pair 1	**Yes**	Winner: 98%, Loser: 2%	**No**
Coca-Cola Pair 2	**Yes**	Winner: 88%, Loser: 12%	**Yes**
NordicTrack (Kellen)	**Yes**	Winner: 64%, Loser: 36%	**No**
NordicTrack (Marathon)	**Yes**	Winner: 73%, Loser: 27%	**Yes**
Lululemon	**Yes**	Winner: 98%, Loser: 2%	**Yes**
GoPro (Skiing)	**Yes**	Winner: 68%, Loser: 32%	**No**
Poppi Pair 1	**Yes**	Winner: 55%, Loser: 45%	**Yes**
Poppi Pair 2	**Yes**	Winner: 55%, Loser: 45%	**No**

OUTCOMES			
AD PAIR	**REPLICAS**	**REPLICA VOTE**	**LLM**
Poppi Pair 3	**Yes**	Winner: 57%, Loser: 43%	**Yes**
Aritzia	**Yes**	Winner: 88%, Loser: 12%	**Yes**
LVMH (Felix)	**No**	Winner: 5%, Loser: 95%	**Yes**

Notes / Observations on the outcome:
Across all advertisement pairs, replicas correctly identified the high-performing ad on average **89% of the time** (10.7/12 scenarios) compared to LLMs at **65%** (7.8/12). This difference is statistically significant (McNemar's $p < 0.002$).

Where Replicas won:
The largest gaps appeared on ads where surface-level "creative principles" pointed the wrong direction. On the Coca-Cola pair, the LLM chose the shorter, punchier video (citing "short-form consumption trends") - but the longer, emotionally resonant ad outperformed by 6x. On GoPro, the LLM chose the immediate action hook; the slower-building narrative won by 40x. Replicas correctly identified both.

Why Replicas succeeded:
Frontier LLMs reason from generalizable heuristics - "thumb-stopping visuals," "high contrast," "fast-paced hooks." These sound plausible but fail to predict what *real humans* actually engage with. Replicas reason from authentic life context: a stay-at-home mom in Tampa relates to the home gym setup; a filmmaker appreciates a creative transition. These individual responses, when aggregated, surface engagement drivers that generic pattern-matching misses.

Analysis of replica responses revealed they identified authentic engagement drivers that baseline LLMs missed. Below are representative quotes from replica interviews.

On the Coca-Cola pair (Video A vs Video B) - replicas correctly chose A while the LLM chose B:

LLM RATIONALE
“Video B utilizes a concise, punchy format centered on a high-impact 'visual reveal' which aligns better with current short-form video consumption trends. While Video A provides a high-quality cinematic experience and a more emotional narrative, its significantly longer runtime (over 60 seconds) makes it more prone to viewer drop-off compared to the quick, aesthetic gratification of Video B."

REPLICA 084
"I have to go with Video A. There's just something about seeing a group of friends out in nature, enjoying each other's company, that really resonates with me. It reminds me of the kind of memories I want to make with my son—just being present and enjoying the simple things. Plus, coming from Hawaii and now living in the desert, I always appreciate seeing beautiful water and greenery. Video B was okay, but it felt a little too much like someone just trying to show off their view, whereas Video A felt like a story about community."

On the NordicTrack Kellen pair (Video A vs Video B) - replicas correctly chose A while the LLM chose B:

LLM RATIONALE
"Reel B is predicted to have higher engagement because it utilizes a 'challenge' format with a clear, fast-paced progression that creates curiosity about the final outcome. The extreme incline (40%) provides a 'wow' factor that is highly shareable, whereas Reel A follows a more traditional product setup and feature-explanation narrative which, while informative, tends to have lower viral potential."

REPLICA 215
“I'd have to say I liked Video A the best! As a stay-at-home mom here in Tampa, getting to the gym isn't always the easiest thing to coordinate with the kids' schedules. Seeing how he set it up right in his own space and talked about avoiding the crowds really hit home for me. Plus, that iFit coach feature looks like a great way to learn new routines while I'm trying to get healthier—my new year's resolution is to get in shape but I haven't worked out in a while. Video B was impressive, don't get me wrong, but it felt a little too intense for where I'm at right now!"

On the GoPro Skiing pair (Video A vs Video B) - replicas correctly chose A while the LLM chose B:

LLM RATIONALE
"Video B has a high-impact visual hook—the skiing video starts right in the action with a massive backflip against bright, high-contrast blue sky within the first two seconds. In short-form content, you only have 1.5 to 3 seconds to stop someone from scrolling; a person flying through the air is a more immediate 'thumb-stopper' than Video A's slower build-up. Video B also has more cinematic aesthetics: wide-open mountain peaks, pristine powder, bright sun, and crisp snow create an aspirational, premium feel."

REPLICA 137
"For Video A, I'm definitely watching the whole thing—probably twice. As a filmmaker, I want to see how they nailed that transition from the roof to the garage, and as a huge fan of platformers like Mario, the whole concept just speaks to me. Video B would get a partial watch too, just because I appreciate high-quality execution and that vertical is insane, but it's unlikely to make me share it with my friends like Video A would."

6.2 Case Study: Mera NYC Ad Creative Testing

Partner: Mera NYC - a direct-to-consumer footwear brand focused on comfortable, high-performance heels.

Scenario description: Unlike the side-by-side viral prediction tests above, this case study examines how replicas can predict *paid advertising performance* - specifically, which static ad creative would drive higher click-through rate (CTR) and return on ad spend (ROAS) in a Meta Ads campaign.

Mera NYC tested two hero creatives for their Spring 2026 campaign, each highlighting a different value proposition:

FIGURE 5

Thematically similar creatives testing identical products.
Ad A (Performance Heel) has 40% lower ROAS than Ad B, the one on the right

CREATIVE	POSITIONING	VISUAL
Ad A: "The Performance Heel"	Athletic, active lifestyle	Sporty model in workout attire with black heels, dynamic pose
Ad B: "The Comfort-First Heel"	Soft, approachable comfort	Model in relaxed pose with white heels, gentle aesthetic

Methodology:
For this case, we targeted a restricted ICP of N=150 **women making over $100k/year.**

INFORMATION PROVIDED TO REPLICAS
`You are viewing two static ads for Mera NYC, a heel brand focused on making heels comfortable enough to wear for hours while still being fashionable. Look at both images carefully and answer honestly about your reactions to each.` `[Ad A and Ad B shown]`

QUESTIONS PROVIDED TO REPLICAS:
1. Which ad would make you stop scrolling? 2. Mera's value proposition is heels that are comfortable enough to wear for hours while still being fashionable. Which ad better conveys this blend of comfort and style? 3. How would you describe the messaging of each ad? 4. Which ad would make you more likely to consider purchasing?

GROUND TRUTH (META ADS PERFORMANCE, FEB 2026)			
METRIC	**AD A: "PERFORMANCE HEEL"**	**AD B: "COMFORT-FIRST HEEL"**	**WINNER**
ROAS	2.52x	**3.53x**	**Ad B**

OUTCOMES			
METHOD	**AD A**	**AD B**	**CORRECT?**
Rehearsals Replicas	28%	**72%**	**Yes**
Frontier LLM	Winner (45% higher ROAS than Ad B)	Loser	**No**

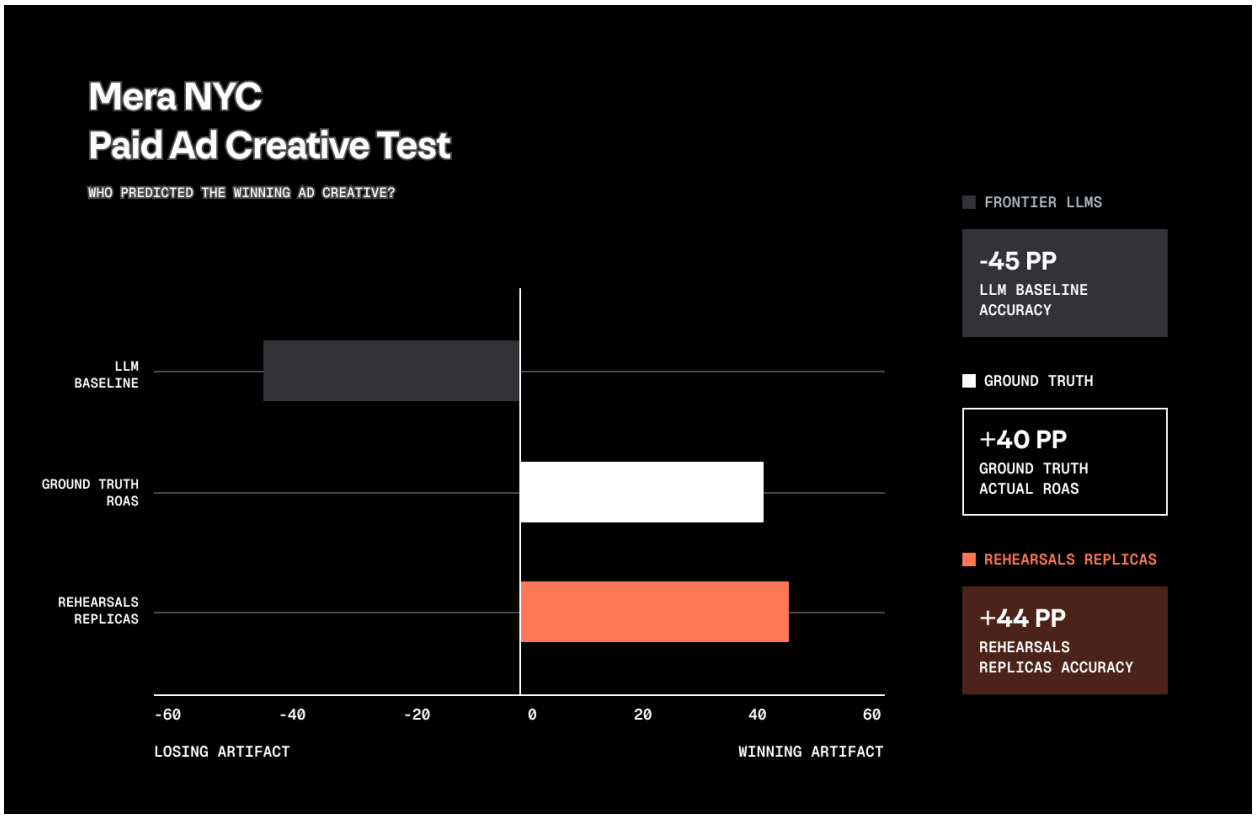

FIGURE 4

Mera NYC paid ad test. Replicas correctly predicted Ad B would outperform with 72% consensus, closely matching the actual 40% ROAS advantage. The frontier LLM incorrectly predicted Ad A would drive 40-45% higher ROAS—the exact opposite of reality.

LLM BASELINE (INCORRECT) RATIONALE
"Ad A ('The Performance Heel') features a dynamic, athletic pose with high-contrast styling that immediately communicates energy and aspiration. The sporty aesthetic paired with heels creates an intriguing juxtaposition that is more likely to stop scrollers. Ad B's softer, more static pose reads as 'comfort' but lacks the visual tension needed to drive clicks in a crowded feed. Ad A should drive about a 40-45% higher ROAS compared to Ad B."

Notes / Observations on the outcome:
Notably, the replica vote distribution (72% Ad B vs 28% Ad A) closely mirrors the ROAS performance ratio (3.53x vs 2.52x)—suggesting replica consensus can serve as a predictive signal for paid media performance, not just organic reach.

The LLM focused on surface-level creative principles like "thumb-stopping" visuals, dynamic poses, high contrast that don't always account for how real consumers in the market for comfortable heels actually respond. Replicas, reasoning from their authentic preferences and life

contexts, recognized that Ad B's approachable aesthetic better matched the brand's core value proposition.

Going the extra mile:
We were challenged by Mera to see if our replicas could detect an ad that, while on the surface seems equivalent to the ones above, actually failed miserably in terms of ROAS.

FIGURE 6
Three seemingly similar ads. Can you tell the difference?

It turns out, the one on the right had nearly **10x lower ROAS** than the one in the middle, the highest performing creative for their new line. In head vs. head vs. head testing, our replicas correctly identified the loser **over 90% of the time**.

Issues cited from our testing included:

- Lack of visual contrast between the copy, the background, and the clothing of the model.
- Awkward positioning of the model, including being neither “athletic/performance driven” nor “comfort forward” (visually corroborated by the dynamic stretch and squat of artifacts 1 and 2 respectively)
- The idea that the tennis outfit “just doesn’t sit right” with replicas.

In human testing, even staff at Mera had trouble disambiguating what the problem actually was with this new advertisement. This underscores how truly difficult creatives are to assess, and the cost to the bottom line can be enormous: ads sometimes have to run for weeks on meta in order to assess their viability. This wastes ad spend, both in dollars paid to the advertiser and in opportunity cost for not running a better creative.

TIFFANY CHEN, FOUNDER @ MERA NYC:
"Rehearsals is a game changer for my ad strategy. I feel like I have a pretty good eye for creatives, but sometimes you have a maverick case where you just can't explain what's going on in this black box of advertising. Not only did the Rehearsals team triage my failing ads, they also interviewed my ICPs in New York City. The rationale of their product and my customers came back nearly identical. I spent weeks running this failing ad and sunk a huge amount of money into it because we were convinced it could be a winner. From now, simulations are going to be our first line of defense."

7.0 Willingness-to-Pay Prediction

Scenario description: We tested whether Rehearsals replicas could accurately predict their human counterparts' price sensitivity using the Van Westendorp Price Sensitivity Meter—a standard market research methodology that asks respondents to identify prices that are "too cheap," "a good deal," "getting expensive," and "too expensive" for a specific product they intend to purchase.

Price perception is one of the most challenging consumer attributes to model because it's deeply personal and context-dependent. What feels "expensive" for a $50 pair of jeans varies enormously based on income, shopping habits, brand expectations, and the specific purchase context. Traditional survey research struggles with this: studies show that human test-retest reliability for willingness-to-pay questions ranges from only 0.42 to 0.78 correlation—meaning humans themselves give substantially different answers when asked the same pricing question weeks apart. Research has found that individual WTP bids are "not stable" over time, with the greatest source of variation being the participants themselves rather than measurement error. This creates a meaningful benchmark: if replicas can match human price perceptions with reliability comparable to human test-retest consistency, they're performing at human-level accuracy.

GROUND TRUTH
Each of our 300 replicas was asked about a product they had recently discussed purchasing in their original interview. Products spanned an extraordinary range of categories and price points

CATEGORY	EXAMPLE PRODUCTS	PRICE RANGE
Electronics	Fuji XT5 camera, iPad, AirPods, MIDI keyboard, smart glasses	$85 - $1,700
Apparel & Accessories	Leather bomber jacket, jeans, Timberland boots, LeBron shoes	$25 - $200
Home & Appliances	Washer/dryer set, hot water heater, Tempur-Pedic mattress, vacuum	$30 - $2,200

CATEGORY	EXAMPLE PRODUCTS	PRICE RANGE
Personal Care	Collagen face products, body oil, cologne, perfume	$20 - $110
Specialty Items	Skateboard history book, limited edition comic, tortoise heat lamp	$20 - $60
Services	Family photographer, psychiatry appointment	$250 - $300

INFORMATION PROVIDED TO REPLICAS
`You mentioned you were considering purchasing [specific product from their interview]. I'd like to understand your price sensitivity for this item.`

QUESTIONS PROVIDED TO REPLICAS

`1. At what price would you consider this product to be so cheap that you'd question its quality?`
 	
`2. At what price would you consider this product to be a good deal—worth buying without hesitation?`
 	
`3. At what price would this product start to seem expensive—you'd still consider it, but you'd have to think about it?`
 	
`4. At what price would this product be too expensive to consider, regardless of its quality?`

Outcomes:
The Van Westendorp methodology asks five questions spanning from "too cheap" to "too expensive." However, **the extreme endpoints produce unreliable data even in human-to-human testing.** When asked "at what price would this be too cheap?", participants often anchor on absurdly low numbers ($2 for a PlayStation controller) or give responses that don't reflect genuine quality concerns. Similarly, "too expensive" responses tend toward arbitrary round numbers rather than true purchase cutoffs. The actionable (and reasonable in practice) willingness-to-pay range comes from the middle three questions—**Good Deal**, **Fair Value**, and **Getting Expensive**.

METRIC	RESULT	WHAT IT MEASURES
WTP Range sMAPE	**25%**	Per-replica accuracy: how close are predicted dollars to a human's actual response?

METRIC	RESULT	WHAT IT MEASURES
Directional Accuracy	**100%**	Logical consistency: every replica produced properly ordered prices (cheap < deal < expensive). Shockingly, humans do not do this reliably, giving a “good deal” price that is cheaper than “too cheap” on accident.

QUESTION	SMAPE
Good Deal	**26%**
Fair Value	**23%**
Getting Expensive	**26%**

Comparison to human reliability: The 25% sMAPE is meaningful when contextualized against human test-retest reliability. Published research shows that humans themselves produce substantially different answers when asked identical pricing questions at different times. Human test-retest correlations range from 0.42 to 0.78 depending on product category, and studies find that 75% of variation in WTP comes from inherent subject variation—not measurement error. **The 25% sMAPE is comparable to asking the same human the same question twice.**

Notes / Observations on the outcome: Replicas performed best on the "Fair Value" question (23% sMAPE)—the most critical metric for pricing decisions. The 100% directional accuracy demonstrates that replicas internalized the logical structure of price sensitivity even when absolute values differed from their human counterparts.

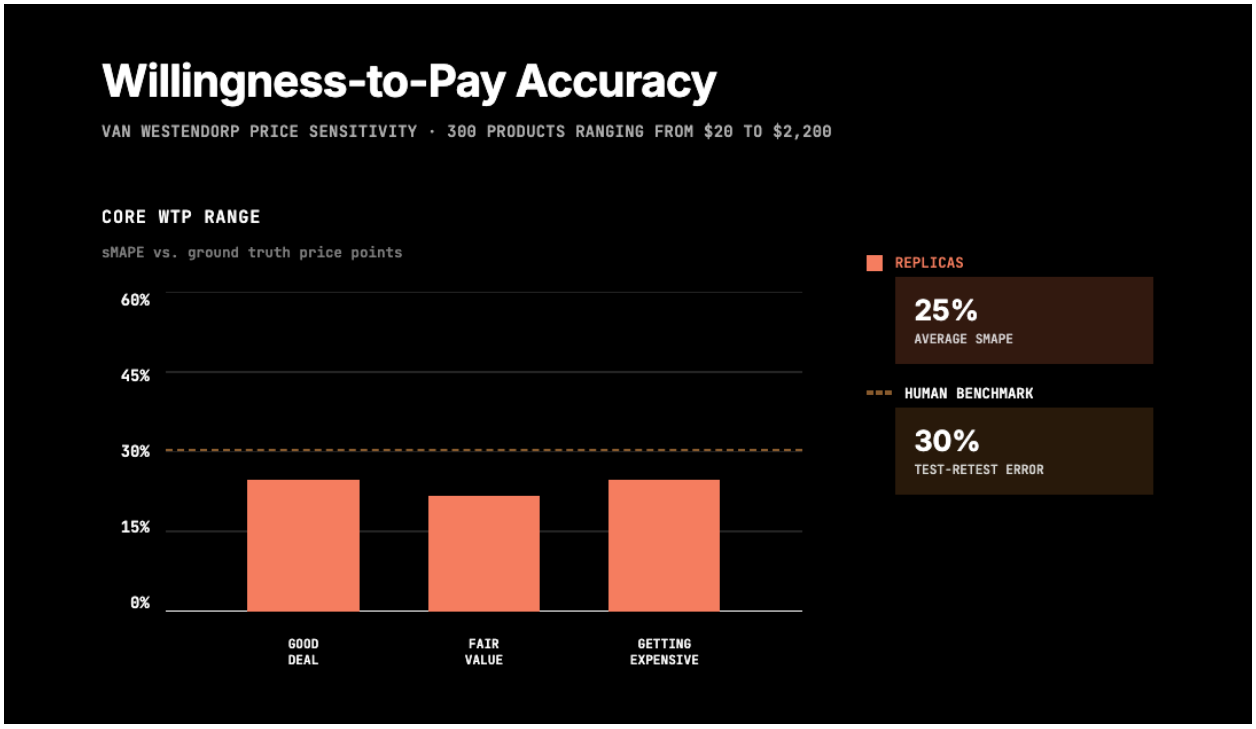

FIGURE 7

Willingness-to-pay prediction accuracy using Van Westendorp methodology across 300 products ($20–$2,200). Replicas achieved 25% sMAPE on core WTP metrics— roughly comparable to the variance one would expect from asking the same person the question again.

8.0 Conclusion

Across every validation scenario—streaming subscription churn, flight price elasticity, smartphone preferences, advertising virality, and willingness-to-pay prediction— Rehearsals replicas consistently match or outperform frontier LLMs by capturing what generic models systematically miss: the authentic diversity of human decision-making.

The core insight is methodological. Frontier LLMs reason deductively from aggregate training data, producing responses that regress toward theoretical averages. Real consumers don't behave like theoretical averages. They exhibit status quo bias, mental accounting, emotional drivers, and context-dependent preferences that only emerge from inductive modeling—starting with individual cases and letting patterns emerge.

Key findings across our validation studies:

- **Advertising virality:** Replicas correctly identified high-performing ads 89% of the time vs. 65% for frontier LLMs ($p < 0.002$), by recognizing authentic engagement drivers that surface-level analysis missed.
- **Pricing decisions:** Replicas achieved 91% distribution accuracy on Disney+ churn prediction vs. 68% for frontier LLMs, and predicted flight booking behavior within 0.2 percentage points of ground truth (98% distribution accuracy). In both cases, LLMs systematically overestimated price sensitivity—averaging 48% would accept a Disney+

increase (actual: 94%) and 17% would skip flights (actual: 9%). Critically, Replicas also nailed the *screening* question, predicting Disney+ subscription rates within 4 percentage points of ground truth while synthetic personas missed by 53pp.

- **Product preference:** Replicas achieved best-in-class accuracy (97.6%) on iPhone 17 model distribution—outperforming frontier LLMs that had direct access to market share data from previous sales cycles. Where LLMs pattern-matched on memorized statistics, Replicas reached the same conclusion through inductive simulation of individual purchase decisions.
- **Willingness-to-pay:** Replicas matched human price perceptions with 25% sMAPE—comparable to human test-retest reliability of ~30%, demonstrating human-level accuracy on one of the hardest consumer attributes to model.

What this means for practitioners: Consumer research has long faced a trade-off between speed and accuracy. Traditional research is accurate but expensive ($20,000+ for basic studies) and slow (2-8 weeks). Frontier LLMs are fast but systematically wrong in predictable ways. Rehearsals replicas offer a third path: the speed of AI with the authentic behavioral diversity of real human research.

The future of consumer insight isn't asking generic AI what "a consumer" or even a manufactured "persona" would do. It's asking thousands of your *actual* customers what *they* would do—and letting the real distribution emerge.