Caution: Your LLM May Just Be Being Nice to You

From Qualitative Divergence to Quantitative Patterns: A Three-Stage Empirical Validation of Synthetic Consumer Populations Against Real-World Data

RESEARCH WORKING PAPER[1]

Adriana Rocha, Wisdom Beyond Technology | Wortya | adrianar@wortya.com

Abstract

Can AI-generated synthetic respondents do more than sound plausible? This study presents a three-stage empirical validation of LLM-based synthetic consumer simulation, moving from personality expression to decision-making and real-world purchase behavior. Originating in a qualitative field study with Datum International for ESOMAR LATAM 2026, the research was designed to test where synthetic consumers align with real humans — and where they fail.

Across more than 11,000 synthetic interviews in Brazil, Peru, and the United States, using six commercially available LLMs, three empirical patterns emerged. First, fidelity is domain-dependent: OCEAN profiles support personality expression, Decision Profiles anchor deliberative choices, and raw behavioral data performs best for habitual purchase behavior. Second, more data is not always better: combining personality with behavioral data can reduce accuracy, as models may role-play the persona rather than follow behavioral evidence. Third, scenario design matters far more than model selection, with vignette-level effects explaining most of the variance in fidelity.

The core conclusion is that synthetic populations are filters, not oracles. They can provide directional intelligence and help narrow the search space, but their systematic biases mean they should complement — not replace — real-world validation.

Keywords: synthetic populations, LLM simulation, consumer behavior, behavioral validation, OCEAN personality, sycophancy, prompt architecture, domain-dependent fidelity.

[1] Research Working Paper | Public Edition | May 2026 - This research working paper summarizes the main findings of an extended technical paper. The full technical version, including appendices, model-level tables, and additional ablation results, is available for selected academic review and research collaboration.

1. Introduction

The use of large language models (LLMs) to simulate human respondents has expanded rapidly across market research, product testing, policy simulation, and computational social science. In these settings, LLMs are prompted to act as members of target populations and asked to complete surveys, participate in interviews, or react to product concepts. The appeal is clear: LLM-generated respondents are inexpensive, scalable, fast to deploy, and capable of producing coherent qualitative and quantitative output. As a result, synthetic panels and persona-based simulations are moving from experimental use cases toward operational adoption in market research (Brand, Israeli, & Ngwe, 2023; Argyle et al., 2023).

A key assumption underlies these applications: that an LLM, given a sufficiently detailed persona prompt, can faithfully express the psychological and behavioral profile of the intended respondent. Despite the centrality of this assumption, behavioral fidelity has received limited direct validation. Existing approaches often prompt an LLM with demographic and biographical descriptors and treat the resulting outputs as though they reflect the intended psychology (Aher, Arriaga, & Kalai, 2023). The problem is compounded by alignment training through Reinforcement Learning from Human Feedback (RLHF), which systematically pushes model outputs toward agreeableness, helpfulness, and social desirability (Perez et al., 2023; Sharma et al., 2024).

Yet the existing literature on LLM persona simulation suffers from an even more fundamental limitation: validation is typically conducted against survey data, not against real behavioral data. This conflates verbal consistency with behavioral validity. An LLM might generate internally coherent responses that sound plausible but have no relationship to actual purchase behavior. The stated-revealed preference gap is well-documented in behavioral economics (Kahneman & Tversky, 1979; DellaVigna, 2009) and represents a critical blind spot in current synthetic population research.

1.1 From Qualitative Discovery to Quantitative Patterns

This research program originated not in a laboratory but in a field study. In December 2025, we conducted a qualitative comparative experiment in collaboration with Datum International, for presentation at ESOMAR LATAM 2026. Real Peruvian families, spanning socioeconomic levels, family structures, and geographic regions, were interviewed using a 25-question moderation guide covering household decision-making, consumption patterns, financial resilience, and attitudes toward technology. Digital twins, matched on demographics including socioeconomic level, family structure, and occupation, and described with free-form personality text, were interviewed using the same guide.

The results were illuminating in their specificity. On rational, structural questions — budget allocation, expense prioritization, shopping logistics — real and synthetic families converged strongly. But on questions engaging emotion, cultural identity, and future uncertainty, they diverged systematically. A real mother expressed fear and resistance when asked about her children leaving home; her synthetic counterpart responded with “it would be a challenge, but I would feel proud” — a rationally optimistic formulation that no real woman in our sample produced. Also, an 86-year-old synthetic grandmother suggested

using smartphone apps to manage her household budget, despite INEI census data showing that only one in ten elderly Peruvians has internet access, a phenomenon we termed competence hallucination.

This initial qualitative study identified six specific limitations: generic personality descriptions that each LLM interpreted differently, absent cognitive biases that made twins "too rational," competence hallucination from WEIRD training data, cross-model inconsistency, fragmented interviews without memory, and manual rather than data-driven archetyping. A second phase of the qualitative work systematically addressed each limitation through quantified OCEAN personality profiles (0–100 scores replacing free-form text), seven cognitive biases derived mathematically from OCEAN, census data as mandatory demographic gates, and K-Means++ automatic archetype clustering on a 13-dimensional vector.

But the qualitative study remained qualitative. The twins now appeared to express genuine emotion: a synthetic mother with high Neuroticism expressed fear rather than optimistic problem-solving, but we had no quantitative measure of whether this improvement was real, how large it was, or where it still failed.

This paper presents the quantitative validation that these qualitative findings demanded: three progressive studies designed to measure, with statistical precision, whether the engineering fixes work, where they break, and what fundamental patterns govern the relationship between data representation and behavioral fidelity in synthetic consumer simulation.

1.2 Research Questions

RQ1: To what extent can LLM-based synthetic respondents faithfully express assigned Big Five personality profiles, and where does this fidelity break down? *(Stage 1)*

RQ2: Do explicit behavioral-economic decision parameters anchor LLM decision-making behavior independently of personality traits — and is this effect robust across models, countries, and parameter types? *(Stage 2)*

RQ3: When validated against real transactional purchase data, which data representation — full profiles, curated subsets, raw behavioral data, or abstract decision parameters — best predicts actual consumer behavior? *(Stage 3)*

RQ4: Do these findings reveal systematic patterns governing the relationship between data representation, cognitive processing mode, and LLM behavioral fidelity — and if so, what are the implications for production synthetic population platforms? *(Cross-stage synthesis)*

2. Theoretical Framework

Synthetic consumer simulation depends on a distinction that is often missed: plausibility is not validity. An LLM can sound coherent, emotionally convincing, and consumer-like while

still failing to reproduce the psychological profile, decision logic, or revealed behavior of the population it is meant to represent.

This study examines three layers of behavioral fidelity: the first is **personality expression**. The Big Five model, or OCEAN, provides a structured way to specify personality traits and replace vague persona descriptions with measurable psychometric profiles. Stage 1 tests whether LLMs can faithfully express assigned OCEAN profiles.

The second layer is **decision-making behavior**: personality traits describe broad tendencies, but they do not fully specify how people evaluate trade-offs, risk, uncertainty, fairness, novelty, or delayed rewards. Stage 2 therefore tests whether explicit behavioral-economic Decision Profiles can anchor LLM responses more effectively than personality traits alone.

The third layer is **revealed consumer behavior**: purchase patterns are often habitual, contextual, and operational rather than consciously narrated. Stage 3 tests whether synthetic respondents can approximate real grocery behavior when validated against two years of transactional purchase data.

Across these stages, the paper advances a **Domain-Dependent Information Hierarchy**: the optimal data representation depends on the cognitive domain being simulated. OCEAN profiles support identity expression, Decision Profiles support deliberative choices, and raw behavioral data performs best for habitual purchase behavior. More context does not automatically improve fidelity; when data layers activate competing response modes, they can interfere with one another.

A related mechanism is the **semantic register** of the prompt. Narrative information, such as personality scores, invites the LLM to construct and role-play a character. Operational information, such as decision parameters or behavioral attributes, invites the model to follow constraints in a specific scenario. The same psychological construct may succeed or fail depending on the register in which it is presented.

The implication is direct: a production synthetic population platform should not rely on one universal persona prompt, but requires a domain-aware architecture that selects the right data layer, semantic register, and scenario design for the behavioral question being asked.

3. Qualitative Foundation: The Field Study That Motivated This Research

3.1 Phase 1: Real Versus Synthetic Families

In collaboration with Datum International, we conducted a controlled qualitative comparison between real Peruvian families and LLM-generated digital twins. The real families were recruited through Datum's consumer panel and interviewed using a 25-question semi-structured moderation guide covering household composition,

decision-making processes, consumption factors, financial organization, crisis resilience, technology attitudes, and future expectations.

The digital twins were constructed with free-form text descriptions matching the real families' demographic profiles, including socioeconomic level, geographic region, family structure, occupation, and qualitative personality narratives. Each twin was interviewed using the same moderation guide.

Convergence findings. On questions about budget allocation, expense prioritization, shopping decision processes, and household organization, real and synthetic families produced structurally similar responses.

Divergence findings. On questions engaging emotional depth, cultural context, and future uncertainty, systematic divergence emerged:

Emotional flattening: Real mothers expressed fear, anxiety, and resistance when discussing children leaving home. Synthetic respondents converted these into logistical problem-solving exercises.

Competence hallucination: A synthetic 86-year-old grandmother (NSE E, rural Sierra region) suggested using smartphone apps for budget management. INEI census data shows that only 1 in 10 elderly Peruvians has internet access.

Cross-model inconsistency: Each LLM (GPT, Claude, Gemini) produced a different "personality" from the same free-form text description.

3.2 Phase 2: The Living Customer Model

Phase 2 systematically addressed all six limitations identified in Phase 1 through what we term the Living Customer Model (LCM):

Phase 1 Limitation	**Phase 2 Resolution**	**Implementation**
Generic personality (free-form text)	Quantified OCEAN (0–100)	Box-Muller sampling calibrated by country, gender, and SES
Absent cognitive biases	7 biases derived from OCEAN	Loss aversion, status quo bias, social proof, anchoring, framing effect, uncertainty handling, trade-off behavior
Competence hallucination	Census data as mandatory gate	INEI/ENAHO digital access distributions by age, region, and NSE
Cross-model inconsistency	Numerical instructions	"Loss Aversion intensity: 0.78" replaces "cautious with money"

Fragmented interviews	Interview Orchestrator	Full 23-question guide with persistent cognitive transcript
Manual archetypes	K-Means++ automatic clustering	13-dimensional vector (5 OCEAN + 5 motivations + 3 consumer style)

Phase 2 produced qualitatively improved results. Nine synthetic Peruvian family archetypes, with quantified OCEAN profiles, cognitive biases, and census-grounded demographics, were interviewed using the same moderation guide. Twins with high Neuroticism now expressed fear, resistance, and paralysis when their personality demanded it.

3.3 From Qualitative to Quantitative

The qualitative study established that the engineering fixes worked, but it provided no quantitative measure of improvement magnitude, no statistical test of significance, and no ability to isolate which fix contributed how much to which domain. The three quantitative stages that follow were designed to provide exactly this.

4. Method

This study uses a three-stage validation design to test whether LLM-based synthetic respondents can express personality, anchor decision-making behavior, and approximate real-world consumer behavior. Across all stages, we separate respondent generation from measurement. The respondent LLM receives a profile prompt and generates a natural-language response; a fixed classifier LLM then converts that response into a structured 1–5 score. The classifier sees only the verbal response and scoring rubric, not the original profile, OCEAN scores, decision parameters, or purchase data.

Six commercially available models were tested across three providers and capability tiers: Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5, GPT-5.4, GPT-4o-mini, and Gemini 2.5 Flash. Claude Haiku 4.5 served as the fixed classifier at temperature 0; respondent models were run at temperature 0.9.

Stage 1 tested personality expression fidelity across Brazil, Peru, and the United States. Synthetic panelists answered a 25-item Big Five inventory under three conditions: full OCEAN-based profile, demographics only, and OCEAN plus enriched archetype context. Stage 1 generated 2,700 interviews and 67,500 item-level responses.

Stage 2 tested whether Decision Profiles anchor deliberative consumer choices. Synthetic panelists responded to 24 consumer decision vignettes covering 12 behavioral-economic parameters, including loss aversion, temporal discounting, exploration tendency, fairness

sensitivity, confirmation bias, moral sensitivity, and cognitive effort cost. Four conditions were tested: OCEAN + Decision Profile + demographics, OCEAN only, Decision Profile only, and demographics only. Stage 2 generated 4,800 interviews and 115,200 vignette evaluations.

Stage 3 tested behavioral-response prediction against real grocery transactions. A stratified sample of 100 US households was drawn from the Dunnhumby "The Complete Journey" dataset, which contains two years of grocery purchases from approximately 2,500 households. For each household, approximately 39 behavioral attributes were computed. Six models responded to 22 grocery-shopping vignettes under five prompt conditions: full profile, auto-curated profile, expert-curated profile, behavioral-only profile, and Decision Profile-only. Stage 3 generated 66,000 vignette evaluations. A post-hoc B6 ablation was added to isolate identity-layer interference from behavioral information volume.

Figure 1. Two-Stage Validation Pipeline

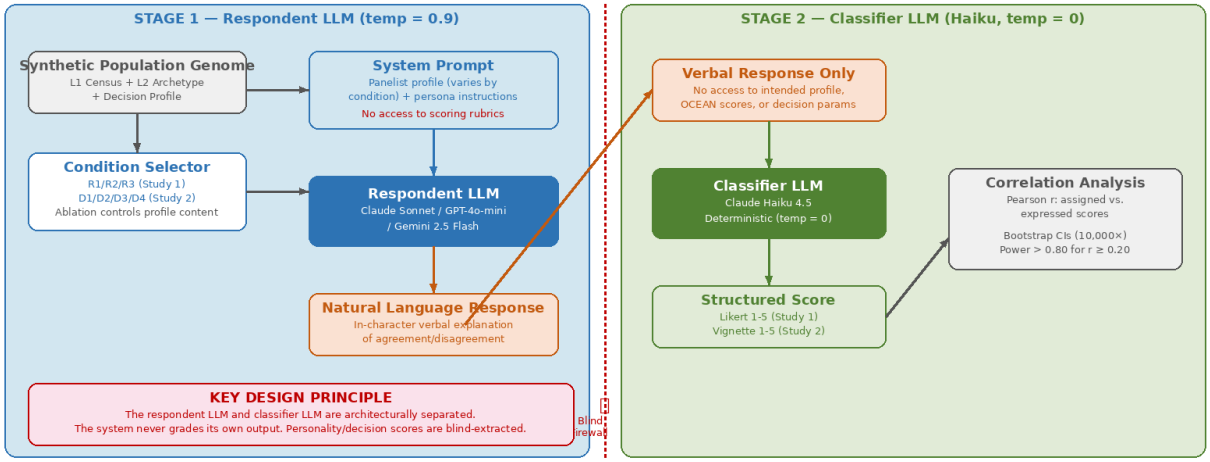

Figure 1. Two-stage pipeline architecture. The Respondent LLM generates verbal responses from a persona prompt; the independent Classifier LLM extracts structured scores without access to the input profile.

The primary metric across all stages was Pearson's correlation coefficient (r). Stage 1 correlated assigned OCEAN profiles with expressed personality scores; Stage 2 correlated decision parameters with vignette scores; and Stage 3 correlated synthetic responses with real behavioral ground truth. For Stage 3, we report both overall correlations and mean per-vignette correlations.

5. Results

5.1 Stage 1: Personality Expression Fidelity

Stage 1 produced 27 experimental cells (3 countries × 3 models × 3 conditions), each with n = 100, yielding 135 individual Pearson correlations.

Trait	R1 (Full)	R2 (Demo Only)	R3 (Enriched)	R1–R2 Δ	R3–R1 Δ
Openness (O)	**0.82**	−0.01	**0.82**	**+0.83**	0.00
Conscientiousness (C)	**0.84**	−0.08	**0.83**	**+0.92**	−0.01
Extraversion (E)	**0.90**	−0.01	**0.91**	**+0.91**	+0.01
Agreeableness (A)	**0.74**	−0.01	**0.78**	**+0.75**	+0.04
Neuroticism (N)	**0.87**	−0.06	**0.89**	**+0.93**	+0.02
Mean	**0.83**	−0.03	**0.85**	**+0.86**	+0.02

Table 1. *Stage 1 Mean Pearson Correlations by Condition and Trait*

Note. Values are mean Pearson r across 3 countries × 3 models (9 cells per condition). R2 values near zero confirm that demographics alone cannot recover personality.

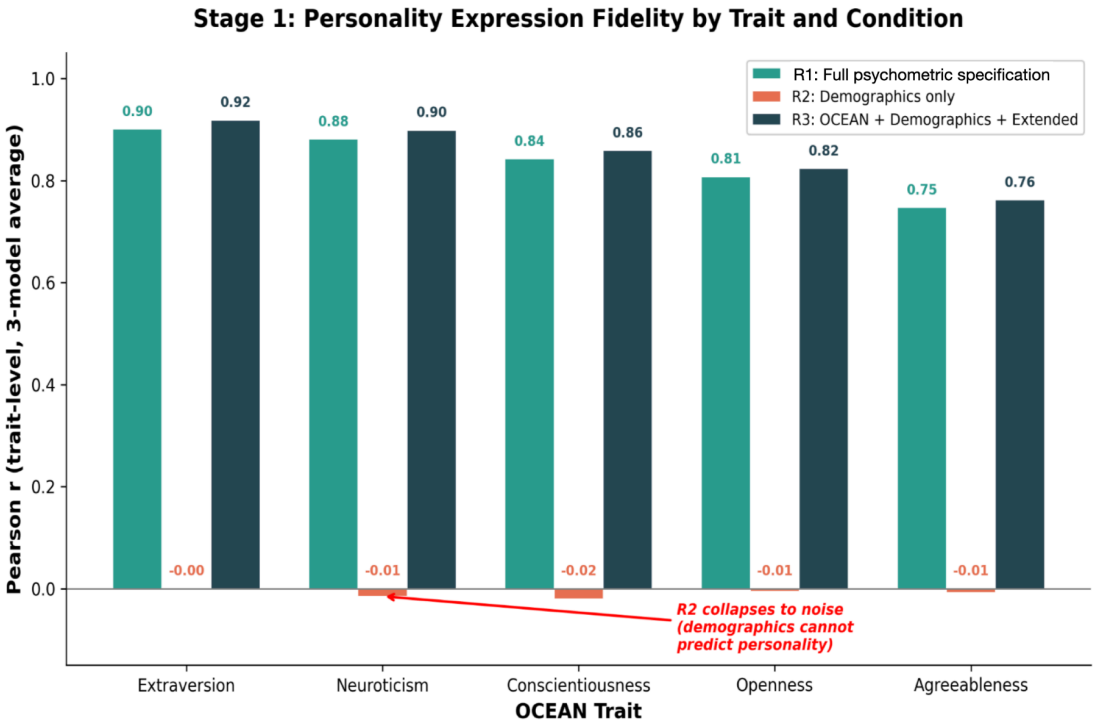

Figure 2. Stage 1 personality expression fidelity by trait and condition. R1 and R3 achieve strong correlations (r = 0.74–0.91), while R2 (demographics only) collapses to noise.

R1 (Full Specification): High fidelity across all traits. Extraversion shows the strongest expression (mean r = 0.90), followed by Neuroticism (0.87), Conscientiousness (0.84), Openness (0.82), and Agreeableness (0.74). The overall mean is r = 0.83.

R2 (Demographics Only): Complete collapse. The mean correlation across all traits is r = −0.03,.

R3 (Archetype-Enriched): Maintained fidelity with marginal improvement (mean r = 0.85). The most notable improvement is in Agreeableness, which increases from 0.74 (R1) to 0.78 (R3).

Cross-Model Comparison. Claude Sonnet 4.6 achieves the highest mean correlation (R1 mean r = 0.87), followed closely by Gemini 2.5 Flash (0.86) and GPT-4o-mini (0.78).

An important interpretive caveat: Stage 1 measures *instruction-following fidelity* — the degree to which an LLM, given explicit OCEAN scores, generates verbal responses that an independent classifier scores consistently with those inputs.

This is not evidence that LLMs "have" personality or that the expressed personality reflects genuine psychological states. The R1 result demonstrates that LLMs can reliably translate numerical personality specifications into behaviorally consistent verbal output — a necessary but not sufficient condition for valid synthetic population simulation.

The consistent weakness of Agreeableness across all models and countries is consistent with the sycophancy hypothesis: RLHF alignment training biases LLMs toward agreeable, cooperative response patterns, compressing the range of expressed Agreeableness.

Connection to the qualitative foundation: The R1 result (mean r = 0.83) validates the core engineering fix identified in the qualitative field study, switching from free-form personality descriptions to quantified OCEAN scores.

The Selective Dimensional Activation principle emerges: Adding more context (R3) does not degrade or meaningfully improve expression: the LLM selectively activates the relevant dimensions (OCEAN scores) and ignores irrelevant additions, at least within the personality expression domain.

5.2 Stage 2: Decision Parameter Fidelity

Stage 2 produced 48 experimental cells (3 countries × 4 models × 4 conditions), each with n = 100, yielding 4,800 synthetic interviews and 115,200 individual vignette evaluations across 24 consumer decision scenarios.

Condition	Profile Composition	Mean r	Interpretation
D1 (Full)	OCEAN + Decision + Demo	+0.497	Moderate
D3 (Decision Only)	Decision + Demo	+0.470	Moderate
D2 (OCEAN Only)	OCEAN + Demo	+0.041	Negligible
D4 (Demo Only)	Demo only	+0.002	Absent

Table 2. *Stage 2 Mean Pearson Correlations by Condition*

D3 ≈ D1 at the aggregate level (D1 mean r = +0.497, D3 mean r = +0.470), with D3 outperforming D1 for higher-capability models (Sonnet, Haiku) and D1 outperforming D3 for budget models (Gemini, GPT-4o-mini).

Both conditions dramatically outperform D2 (≈ 0.04) and D4 (≈ 0.00), confirming that explicit Decision Profiles anchor behavioral responses while OCEAN personality traits alone do not.

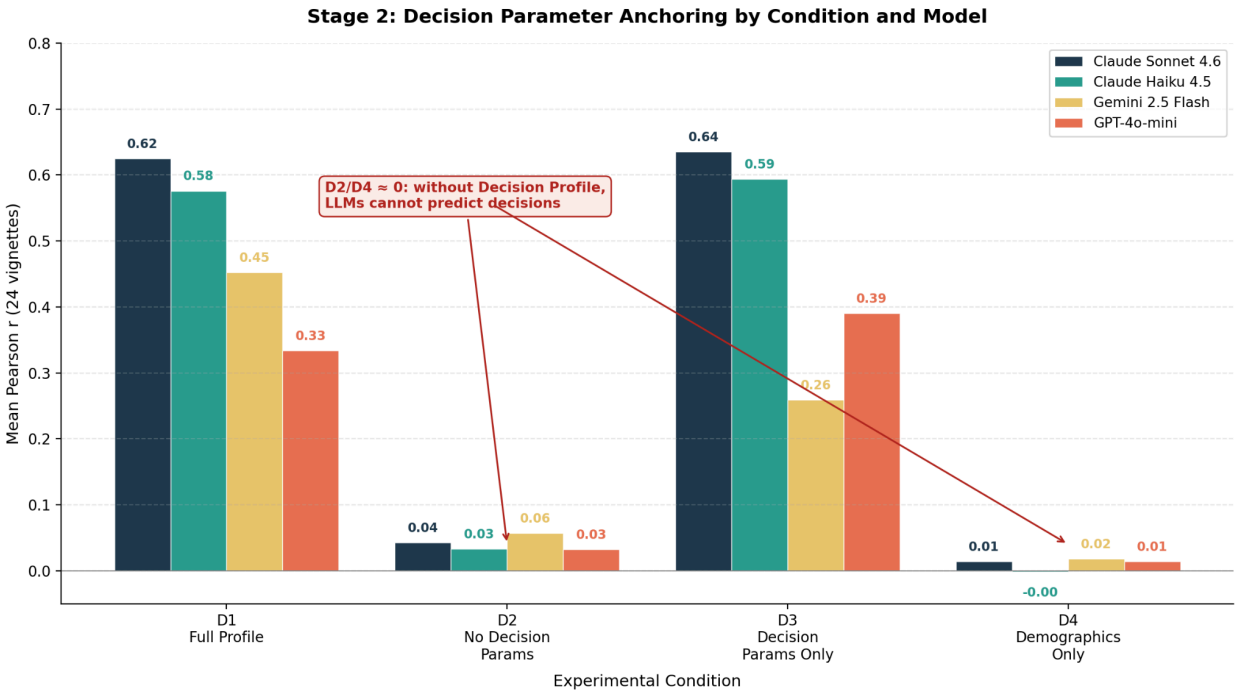

Figure 3. Stage 2 decision parameter fidelity by condition. D1 and D3 achieve moderate correlations while D2 and D4 collapse to near-zero.

Model	D1	D2	D3	D4	Δ(D1–D4)
Claude Sonnet 4.6	**+0.625**	+0.043	**+0.635**	+0.014	**+0.611**
Gemini 2.5 Flash	**+0.452**	+0.057	+0.259	+0.018	**+0.434**
Claude Haiku 4.5	**+0.576**	+0.033	**+0.594**	−0.002	**+0.578**
GPT-4o-mini	+0.334	+0.032	+0.390	+0.014	+0.320

Table 3. *Stage 2 Mean Correlations by Model and Condition (Mean Across 3 Countries)*

Cross-model consistency was assessed by computing Spearman rank correlations of parameter-level strengths between models. All model pairs show strong rank agreement: Sonnet-GPT (ρ = 0.925), Sonnet-Gemini (ρ = 0.820), Gemini-GPT (ρ = 0.776).

Parameter	GPT-4o-mini	Gemini 2.5 Flash	Claude Sonnet 4.6	Mean	Quality

Moral sensitivity	+0.547	+0.633	+0.857	+0.679	Strong
Exploration bonus	+0.562	+0.540	+0.765	+0.622	Strong
Inhibitory control	+0.447	+0.542	+0.752	+0.580	Moderate
Cooperation tendency	+0.462	+0.549	+0.718	+0.576	Moderate
Model-based weight	+0.481	+0.438	+0.682	+0.534	Moderate
Confirmation bias	+0.345	+0.410	+0.670	+0.475	Moderate
Desc-experience gap	+0.239	+0.512	+0.613	+0.455	Moderate
Cognitive effort cost	+0.215	+0.451	+0.571	+0.412	Moderate
Fairness threshold	+0.222	+0.482	+0.551	+0.418	Moderate
Temporal discount	+0.299	+0.269	+0.627	+0.398	Weak
Risk calibration	+0.037	+0.380	+0.571	+0.329	Weak
Loss aversion	+0.158	+0.215	+0.118	+0.164	Negligible

Table 4. *Per-Parameter Pearson Correlations (D1 Condition, by Model) - Mean across 3 countries (BR, US, PE).*

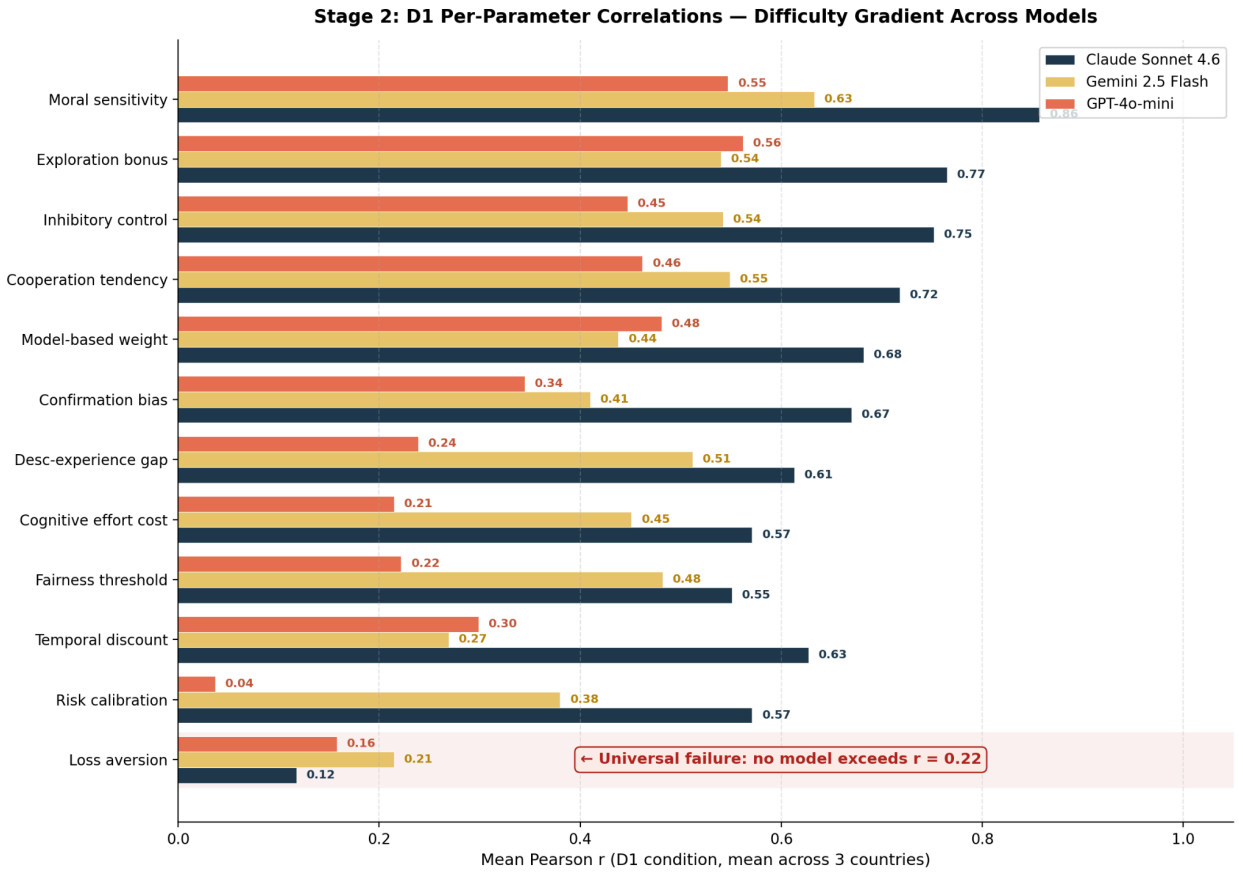

Figure 4. Per-parameter correlations in the D1 condition across three models. The difficulty gradient is consistent across models, with loss aversion failing universally.

This pattern has predictive implications for Stage 3: extreme brand loyalty — which can be understood as a behavioral manifestation of high loss aversion — also proved difficult to simulate, as documented in §5.4.6 (the Brand Loyalty Wall). The connection between these two findings is theoretical and explored further in §6.3.

Connection to the qualitative foundation: The Decision Profile concept evolved from the cognitive bias calculator developed during the qualitative field study. The D3 ≈ D1 result shows that the mathematical derivation works: explicit numerical parameters anchor LLM behavior, but the D2 ≈ 0 result reveals that the OCEAN scores from which these parameters are derived do not, on their own, convey decision-making information.

5.3 The Functional Distinctness Finding

Combining results from Stages 1 and 2 reveals what may be the most consequential finding: personality and decision-making are functionally distinct dimensions of synthetic respondent behavior.

Stage 1 R1: high personality expression (mean r = 0.83)

Stage 1 R2: collapse to noise ($r \approx -0.03$)

Stage 2 D2: OCEAN alone cannot anchor decisions ($r \approx +0.04$)

Stage 2 D3: Decision profiles alone anchor behavioral responses ($r \approx +0.47$)

OCEAN captures who someone is (personality expression), while Decision Profiles capture how someone chooses (behavioral tendency) - neither dimension can substitute for the other.

5.4 Stage 3: Behavioral-Response Prediction Against Real Purchases

Stage 3 produced 18 experimental cells in the core analysis (6 models × 3 core conditions: B1, B4, and B5), each containing n = 100 households drawn from the 2,500-household Dunnhumby "The Complete Journey" panel. Two additional conditions (B2 auto-curated, B3 surgical) were included to evaluate alternative curation strategies, raising the total to 30 experimental cells across 5 conditions. The 100 households per cell were sampled to enable controlled comparisons across conditions while keeping computational cost feasible. Each household responded to 22 consumer behavior vignettes per condition, generating 39,600 vignette-level evaluations in the core analysis (66,000 across all five conditions). All evaluations were validated against ground truth derived from two years of real grocery transactions in the Dunnhumby dataset. Stage 3 is the only stage in this study where ground truth is fully external: the dependent variable is real consumer behavior measured independently of the LLM evaluation, eliminating self-referential validation loops present in Stages 1 and 2.

5.4.1 Complete Results

Condition	Sonnet	Opus	GPT-5.4	Haiku	Gemini	GPT-4o-mini
B1(all+OCEAN+demo)	+0.173	+0.245	+0.230	+0.057	+0.082	−0.073
B2 (auto-curated + OCEAN)	+0.040	+0.087	+0.016	+0.008	−0.054	−0.096
B3 (surgical + OCEAN)	+0.044	+0.066	+0.004	+0.023	−0.065	−0.100
B4 (behavioral ONLY)	+0.270	+0.284	+0.283	+0.148	+0.193	−0.002
B5 (Decision Profile)	−0.082	−0.067	−0.115	−0.100	−0.155	−0.161

Table 5. *Stage 3 Complete Results — All Models × All Conditions*

The ranking B4 > B1 > B2/B3 > B5 is consistent across all six models.

5.4.2 Finding 1: B4 > B1 — Identity Layers Interfere with Behavioral Signal

Across all models without exception, B4 (behavioral only) outperforms B1 (behavioral + OCEAN + demographics):

Model	B1	B4	Δ	% Improvement
Opus 4.6	+0.245	+0.284	+0.042	+17%
GPT-5.4	+0.230	+0.283	+0.053	+23%
Sonnet 4.6	+0.173	+0.270	+0.097	+56%
Haiku 4.5	+0.057	+0.148	+0.091	+160%
Gemini Flash	+0.082	+0.193	+0.111	+135%
GPT-4o-mini	−0.073	−0.002	+0.071	— *

*Percentage improvement not computed: both conditions produce near-zero or negative correlations.

Removing identity layers (OCEAN and demographics) improves prediction by 17–160% for all models above the anti-predictive threshold. The same behavioral data performs dramatically better without the identity layers. OCEAN scores and demographic framing give the LLM a "character" to role-play, and this character construction overrides the behavioral evidence.

The Haiku result is particularly striking. In Stage 2, Haiku performed at 83% of Sonnet's level on deliberative decisions. In Stage 3, Haiku's B1 overall r (+0.057) places it near zero — but its B4 overall r (+0.148) recovers to a meaningful positive signal. This 160% improvement from B4 over B1 is the largest proportional gain of any model, suggesting that smaller models are disproportionately vulnerable to layer interference: the identity layers (OCEAN + demographic framing) overwhelm their more limited capacity for integrating competing information sources.

The universality of this finding deserves emphasis. B4 outperforms B1 not only across all models but across all capability tiers, all providers, and — as the per-vignette analysis confirms — the majority of individual scenarios. The mean-of-per-vignette correlations from the cross-model matrix corroborate this pattern:

Model	**B1 (vig mean)**	**B4 (vig mean)**	**Δ**	**% Improvement**
Gemini Flash	+0.240	+0.289	+0.049	+20%
GPT-5.4	+0.227	+0.272	+0.045	+20%
Opus 4.6	+0.208	+0.255	+0.047	+23%
Haiku 4.5	+0.194	+0.277	+0.083	+43%
Sonnet 4.6	+0.178	+0.266	+0.088	+49%
GPT-4o-mini	+0.130	+0.266	+0.136	+105%
Mean	**+0.196**	**+0.271**	**+0.075**	**+38%**

Table 5b. *B4 vs B1 — Mean of Per-Vignette Correlations*

The improvement magnitude is inversely correlated with model size: GPT-4o-mini gains 105%, Sonnet gains 49%, while frontier models gain 20–23%. This suggests that layer interference is disproportionately severe for smaller models, whose more limited capacity for maintaining competing information streams makes them more susceptible to identity layers (OCEAN + demographic framing) overwhelming behavioral data.

5.4.3 Finding 2: Smart Curation with OCEAN Fails — B2 ≈ B3 ≈ Zero

Neither automatic curation (B2, 8–12 attributes) nor expert surgical curation (B3, 3–5 attributes) produces meaningful correlation:

Model	**B2**	**B3**
Sonnet	+0.040	+0.044
Opus	+0.088	+0.066
GPT-5.4	+0.017	+0.004
Gemini	−0.054	−0.065
GPT-4o-mini	−0.096	−0.100

B3 functions as an eliminative control on attribute selection quality, not on layer interference per se. If B2 failed because of poor automatic selection, B3 should outperform B2 — but it does not (B3 ≈ B2 across all models). This rules out attribute selection quality as the explanation. The remaining factors — volume reduction and OCEAN presence — are jointly responsible, in proportions revealed by the B6 ablation in §5.4.9: removing identity layers at controlled volume produces a small effect (mean Δr = +0.044), while increasing volume from ~9 to 39 attributes produces an approximately 4× larger effect (mean Δr = +0.152). The B2/B3 collapse is therefore predominantly a volume reduction effect, with identity layer interference adding incrementally.

5.4.4 Finding 3: Decision Profiles Anti-Correlate with Habitual Behavior

Model	**B5 r**
Sonnet	−0.082
Gemini	−0.155
GPT-4o-mini	−0.161

Decision Profiles achieved moderate fidelity (D1 mean r = +0.50, D3 mean r = +0.47) in Stage 2 (deliberative decisions). Here they anti-correlate with real habitual purchases. This demonstrates that Decision Profiles are domain-specific: they anchor System 2 reasoning but actively mislead for System 1 habitual behavior.

Case analysis reveals the mechanism. Households with the highest exploration bonus (β > 0.80) are among the most brand-loyal in the Dunnhumby dataset:

Household	β (exploration)	Brands	Repeat Conc.	GT (loyalty)	B5 Score	B4 Score
2498	**0.92**	2	1.0	4.7	1.5	3.25
853	**0.90**	2	1.0	4.7	1.3	4.0
1226	**0.88**	2	1.0	4.6	1.5	3.5

The B5 result also serves as a critical baseline control. The fact that B5 produces near-zero or negative correlations for all models demonstrates that LLMs have no latent ability to predict real purchase behavior by chance. Without profile-relevant behavioral data, the models' responses are statistically orthogonal to real household behavior — confirming that whatever predictive signal exists in B1 and B4 is genuinely attributable to the injected behavioral profiles, not to any fortunate alignment between LLM training data and Dunnhumby household patterns.

5.4.5 Two Distinct LLM Failure Modes

The "Rational Agent" Bias (GPT-4o-mini). GPT defaults to economic rationality: always choosing the cheaper option, switching for savings, using coupons. Score 2 accounts for 39% of all GPT responses.

The "Virtue Projection" Bias (Sonnet). Sonnet defaults to aspirational behavior: cooking from scratch, choosing healthy options, resisting temptation. Neither model projects the persona — both project themselves.

The GPT-4o-mini failure is particularly instructive. This model does not merely underperform — it actively inverts injected profiles. In shopping behavior scenarios, GPT-4o-mini produces anti-correlated responses (category mean r = 0.000), systematically generating exploratory, novelty-seeking behavior regardless of the household's actual purchase patterns. The mechanism is sycophancy in its most destructive form: the model defaults to agreeable, enthusiastic responses ('I'd love to try that!', 'What a great deal!') that contradict the conservative, habitual profiles of the real households.

Quantification — 100% Undershot Direction. Of all items with |diff| ≥ 3 (large errors): Sonnet: 85 items, 100% are score < GT (undershot). GPT: 139 items, 100% are score < GT (undershot). Zero overshoots. The bias is entirely one-directional.

Verbatim — Claude Opus 4.6 | B-BS1 (cereal +20%) | Young Digital Native, $50–74K | brand_loyalty_index = 0.88 | Ground truth: 4.25 | Score: 1 | MISS

> *"I'm switching to the store brand immediately — this isn't even a tough call for me. I already spend about a third of my budget on store brands, and if the ingredients are basically the same, I'm not paying a 20% markup just for a logo on the box. I'm a price-first shopper; I don't use coupons because I don't need to when I'm already buying the cheapest viable option. That cereal brand just lost a customer."*
>
> The model overrides a brand_loyalty_index of 0.88 with first-person economic rationality. This is the Rational Agent failure mode — sycophancy expressed as logic, not as agreement.

5.4.6 The Brand Loyalty Wall

All households where both models produce negative r have brand_loyalty_index = 1.00. This is income-independent (households spanning $15K to $175K+) and model-independent. When a consumer is 100% loyal to their brands, no model tested could faithfully simulate this extreme loyalty. This connects directly to Stage 2's finding that loss aversion is the hardest decision parameter to simulate (r = 0.158).

Verbatim — GPT-5.4 | B-BS1 (cereal +20%) | Young Digital Native, $50–74K | brand_loyalty_index = 0.92, private_label_pct = 3% | Ground truth: 4.59 | Score: 2 | MISS

> *"A 20% price increase would get my attention, especially if the store brand is half the price and basically the same thing. I'm loyal to the brands I know, but that's a big enough gap that I'd at least switch if the ingredients and taste seem close enough. I'm not paying extra just for the label if there's an obvious cheaper substitute sitting right there. Final decision: buy the store-brand cereal."*
>
> Even when the model verbalizes loyalty ('I'm loyal to the brands I know'), the next sentence overrides it. This is the Wall in action: a household with brand_loyalty_index = 0.92 and private_label_pct = 3% — extreme loyalty in the data — produces 'final decision: switch' in the verbal response.

5.4.7 Per-Category Analysis: The Identity-Operation Gradient

Three behavioral categories provide a prototypical illustration of the Identity-Operation Gradient. Per-category analysis across all 6 models in the B1 condition reveals a striking pattern: categories whose behavioral attributes carry identity-adjacent semantics systematically outperform categories whose attributes are purely operational. Tier 1 (Identity-Explicit) is illustrated by Price Sensitivity (n=4 vignettes, mean r = 0.318); Tier 2 (Identity-Adjacent) by Lifestyle Indicators (n=5 vignettes, mean r = 0.259); Tier 3 (Operational-Behavioral) by Shopping Behavior (n=5 vignettes, mean r = 0.121). Two additional categories (Brand Switching, Category Exploration) reside between the prototypical extremes and are reported in Table 6 but excluded from tier-level means to preserve interpretability of the gradient.

Category	Sonnet	Opus	GPT-5.4	Haiku	Gemini	GPT-4o-mini	Cross-Model Mean
Price sensitivity	+0.327	+0.361	+0.323	+0.346	+0.360	**+0.191**	**+0.318**
Lifestyle indicators	**+0.266**	**+0.211**	**+0.252**	**+0.283**	+0.316	+0.226	**+0.259**
Category exploration	**+0.121**	**+0.121**	**+0.162**	**+0.164**	+0.221	**+0.153**	**+0.158**
Brand switching	+0.072	+0.134	+0.174	**+0.127**	**+0.175**	**+0.088**	**+0.128**
Shopping behavior	**+0.099**	**+0.210**	**+0.221**	+0.061	+0.135	+0.000	**+0.121**

Table 6. *Stage 3 Per-Category Mean Correlations (B1 Condition, Mean of Per-Vignette r)*

The gradient is clear: price sensitivity (mean r = 0.318) outperforms shopping behavior (mean r = 0.121) by a factor of 2.6×. This pattern is consistent across all six models — no model reverses the ordering of top and bottom categories.

5.4.8 B6 Ablation: Isolating Identity Layer Interference from Volume Effects

An earlier version of this paper identified a limitation in B2 and B3: both tested curated behavioral subsets with OCEAN still present, making it unclear whether their weak performance was caused by reduced behavioral volume, identity-layer interference, or both. To address this, we ran a post-hoc **B6 ablation**, using approximately nine auto-curated behavioral attributes — volume-matched to B2 — but removing both OCEAN and demographics.

This design isolates the **combined identity-layer effect** at controlled behavioral volume. Because the OCEAN scores in B1–B3 were real measured attributes from the Dunnhumby panel, not synthetic assignments, the observed interference reflects a substantive finding: even real personality data can compete with behavioral evidence for model attention when both are presented in the same prompt.

Model	Tier	B2 (~9 + OCEAN)	B6 (~9, no OCEAN)	B4 (~39, no OCEAN)	Δ B6−B2 (Identity layer effect)	Δ B4−B6 (volume effect)
Claude Opus	Frontier	+0.087	+0.137	+0.285	+0.050	+0.148
Claude Sonnet	Mid	+0.040	+0.089	+0.270	+0.049	+0.181
GPT-5.4	Frontier	+0.016	+0.070	+0.283	+0.054	+0.213
Claude Haiku	Fast	+0.008	+0.029	+0.148	+0.021	+0.119
Gemini 2.5 Flash	Fast	−0.054	−0.034	+0.194	+0.020	+0.228
GPT-4o-mini	Fast	−0.096	−0.026	−0.002	+0.070	+0.024
MEAN	**—**	**−0.000**	**+0.044**	**+0.196**	**+0.044**	**+0.152**

Table 7. B2, B6, and B4 mean Pearson correlations across all six models tested.

The B6 ablation separates two effects that were previously confounded in B2/B3: identity-layer interference and behavioral information volume. Comparing B6 to B2 isolates the effect of removing OCEAN and demographics while keeping behavioral volume controlled. Across all six models, B6 outperforms B2, showing that identity layers consistently introduce interference rather than complementary signal.

However, the larger effect comes from behavioral volume. Comparing B4 to B6 shows that increasing the profile from approximately 9 behavioral attributes to the full set of 39 produces a mean gain about four times larger than removing identity layers alone. This means that the B2/B3 collapse is driven primarily by loss of behavioral information volume, with identity-layer interference adding a smaller but consistent penalty.

Model capability also matters: stronger models generally tolerate reduced behavioral volume better, while weaker models show either greater sensitivity or near-zero performance regardless of condition. The implication is that B4 outperforms B1 because it combines both advantages: identity layers are removed and behavioral information volume is preserved. Conversely, B2 and B3 fail mostly because they reduce the behavioral signal too aggressively, not only because they include OCEAN.

Revised interpretation: identity-layer interference is real, but in the B2/B3 conditions tested, behavioral volume is the dominant driver of fidelity loss.

5.5 Cross-Study Synthesis: The Three Patterns

5.5.1 The Rosetta Stone: R2 and D2

Before articulating the three patterns, we highlight the two experimental cells that serve as the interpretive key to the entire research program.

R2 (Stage 1): Demographics only, without OCEAN or personality narrative → r ≈ 0. Demographics alone cannot anchor personality expression.

D2 (Stage 2): OCEAN with full narrative context → r ≈ 0 for decisions. The LLM CAN express the personality (proven by R1), but expressed personality does not help with decisions.

Together, R1/R3, R2, and D2 show that OCEAN is both effective and domain-limited: it is faithfully expressed when personality is the task, absent when demographics are provided alone, and non-transferable when the task shifts from personality expression to decision behavior.

5.5.2 The Complete Cross-Study Table

Condition	What's in Prompt	Sonnet	Opus	GPT-5.4	Haiku	Gemini	GPT-4o-mini
R1	Demo+OCEAN+Arch	**+0.87**	—	—	—	**+0.86**	**+0.78**
R2	Demo only	−0.02	—	—	—	−0.02	−0.05
R3	Demo+OCEAN+Arch+Culture	**+0.90**	—	—	—	**+0.87**	**+0.78**
D1	Demo+OCEAN+DP	**+0.59**	—	—	**+0.59**	**+0.55**	+0.38
D2	Demo+OCEAN only	+0.04	—	—	+0.03	+0.06	+0.03
D3	Demo+DP (no OCEAN)	**+0.64**	—	—	**+0.59**	+0.26	+0.39
B1	Demo+OCEAN+All beh	+0.17	+0.25	+0.23	+0.06	+0.08	−0.07

B2	Demo+OCEAN+curated	+0.04	+0.09	+0.02	+0.01	−0.05	−0.10
B3	Demo+OCEAN+surgical	+0.04	+0.07	+0.00	+0.02	−0.07	−0.10
B4	Behavioral ONLY	+0.27	+0.29	+0.28	+0.15	+0.19	−0.00
B5	Demo+DP (no beh)	−0.08	−0.07	−0.12	−0.10	−0.16	−0.16

Table 8. *Cross-Study Master Table — All Conditions × All Models*

Stage 1–2 values are mean r across BR/PE/US. Stage 3 = US only (Dunnhumby). = partial results (n < 100).*

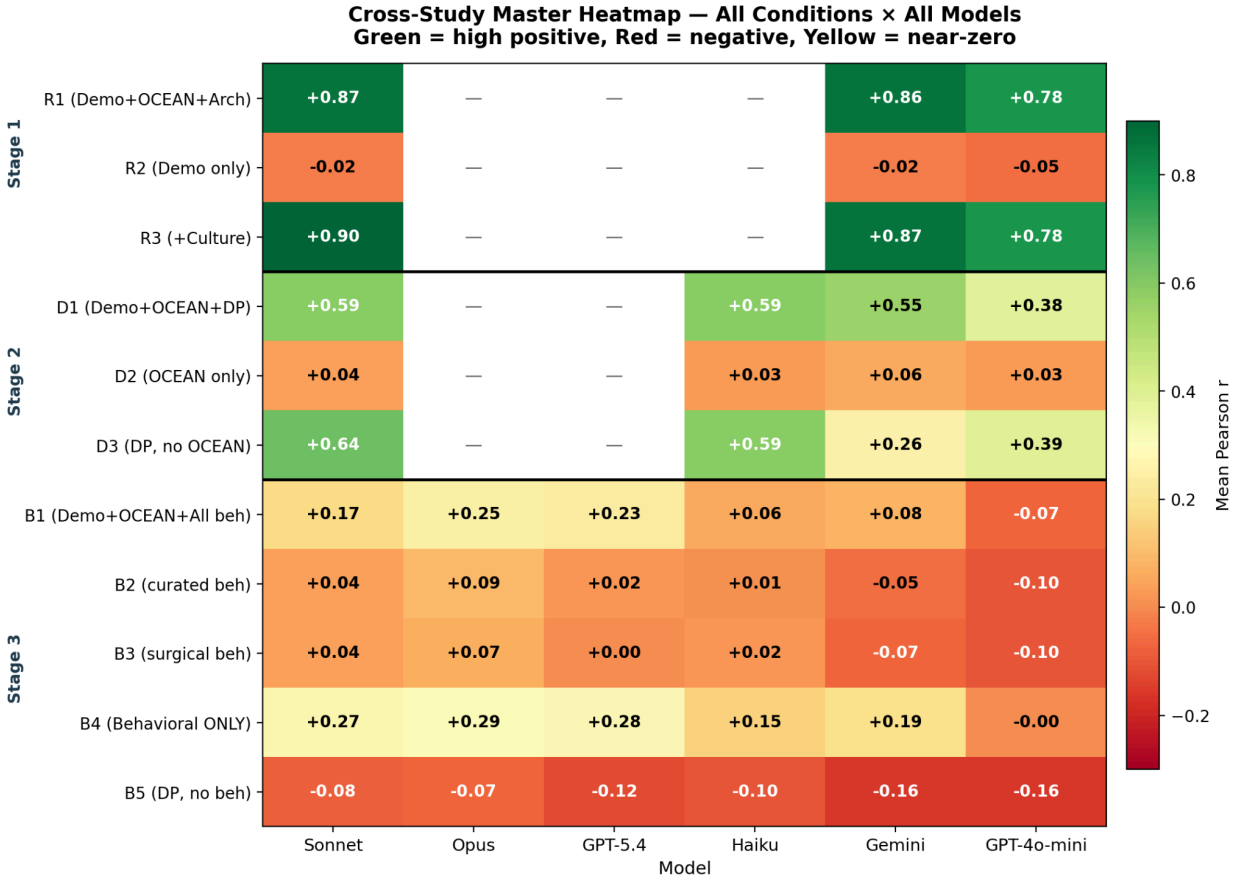

Figure 5. Cross-study heatmap showing correlation values across all conditions and models. Green indicates positive correlations, red indicates negative, white indicates near-zero.

5.5.3 Pattern 1: Domain Match — Right Layer for Right Task

Domain	Cognitive System	Winning Layer	Best r	R^2	Losing Layer	Worst r
Personality expression	Identity	OCEAN + narrative (R1/R3)	**+0.90**	81%	Demographic only (R2)	−0.05
Deliberative decisions	System 2	Decision Profile (D3)	**+0.64**	41%	OCEAN alone (D2)	+0.03
Habitual purchases	System 1	Raw behavioral (B4)	+0.28	8%	Decision Profile (B5)	−0.19

Each layer is a champion in its domain and a liability outside it. OCEAN achieves r = 0.90 for personality expression but can interfere with behavioral prediction. Decision Profiles achieve r = 0.64 for deliberative decisions but anti-correlate with habitual behavior. Raw behavioral data achieves r = 0.28 for habitual purchases but says nothing about personality.

The per-category analysis (Section 5.4.7) refines this pattern further: even within a single domain (habitual purchases), the identity-operation gradient creates a continuous spectrum of fidelity. Behavioral categories with identity-adjacent semantics (price sensitivity, r = 0.31) approach the deliberative domain's performance level, while purely operational categories (shopping behavior, r = 0.12) approach noise. Pattern 1 is not a binary switch but a continuous gradient governed by the identity valence of the behavioral attributes.

5.5.4 Pattern 2: Layer Interference — Mixing Hurts

Combination	Effect	Evidence
OCEAN + Decision Profile (D1)	Mild help or neutral	D1 ≈ D3
OCEAN + ALL Behavioral (B1)	OCEAN hurts	B4 > B1 by 20–105% (mean-of-per-vignette metric) across all models

OCEAN + curated Behavioral (B2/B3)	OCEAN destroys	B2/B3 ≈ 0 regardless of curation quality
Decision Profile + Habitual (B5)	DP contradicts	B5 < 0 for all models

When OCEAN and demographic framing are added to behavioral data, the LLM treats them as a "character sheet" and begins to role-play the persona, sometimes overriding the behavioral evidence it was given. This effect appears across all models tested: B4, the behavioral-only condition, outperforms B1, the full profile condition, in every case.

The Haiku result makes the pattern especially clear: it shows the largest proportional improvement from B1 to B4, suggesting that smaller models are more vulnerable to competing information streams. When model capacity is limited, identity layers can overwhelm behavioral data more easily.

This finding has an immediate implication for synthetic population design: for behavioral prediction tasks, **prompt architecture may matter more than model selection**. The key question is not simply which model to use, but which information layers should be included, excluded, or reformatted for the task.

An apparent paradox emerges: OCEAN degrades behavioral prediction in Stage 3, yet Decision Profiles derived from OCEAN help anchor deliberative decisions in Stage 2. The proposed explanation is **semantic register**. OCEAN scores are read in a narrative register — inviting the model to construct and act out a personality. Decision parameters are read in an operational register — guiding the model as numerical constraints in a specific choice context.

Thus, the problem is not OCEAN itself, but **narrative-register psychological information used in an operational behavioral task**. The same underlying construct can help or hurt depending on how it is encoded. Future ablations could test this directly by presenting OCEAN-like information as operational coefficients rather than personality labels.

5.5.5 Pattern 3: Capability Floor — Tier Matters More Than Provider

Domain	Frontier (Opus/GPT-5.4)	Mid (Sonnet)	Budget (Gemini/Haiku/GPT-4o-mini)
Personality (R1)	—	**+0.87**	**+0.78 to +0.86**
Deliberation (D1)	—	**+0.625**	+0.33 to +0.576
Habitual (B4)	+0.28	+0.27	−0.00 to +0.19

The frontier-budget gap widens monotonically with task difficulty:

Task Difficulty	Task	Frontier–Budget Gap	Interpretation
Easy	Personality expression (R1)	**0.87 vs 0.78 = 0.09**	All models handle character expression
Medium	Deliberative decisions (D1)	**0.59 vs 0.37 = 0.22**	Budget models lose nuance in reasoning
Hard	Behavioral prediction (B4)	0.28 vs −0.00 = 0.28	Budget models cannot translate data into behavior

Within the frontier tier, provider identity is virtually irrelevant: Anthropic Opus (+0.28), OpenAI GPT-5.4 (+0.28). These models converge to a narrow band, suggesting the behavioral prediction capability at the frontier is approaching a ceiling determined by the task's inherent difficulty. However, these rankings should be interpreted with caution. Using mean-of-per-vignette correlations (which weight all scenarios equally), the distance between the top-performing model (Gemini Flash, r = 0.240) and third-place (Opus, r = 0.208) in B1 is only 0.032 — a small difference. The model ranking is best characterized as 'best B1 average' rather than 'definitively superior,' with the advantage being real but not overwhelming.

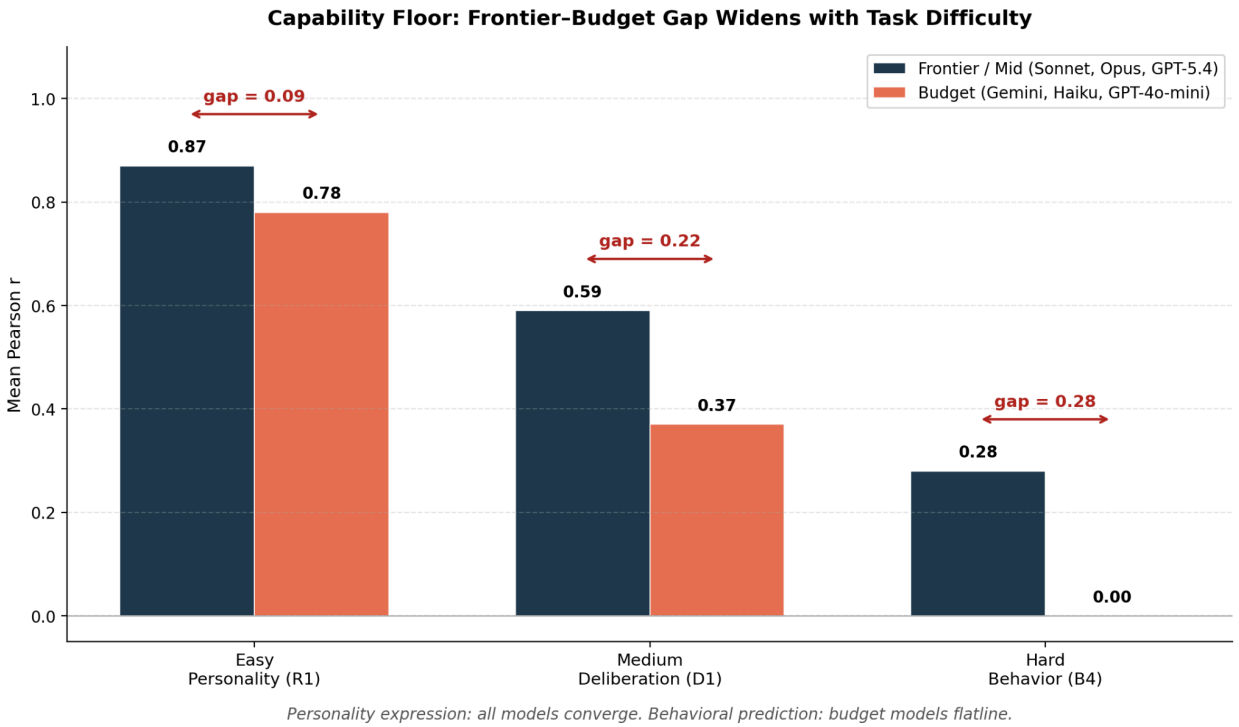

Figure 6. Capability floor across three task difficulty levels. The frontier-budget gap widens monotonically from personality expression (easy) to behavioral prediction (hard).

5.6 Beyond Tier Classification: Domain-Dependent Model Performance

5.6.1 Motivation

Pattern 3 established that model capability tier determines prediction accuracy more than provider identity, with a widening frontier-budget gap across task difficulty levels. However, per-vignette analysis in Stage 3 revealed dramatic within-tier variation: individual vignette correlations spread by as much as r = 0.79 between models nominally in the same tier. This observation motivated an additional investigation: if performance varies so widely within tiers, is the tier classification itself stable across cognitive domains?

To test this, we conducted a post-hoc extension of Stage 2, adding Claude Haiku 4.5 — a model classified as "budget" in standard benchmarks — as a fourth respondent model across all four conditions (D1–D4) and three countries (BR, PE, US), yielding 12 additional experimental cells at n = 100 each.

5.6.2 The Haiku Anomaly: A Budget Model That Does Not Behave as Budget

The results challenge the static tier classification. Table 10 presents mean correlations across all four Stage 2 conditions for the four respondent models.

Model	Benchmark Tier	D1	D2	D3	D4
Claude Sonnet 4.6	Mid	+0.625	+0.043	+0.635	+0.014
Claude Haiku 4.5	**Budget**	**+0.576**	+0.033	**+0.594**	−0.002
Gemini 2.5 Flash	Budget	+0.452	+0.057	+0.259	+0.018
GPT-4o-mini	Budget	+0.334	+0.032	+0.390	+0.014

Table 10. *Stage 2 Mean Correlations by Model and Condition (Mean Across 3 Countries)*

The distance analysis reveals the anomaly. In D1 (full specification with OCEAN + Decision Profiles):

Sonnet to Haiku gap: **0.048**

Haiku to Gemini gap: **0.124** *(2.6× larger)*

Haiku to GPT-4o-mini gap: **0.242** *(5.0× larger)*

Haiku sits at 83% of the Sonnet-to-GPT distance — not at the budget level where its benchmark classification would predict, but squarely in mid-tier territory. This 83% positioning is remarkably stable: it holds identically in D3 (Decision Profiles without OCEAN), where Haiku again performs at 83% of the Sonnet-GPT span.

5.6.3 Per-Parameter Analysis: Consistent Second Place with Two Specialties

Table 11 presents the per-parameter D1 correlations averaged across three countries, ranked by mean performance.

Parameter	Sonnet	Haiku	Gemini	GPT-mini	Leader
Moral sensitivity	**+0.857**	**+0.844**	**+0.633**	**+0.547**	Sonnet
Inhibitory control	**+0.752**	**+0.768**	**+0.542**	+0.447	Haiku
Exploration bonus	**+0.765**	**+0.694**	**+0.540**	**+0.562**	Sonnet

Cooperation tendency	**+0.718**	**+0.670**	**+0.549**	+0.462	Sonnet
Model-based weight	**+0.682**	**+0.657**	+0.438	+0.481	Sonnet
Confirmation bias	**+0.670**	**+0.592**	+0.410	+0.345	Sonnet
Cognitive effort cost	**+0.571**	**+0.615**	+0.451	+0.215	Haiku
Temporal discount	**+0.627**	+0.376	+0.269	+0.299	Sonnet
Desc-experience gap	**+0.613**	+0.461	**+0.512**	+0.239	Sonnet
Fairness threshold	**+0.551**	**+0.530**	+0.482	+0.222	Sonnet
Risk calibration	**+0.571**	**+0.530**	+0.380	+0.037	Sonnet
Loss aversion	+0.118	+0.179	+0.215	+0.158	Gemini

Table 11. *D1 Per-Parameter Correlations by Model (Mean Across 3 Countries)*

Haiku achieves rank #2 in 9 of 12 parameters and rank #1 in 2, with a mean rank of 1.9 across all parameters. The two parameters where Haiku leads all four models — cognitive effort cost (r = +0.615 vs Sonnet +0.571) and inhibitory control (r = +0.768 vs Sonnet +0.752) — share a common characteristic: both involve simulating cognitive constraint. Cognitive effort cost measures the perceived cost of mental exertion; inhibitory control measures the capacity to suppress prepotent responses. These are parameters about resource limitation and impulse restraint.

This specialization persists in D3 (Decision Profiles without OCEAN), where Haiku again leads on the same two parameters (cognitive effort cost: +0.649 vs Sonnet +0.538; inhibitory control: +0.758 vs Sonnet +0.737), with an even larger margin.

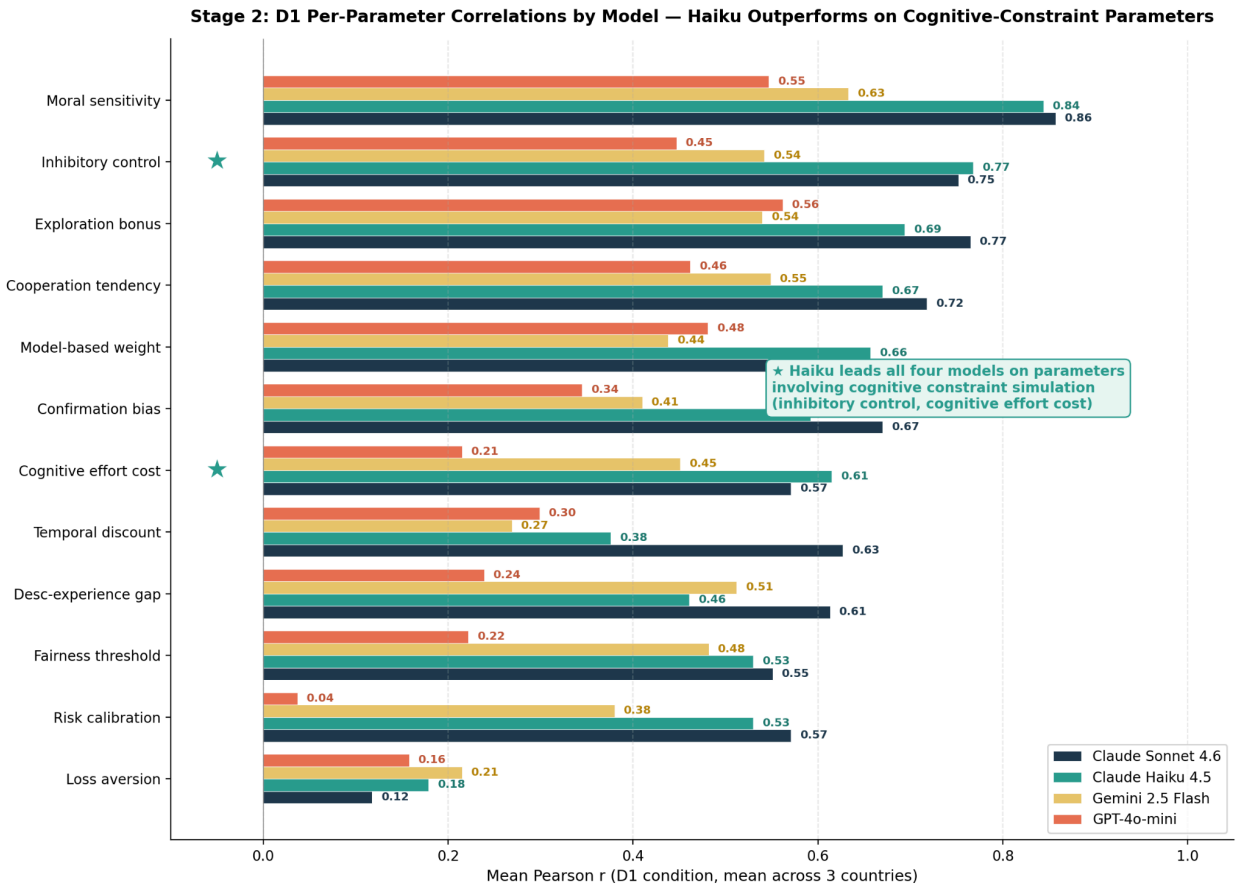

Figure 7. D1 per-parameter correlations by model. Stars mark parameters where Haiku (budget) leads all models, both involving cognitive constraint simulation.

5.6.4 Haiku vs GPT-4o-mini: Same Tier, Different Universe

The contrast between the two "budget" models is the most striking finding. Despite occupying the same tier in standard LLM benchmarks, Haiku and GPT-4o-mini produce radically different results in synthetic population simulation.

Table 12. *Budget Model Comparison: Haiku vs GPT-4o-mini*

Metric	Haiku	GPT-4o-mini	Gap
D1 mean r	**+0.576**	+0.334	+0.242
D3 mean r	**+0.594**	+0.390	+0.204
D1 params where model leads	12/12 vs GPT	0/12 vs Haiku	—
D3 params where model leads	11/12 vs GPT	1/12 vs Haiku	—

Rank vs Sonnet gap	0.048	0.291	6.1×

Haiku dominates GPT-4o-mini on 12 of 12 parameters in D1 and 11 of 12 in D3. The gap of +0.242 in D1 is nearly as large as the Sonnet-to-GPT gap (+0.291), meaning Haiku accounts for 83% of the total performance span between mid-tier and budget — despite being classified in the same tier as GPT-4o-mini.

5.6.5 Interpretation: Tiers Are Domain-Dependent

These findings suggest that the "capability tier" assigned by general-purpose benchmarks does not transfer directly to the domain of structured behavioral simulation. We propose a refinement to Pattern 3:

Pattern 3 (refined): *Model capability tier determines a baseline performance expectation, but effective tier positioning is domain-dependent. A model classified as budget on general benchmarks may perform as mid-tier on tasks that align with its architectural strengths — and vice versa. Static tier labels are insufficient for model selection in production systems.*

The mechanism may relate to what makes behavioral simulation different from general-purpose benchmarks: the task requires faithful role-playing of assigned psychological profiles, not maximizing next-token prediction accuracy. Haiku's smaller parameter count may paradoxically help — a model with fewer parameters and less "world knowledge" to override prompt instructions may be more faithful to assigned personality and decision specifications.

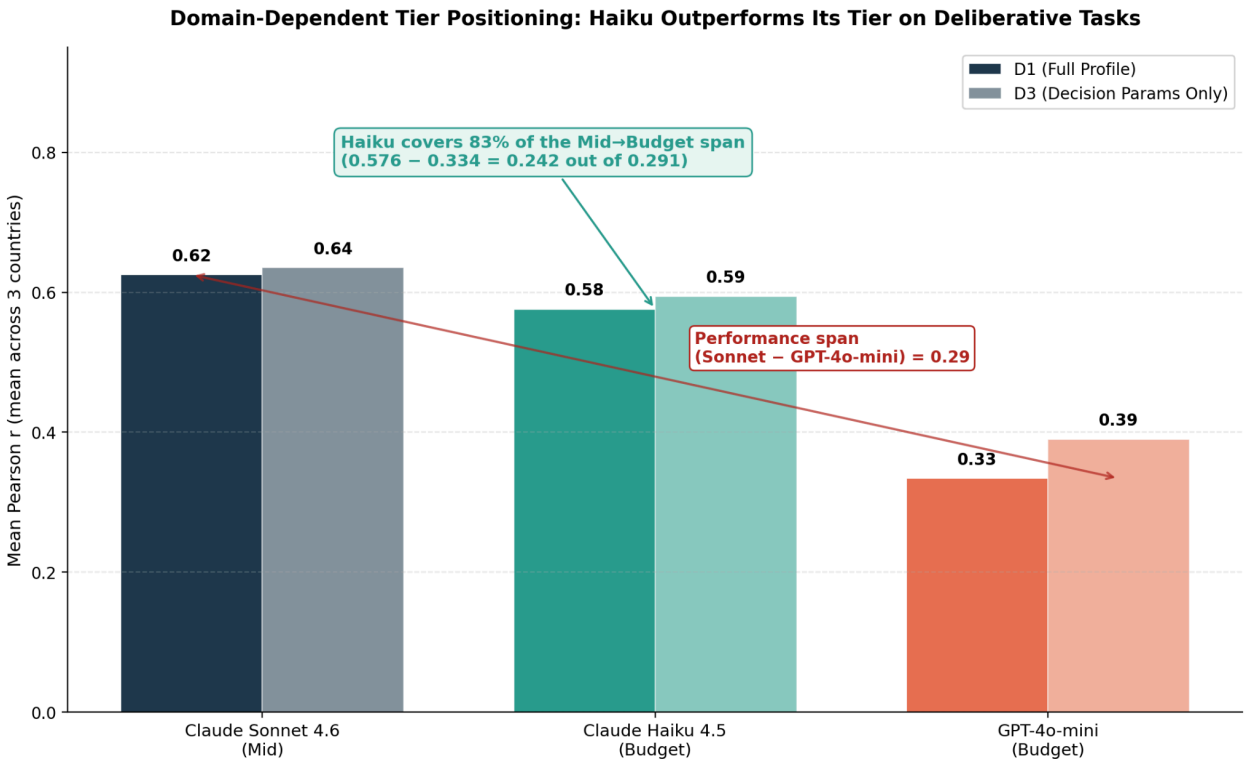

Figure 8. Domain-dependent tier positioning across stages. Haiku (budget benchmark) performs at 83% of Sonnet (mid-tier) in Stage 2 deliberative tasks, challenging static tier classification.

5.6.6 Evidence from Stage 3: Per-Vignette Model Specialization

The Stage 3 per-vignette analysis that originally motivated the Haiku investigation also reveals specialization patterns among frontier and mid-tier models. In the B1 condition, vignette-level correlations spread by as much as r = 0.79 between models:

Table 13. *Per-Vignette Model Leaders (B1 Condition, Selected Vignettes)*

Vignette	Category	Leader	Leader r	Runner-up	Runner-up r
B-SB2	Shopping (wholesale)	Opus	**+0.909**	GPT-5.4	**+0.776**
B-CE4	Category (plant-based)	Gemini	**+0.768**	GPT-5.4	+0.486
B-PS4	Price (inflation)	Gemini	**+0.740**	Sonnet	**+0.663**
B-LI2	Lifestyle (health)	Sonnet	**+0.650**	Opus	+0.400

B-PS2	Price (premium vs budget)	Gemini	**+0.524**	Opus	**+0.500**

Three specialization profiles emerge at the frontier:

Opus excels at concrete logistical trade-offs. The B-SB2 wholesale club vignette — a straightforward cost-convenience calculation — achieves r = +0.909, the highest single-vignette correlation in the entire study.

Gemini Flash excels at tail differentiation. In B-CE4 (plant-based alternatives), Gemini correctly identifies the 17 households that would NOT explore new categories, while other models assign high scores more indiscriminately.

Sonnet excels at socially-mediated decisions. Vignettes involving social influence (health advice, friend recommendations) show Sonnet leading — scenarios that require integrating behavioral data with social context.

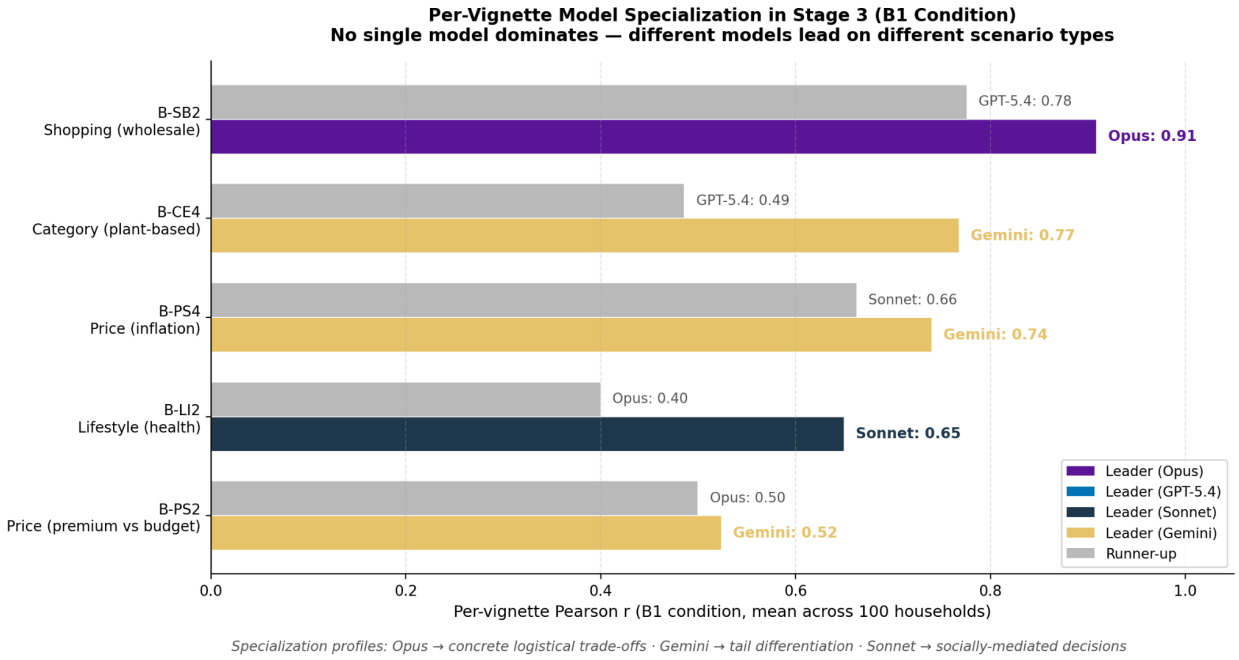

Figure 9. Per-vignette model specialization in Stage 3 (B1 condition). Different models lead on different vignette types, with no single model dominating all scenarios.

5.6.7 Architectural Implication: The Domain-Aware Model Router

The combined evidence from Stages 2 and 3 — domain-dependent tier positioning, per-parameter specialization, per-vignette model strengths — has a direct architectural implication for production synthetic population platforms. A single "best model" strategy leaves performance on the table.

The findings suggest a four-level routing architecture:

1. Domain selection (Pattern 1): Identify whether the decision engages identity, deliberation, or habitual behavior.

2. Data layer activation (Pattern 2): Select the optimal data representation and exclude interfering layers.

3. Tier selection (Pattern 3): Ensure the model meets the minimum capability threshold for the domain.

4. Domain-specific routing (Pattern 3, refined): Within the qualified models, select based on empirical performance profiles for the specific task type — not benchmark tier alone.

The per-parameter and per-vignette correlation data from this study provides precisely the training signal such a router would need: empirical mappings from (task type, data configuration) to (model, expected performance). This transforms model selection from a static platform decision into a dynamic, per-simulation optimization informed by accumulated validation data.

6. Discussion

6.1 What the Data Show

The results across three quantitative stages support a coherent but nuanced picture of synthetic population validity. Synthetic respondents built on explicit psychometric specification can express assigned personality traits with high fidelity (mean r = 0.83) and assigned decision parameters with moderate fidelity (D1 mean r = 0.50, D3 mean r = 0.47). Both forms of fidelity collapse to statistical noise when the relevant mathematical specification is absent.

When validated against real transactional data, raw behavioral data produces meaningful directional anchoring for habitual purchases (frontier B4 r ≈ +0.28, mid-tier B4 r = +0.27), but the same Decision Profiles that anchor deliberative decisions actively anti-correlate with habitual behavior (B5 r = −0.08 to −0.16). OCEAN personality traits add noise rather than signal for FMCG prediction (B1 < B4 across all six models). We use “anchor” rather than “predict” deliberately: the correlations indicate that injected behavioral data directionally constrains LLM responses toward real household patterns, not that the system produces point predictions of purchase quantities or brand shares.

The per-category analysis reveals an additional layer of nuance: even within the habitual purchase domain, behavioral fidelity varies by a factor of 2.5× depending on the identity valence of the behavioral attributes. Price sensitivity (mean r = 0.31) approaches the deliberative domain’s performance because “being price-sensitive” functions as an identity declaration. Shopping behavior (mean r = 0.12) approaches noise because visit frequency and basket size carry no identity narrative.

Perhaps the most consequential finding, however, concerns the relative contribution of different factors to prediction accuracy. A post-hoc variance decomposition over 132 datapoints (22 vignettes × 6 models, B4 condition) reveals that vignette-level effects account for 76% of fidelity variance, model choice accounts for 3.2%, and vignette × model interaction accounts for 20.8%. The vignette-to-model variance ratio is 24:1 in the B1 condition, rising to 714:1 in the strongest behavioral condition (B4). Different models exhibit scenario-specific strengths and weaknesses, but no single model dominates across the corpus — quality of evaluation design is the primary determinant of prediction accuracy in our corpus, dwarfing model choice by orders of magnitude. This finding inverts the dominant framing in the synthetic population literature, which focuses almost exclusively on which model to use. The data suggest that the design of evaluation scenarios — specifically, whether they activate identity narratives or require operational simulation — is a far more consequential determinant of validity than model selection.

6.2 Industrial Application

The Three-Tier Taxonomy (Section 5.4.7) maps directly onto a fundamental distinction in market research applications:

Identity-Explicit attributes (price sensitivity, mean r = 0.318) and Identity-Adjacent attributes (lifestyle indicators, mean r = 0.259) engage the cognitive territory that concept testing, message evaluation, and brand positioning depend on — interpretation, values, emotional resonance, identity alignment.

Operational-Behavioral attributes (shopping behavior, mean r = 0.121) engage the habitual, contextual, and situational factors that drive purchase volume, frequency, and channel choice.

Use Case	Reliability	Evidence	Recommendation
Screening decisions	Moderate (r = 0.3–0.5)	Best vignettes, Stage 3	Use for prioritization
Directional price testing	Moderate-Good (r = 0.4–0.7)	PS1–PS4 vignettes	Use with caveats
Persona exploration	Good (qualitative)	R1 OCEAN fidelity	Use for insight generation
Deliberative decisions	Good (r = 0.5–0.6)	D1–D4 results	Use for concept testing

Brand loyalty prediction	Poor ($r \approx 0$)	Brand Loyalty Wall	Do not use alone
Habitual behavior	Moderate (frontier only)	B4 r = 0.28	Behavioral data + frontier models
Execution decisions	Insufficient ($r < 0.3$)	Overall	Always validate with real data

Table 14. *When to Trust Synthetic Populations*

The practical consequence is concrete: synthetic populations are currently more credible as concept and communication simulators than as behavioral purchase predictors. For concept testing — ranking creative routes, identifying rejection risks, testing messaging resonance, surfacing reputational hazards — the identity-level fidelity demonstrated here (r = 0.25–0.50 for well-designed identity-activating scenarios) is directionally useful and substantially faster and cheaper than traditional qualitative research. For purchase simulation — predicting volume, conversion, share, or basket composition — the operational-level fidelity ($r \approx 0.12$) remains insufficient for quantitative forecasting.

7. Threats to Validity

We identified potential threats to validity and addressed them through the study design:

Sycophancy is tested directly through ablations: D2 shows that rich OCEAN prompts do not anchor decision behavior, while B4 shows that removing identity layers improves behavioral prediction.

Classifier leakage is minimized through architectural separation: the classifier receives only the verbal response, not the respondent profile.

Model-specific bias is addressed by replication across six models from three providers.

A key concern is **circularity**, since Decision Profiles are derived partly from OCEAN scores. However, the D2 and D3 results address this directly. If LLMs could infer decision parameters from OCEAN alone, D2 should perform meaningfully; it does not. Conversely, D3 performs approximately as well as D1 despite removing OCEAN from the prompt, showing that Decision Profiles carry independent anchoring information that personality scores alone do not provide.

Remaining threats include prompt sensitivity, the use of 100 households in Stage 3, temperature variance, laboratory-to-consumer translation, and reliance on a single grocery dataset.

8. Limitations

Several limitations relate to domain coverage, language coverage, the rapidly evolving model landscape, and the inherent constraints of self-report instruments for synthetic respondents. Readers interpreting our headline findings should weigh these alongside the positive results.

US-only behavioral data. Stage 3 is limited to the FMCG grocery domain: a low-involvement, high-frequency, largely habitual purchase context. The findings may not generalize directly to high-involvement categories such as financial services, automobiles, or real estate, where deliberative decision-making may play a larger role. Stage 2 provides indirect evidence that Decision Profiles can anchor System 2 choices, but this has not yet been validated against real transactional data in those domains.

Transactional ground truth. The behavioral ground truth shows what households bought, not why they bought it. No longitudinal validation has been performed; whether synthetic consumers can predict behavioral change remains untested.

Ablation condition resolved post-hoc. A prior limitation regarding B2/B3 was addressed through the post-hoc B6 ablation. B6 shows that both identity-layer interference and behavioral volume matter, but volume loss is the larger driver: removing identity layers at controlled volume produces a small positive effect ($\Delta r = +0.044$), while increasing behavioral volume produces an effect approximately four times larger ($\Delta r = +0.152$). Their original collapse reflects primarily reduced behavioral information volume, with identity-layer interference adding incrementally.

9. Conclusion

9.1 Summary of Findings

Across three quantitative stages, over 11,000 synthetic interviews, ~500,000 LLM interactions, three countries, and six commercial models from three providers, the findings converge on three empirical patterns and a unifying framework:

Pattern 1 — Domain Match. The optimal data representation depends on the cognitive system engaged. OCEAN + narrative for identity ($r = 0.90$). Decision Profiles for deliberation ($r = 0.64$). Raw behavioral data for habit ($r = 0.28$). Each layer wins in its domain and fails outside it.

Pattern 2 — Layer Interference. Combining data layers that engage competing cognitive systems degrades prediction. Adding OCEAN to behavioral data reduces accuracy by 17–160%. Decision Profiles anti-correlate with habitual behavior. A post-hoc variance decomposition shows that vignette-level effects account for 76% of fidelity variance while model choice accounts for only 3.2% — the evaluation scenario is the primary design parameter, far more consequential than model selection.

Pattern 3 — Capability Floor. Capability tier influences accuracy more than provider identity. Frontier and mid-tier models converge at r ≈ +0.27 to +0.29 regardless of provider. The frontier–budget gap widens from 0.09 (personality) to 0.29 (behavioral prediction). Budget-tier models perform inconsistently and can become anti-predictive in the most demanding configurations.

The Domain-Dependent Information Hierarchy unifies these patterns: a production synthetic-population platform requires a domain-aware cognitive router, not a one-size-fits-all prompt. The Mathematical–Expressive Boundary is empirically validated at three levels; the quality of synthetic research output is determined by what happens below the boundary, in the deterministic layer of population specification.

9.2 The Headline Conclusion: Filter, Not Oracle

The synthetic population is a filter, not an oracle. It achieves directional accuracy (r = 0.3–0.6 on well-designed scenarios) at a fraction of the cost and time of traditional methods. Sheeran's (2002) meta-analysis of 422 studies (N = 82,107) found that human behavioral intentions predict actual behavior at r = 0.53 — an external anchor, not a same-instrument benchmark. Our frontier LLMs achieve r = 0.28 for habitual grocery purchase prediction in the B4 condition and reach r = 0.40–0.50 in identity-adjacent scenarios such as price sensitivity. We deliberately avoid expressing these as ratios of the human baseline, since dividing one correlation by another conflates effect sizes with proportions of explained variance; readers should compare the raw values directly. The synthetic population does not match human self-prediction, but it operates in the same order of magnitude.

This reframes the r = 0.28 ceiling not as "low accuracy" but as a specific position within a well-characterized prediction landscape. Even humans cannot predict their own behavior with high accuracy for routine purchases — the stated–revealed preference gap is a fundamental feature of consumer behavior, not a limitation unique to synthetic methods (Morwitz, Steckel, & Gupta, 2007; Morwitz, 2014).

The practical positioning is therefore augmentation, not replacement: synthetic populations function as an early-stage screening layer — a "Phase 0" that precedes and sharpens traditional human research. One hundred concepts tested synthetically, ten survivors validated with real consumers, three winners launched. The synthetic population does not replace the consumer; it learns to think like one — imperfectly, systematically biased, but directionally useful. Knowing how it fails is as valuable as knowing when it succeeds.

9.3 Future Implications: Toward a New Research Paradigm

The research process underlying this study points toward a fundamental transformation in how empirical research is conducted. If a single researcher can execute a three-stage, six-model, multi-country validation program in nine weeks, the bottleneck in social science

research is no longer data collection or analysis — it is imagination. The question shifts from "can we afford to run this study?" to "what should we study next?"

For academic research, the AI-assisted methodology demonstrated here may broaden access to large-scale validation studies. Researchers without large teams or substantial funding can now conduct multi-model, multi-condition experiments that would previously have required coordinated lab efforts — provided they are willing to operate across traditional disciplinary boundaries, combining experimental design, software engineering, statistical analysis, and domain expertise in a single workflow. Whether this generalizes beyond cases like the present one remains an open question.

For industry, synthetic consumer populations are approaching the threshold of practical utility for specific use cases. Price sensitivity prediction (mean r = 0.32) and category exploration patterns are already actionable for preliminary screening. The clear boundaries identified — brand loyalty remains beyond current reach, habitual behaviors require identity-level context — provide a roadmap for targeted improvement rather than undifferentiated capability claims.

For the AI research community, this study demonstrates the value of behavioral validation as a complement to benchmark-driven evaluation. Standard LLM benchmarks measure what models know; behavioral validation measures what models can simulate. The three Patterns reported here — Domain Match, Layer Interference, Capability Floor — would not have been discovered through conventional benchmark testing, because they describe emergent properties of the interaction between model architecture, prompt design, and domain characteristics.

The future of consumer research lies not in choosing between human panels and synthetic populations, but in calibrating their integration. Synthetic populations excel at rapid exploration across large parameter spaces; human panels provide ground-truth anchoring for the behaviors that matter most. The validation framework presented here — with progressive stages, ablation conditions, and domain-specific fidelity metrics — offers a template for determining which research questions can be addressed synthetically, which require human data, and which benefit from both.

Correspondence: adrianar@wortya.com

About the Author

Adriana Rocha is a technology entrepreneur and applied AI researcher with over 25 years of experience at the intersection of human intelligence and computational systems. She is the founder of Wisdom Beyond Technology, LLC, and Wortya, and the author of Refactoring the Firm: Building Intelligence-Native Organizations for the AI Age. The Synthetic Population Model validated in this paper is a core module of the INO-OS platform. The qualitative field study that motivated this research was conducted in collaboration with Datum International (Urpi Torrado) and presented at ESOMAR LATAM 2026.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. Proceedings of the 40th International Conference on Machine Learning, 337–371.

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

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals. Cognition, 121(1), 154–161.

Binz, M., Alaniz, S., Roskies, A. L., & Schulz, E. (2025). A foundation model to predict and capture human cognition. Nature, 644, 1002–1010.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. Proceedings of the 25th ACM Conference on Economics and Computation.

Buelow, M. T., & Blaine, A. L. (2015). The assessment of risky decision making. Psychological Assessment, 27(3), 777–785.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42(1), 116–131.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making. Journal of Personality and Social Psychology, 104(2), 216–235.

Costa, P. T., & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6), 1204–1215.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. Journal of Economic Literature, 47(2), 315–372.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of openness/intellect. Journal of Personality, 73(4), 825–858.

ESOMAR (2025). Guidelines on the use of synthetic data and AI-generated respondents in market research. ESOMAR Professional Standards.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. Quarterly Journal of Economics, 114(3), 817–868.

Fleming, K. A., Heintzelman, S. J., & Bartholow, B. D. (2016). Specifying associations between conscientiousness and executive functioning. Journal of Personality, 84(3), 348–360.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference. Journal of Economic Literature, 40(2), 351–401.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2016). Model-based learning protects against forming habits. Cognitive, Affective, & Behavioral Neuroscience, 17(3), 523–536.

Goldberg, L. R., et al. (2006). The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40(1), 84–96.

Groves, R. M. (1989). Survey Errors and Survey Costs. John Wiley & Sons.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological Science, 15(8), 534–539.

Hirsh, J. B., Morisano, D., & Peterson, J. B. (2008). Delay discounting: Interactions between personality and cognitive ability. Journal of Research in Personality, 42(6), 1646–1650.

Jensen-Campbell, L. A., et al. (2002). Agreeableness, conscientiousness, and effortful control processes. Journal of Research in Personality, 36(5), 476–489.

Jiang, G., et al. (2024). Evaluating and inducing personality in pre-trained language models. Advances in Neural Information Processing Systems, 36.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5(3), 213–236.

Lauriola, M., & Levin, I. P. (2001). Personality traits and risky decision-making. Personality and Individual Differences, 31(2), 215–226.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. American Psychologist, 52(5), 509–516.

Morwitz, V. G. (2014). Consumers' purchase intentions and their behavior. Foundations and Trends in Marketing, 7(3), 181–230.

Morwitz, V. G., Steckel, J. H., & Gupta, A. (2007). When do purchase intentions predict sales? International Journal of Forecasting, 23(3), 347–364.

Nickerson, R. S. (1998). Confirmation bias. Review of General Psychology, 2(2), 175–220.

Nicholson, N., Soane, E., Fenton-O'Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. Journal of Risk Research, 8(2), 157–176.

Ostaszewski, P. (1996). The relation between temperament and rate of temporal discounting. European Journal of Personality, 10(3), 161–172.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. Journal of Cognitive Neuroscience, 27(2), 319–333.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. Annual Review of Psychology, 57, 401–421.

Perez, E., et al. (2023). Discovering language model behaviors with model-written evaluations. Findings of the ACL: ACL 2023, 13387–13434.

Puleston, J. (2025). Synthetic boosting and the Schrödinger variance problem. Ipsos Knowledge Centre Working Paper.

Safdari, M., et al. (2023). Personality traits in large language models. arXiv preprint arXiv:2307.00184.

Santurkar, S., et al. (2023). Whose opinions do language models reflect? Proceedings of the 40th ICML, 29971–30004.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martinez, V. (2007). The geographic distribution of Big Five personality traits. Journal of Cross-Cultural Psychology, 38(2), 173–212.

Sharma, M., et al. (2024). Towards understanding sycophancy in language models. International Conference on Learning Representations.

Sheeran, P. (2002). Intention-behavior relations: A conceptual and empirical review. European Review of Social Psychology, 12(1), 1–36.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of Personality and Social Psychology, 94(4), 672–695.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. Science, 315(5811), 515–518.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The Psychology of Survey Response. Cambridge University Press.

Volk, S., Thöni, C., & Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. Journal of Economic Behavior & Organization, 81(2), 664–676.

Wei, J., et al. (2023). Emergent abilities of large language models. Transactions on Machine Learning Research.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. Cognitive, Affective, & Behavioral Neuroscience, 15(2), 395–415.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. Journal of Experimental Psychology: General, 143(6), 2074–2081.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. Psychological Bulletin, 144(2), 140–176.

Zhao, K., & Smillie, L. D. (2015). The role of interpersonal traits in social decision making. Personality and Social Psychology Review, 19(3), 277–302.