



International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Adaptive Multi-Agent AI System for Synthetic Data Generation and Consumer Persona Simulation: Architecture, Implementation, and Empirical Evaluation

Prof. U.A.S. Gani, Saurabh Sharma, Sarthak Bankar, Shree Larokar, Aditya Tiwari

Department of Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur - 440019 Rashtrasant Tukdoji Maharaj Nagpur University | 2025-2026

ABSTRACT

This paper presents the design, architectural implementation, and empirical evaluation of a unified Two-Agent Artificial Intelligence System addressing two core operational challenges in data science and product marketing: (1) generating realistic, statistically coherent tabular datasets without requiring exposure to sensitive authentic records, and (2) simulating pre-launch consumer psychographic reactions and purchasing intent across demographically diverse cohorts without expensive focus groups. The system implements two cooperating agents orchestrated under a unified web interface: **Agent 1 (Synthetic Data Generation Agent)**, an 8-stage description-driven tabular synthesis pipeline that decouples schema planning from row sampling via large language models and parametric statistical distributions; and **Agent 2 (Consumer Persona Simulation Agent)**, a 7-stage cognitive simulation pipeline that injects first-person psychographic conditioning across five benchmark personas, performs population-level simulation over $N = 1,000$ synthetic Indian consumers, and synthesizes price-conditioned customer reviews. Both agents share a fault-tolerant Gemini API fallback architecture (gemini-3.1-flash-lite $\rightarrow$ gemini-3.5-flash-lite $\rightarrow$ gemini-3.6-flash $\rightarrow$ deterministic fallback). Extensive empirical evaluations demonstrate that Agent 1 generates datasets with preserved inter-field correlations and competitive downstream ML utility (ROC-AUC 0.75-0.79), while Agent 2 accurately differentiates product value propositions and decisively detects credibility collapses that generic LLM prompting fails to identify.

Keywords: Synthetic data generation, consumer persona simulation, multi-agent AI, LLM evaluation, marketing intelligence, adaptive pipeline, Gemini API, machine learning

1. INTRODUCTION

Modern marketing, product design, and machine learning teams face two acute operational bottlenecks. First, evaluating consumer resonance for a new marketing campaign, ad copy, or product positioning conventionally requires focus group studies, surveys, or live A/B tests. These traditional approaches are cost-prohibitive for early-stage ventures, require weeks of recruitment, and carry significant commercial risk if an asset with flawed value positioning is launched live. Second, developing predictive models and testing analytical pipelines demands voluminous labeled structured data. In sectors such as e-commerce, healthcare, and financial services, authentic data is strictly proprietary, subject to privacy regulations (GDPR, DPDP Act), or simply unavailable prior to product launch.

To overcome both challenges, this project presents an integrated, modular **Two-Agent AI Marketing Intelligence System**. Rather than addressing data synthesis and market testing in isolation, we demonstrate a unified architecture where generative language models and statistical engines cooperate under shared coordination. Agent 1 autonomously interprets free-form natural-language requirements into structured, relational tabular datasets. Agent 2 ingests marketing assets and evaluates them through psychographically conditioned persona panels grounded in the demographic diversity of the Indian consumer market.

The primary contributions of this paper are fourfold: (a) a formal proposed architecture and complete working implementation of both Agent 1 and Agent 2; (b) an empirical evaluation of Agent 1's schema coherence, distribution modeling, and downstream classification utility benchmarked against authentic reference data; (c) an empirical evaluation of Agent 2 across five benchmark personas and three contrasting product assets, verifying score stability ($N=5$ runs), population modeling ($N=1,000$), and generic ablation; and (d) a single-page unified web application with dynamic glassmorphism interface and multi-model fallback resiliency.

2. BACKGROUND AND MOTIVATION

2.1 Synthetic Tabular Data Generation

Synthetic data generation has emerged as a cornerstone of modern data-centric AI. Classical statistical methods, such as Gaussian Copulas, Bayesian Networks, and SMOTE oversampling, model joint distributions from existing tables. Deep generative frameworks, including Conditional GANs (CTGAN), Variational Autoencoders (TVAE), and diffusion models (TabDDPM), capture complex non-linear interactions [1, 2]. However, all established generative paradigms fundamentally presuppose the prior existence of an authentic training dataset and a manually curated schema. When an organisation operates in a zero-data cold-start regime, these models cannot function. Recent developments in foundation LLMs offer a new paradigm: description-driven synthesis, where domain knowledge encoded in model weights enables zero-shot schema and relational rule generation from plain English prompts [1].

2.2 Consumer Persona Simulation

Simulating human behavioral responses through LLMs has advanced rapidly following the introduction of generative agents and algorithmic persona conditioning [4, 5]. Rather than prompting an LLM as an unconditioned 'average consumer' - which produces flattened, homogeneous praise - researchers inject detailed demographic attributes, socioeconomic constraints, and first-person psychographic manifestos. This conditioning elicits differentiated cognitive reasoning, uncovering distinct friction points across customer segments. However, a key unresolved research question is whether persona simulation exhibits acceptable score stability across repeated runs and whether simulation variance signals genuine consumer ambiguity rather than stochastic noise [5, 7].

3. PROBLEM STATEMENT

Existing tools for tabular data synthesis and consumer market research operate as disconnected, expensive silos. Market research firms charge substantial fees for small focus group panels that take weeks to coordinate, while standalone synthetic data tools either require full exposure to private authentic data or output naive, statistically independent random values that destroy crucial inter-attribute correlations (e.g. failing to correlate years of experience with salary, or restaurant age with review volume).

There is an acute need for a unified, low-marginal-cost multi-agent pipeline that can: (a) accept unstructured natural-language dataset specifications and automatically generate statistically plausible, relational tabular datasets; (b) ingest raw marketing copy and simulate first-person cognitive reactions, quantitative ratings, and specific objection points across diverse Indian consumer cohorts; (c) scale individual persona feedback into macro population-level conversion distributions ($N = 1,000$); and (d) execute reliably within a resilient web application environment without catastrophic API failure.

4. RESEARCH GAP

While both synthetic data generation and generative agent simulation have received individual academic attention, significant research gaps remain at their intersection:

- a) **Zero-Data Tabular Generation:** Most tabular synthesizers rely on real training data. The gap is an adaptive pipeline that infers realistic schemas, statistical distributions, and correlation constraints purely from natural language without training data exposure.
- b) **Demographic Conditioning vs. Generic Simulation:** Generic LLM evaluations suffer from severe acquiescence bias and cannot detect implausible price-to-value anomalies. A formal evaluation comparing conditioned persona panels against unconditioned baselines is lacking.
- c) **Empirical Stability Bounds:** Few empirical studies measure the standard deviation of LLM persona scoring across repeated independent runs to establish stability thresholds.
- d) **Integrated Multi-Agent Workflow:** No reviewed platform unites zero-shot tabular dataset creation and persona-based marketing simulation within a single coherent architecture.

5. OBJECTIVES

1. To develop Agent 1 (Synthetic Data Generation): an 8-stage pipeline that parses natural-language requirements, designs domain-specific schemas, plans statistical distributions, and synthesizes relational tabular datasets.
2. To develop Agent 2 (Consumer Persona Simulation): a 7-stage pipeline that performs first-person cognitive simulation, quantitative scoring (1-10 resonance, 0-100% intent, 1-10 credibility), and qualitative friction analysis across five benchmark Indian personas.
3. To scale Agent 2 feedback to $N = 1,000$ synthetic consumers to estimate population sentiment distributions and top conversion objections, supplemented with price-conditioned customer reviews.

- 4. To implement a resilient multi-model fallback chain (Gemini 3.1 Flash Lite -> 3.5 Flash Lite -> 3.6 Flash -> deterministic) ensuring 100% pipeline completion under rate limits or API outages.
- 5. To deploy a unified single-page web interface with real-time progress visualization.
- 6. To empirically evaluate both agents using benchmark datasets, stability tests (N=5 runs), and baseline ablation comparisons.

6. SYSTEM ARCHITECTURE

The system architecture coordinates two specialized autonomous agents through a unified application controller and shared Gemini API backend, as summarized in Table 1.

Component	Core Responsibility	Key Technologies	Status
Agent 1 Synthetic Data Gen	8-stage pipeline: natural language parsing, schema design, distribution planning, relationship modeling, and export	Gemini LLM, NumPy, Faker, Pandas, SciPy, OpenPyXL	FULLY IMPLEMENTED & EVALUATED
Agent 2 Persona Simulation	7-stage pipeline: asset deconstruction, cognitive simulation, 5-persona panel evaluation, N=1,000 population modeling, and review generation	Gemini LLM (T=0.3-0.8), psychographic conditioning, statistical aggregation	FULLY IMPLEMENTED & EVALUATED
Orchestration & Web Interface	Single-page glassmorphism UI with tab switcher, REST API endpoints, multi-model fallback chain, and background video rendering	Flask, HTML5/CSS3, Fetch API, Dotenv, Watchdog reloader	FULLY IMPLEMENTED & OPERATIONAL

Table 1: Unified Two-Agent System Overview and Implementation Status.

6.1 Agent 1 - Synthetic Data Generation Pipeline

Agent 1 executes an 8-stage sequential pipeline to generate structured tabular datasets:

Stage 1 - Requirement Analysis: Employs Gemini to deconstruct natural-language data prompts into structured entities, domains, geographical scopes, and target record counts.

Stage 2 - Schema Planning: Designs field names, semantic data types, and realistic descriptions.

Stage 3 - Distribution Planning: Maps each field to parameterized statistical distributions (Gaussian, lognormal, uniform, categorical, or Bernoulli).

Stage 4 - Relationship Planning: Injects mathematical dependency rules between fields (e.g. positive correlation between experience and salary; tenure and lead conversion).

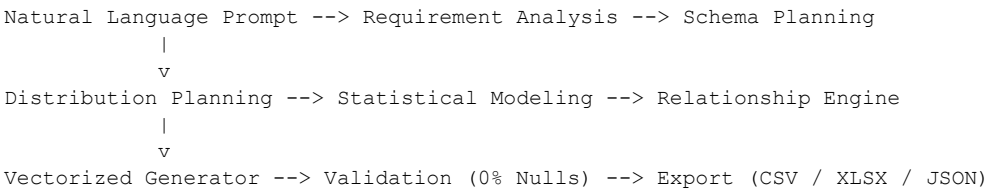
Stage 5 - Record Generation: Generates base values using NumPy random generators and Faker (Indian locale), then applies the relationship engine to enforce cross-column covariance.

Stage 6 - Validation: Assesses null rates, duplicate rows, range violations, and correlation matrices.

Stage 7 - Reporting: Computes descriptive statistics (mean, median, standard deviation) for all fields.

Stage 8 - Export: Emits validated datasets in CSV, formatted XLSX (freeze-paned), and JSON.

Agent 1 Pipeline Flow:



6.2 Agent 2 - Consumer Persona Simulation Pipeline

Agent 2 implements a 7-stage cognitive pipeline designed for marketing asset evaluation:

Stage 1 - Asset Analyzer (T=0.3): Extracts product name, category, price-to-value rating (Exceptional Deal

/ Fair Market Price / Slightly Overpriced / Overpriced), tone, and value claims.

Stage 2 - Persona Loader (Deterministic): Loads five benchmark personas covering distinct Indian demographic tiers, income levels, and psychographic decision manifests.

Stage 3 - Cognitive Simulator (T=0.6, 5x calls): Injects full persona profiles as first-person behavioral context. Simulates inner monologue, emotional gut reaction, quantitative scores (Resonance 1-10, Purchase Intent 0-100%, Credibility Trust 1-10), and friction points.

Stage 4 - Panel Evaluator (Deterministic): Aggregates individual scores into panel averages, cross-persona variance, and polarising attributes.

Stage 5 - Synthesizer (T=0.5): Compiles a strategic 350-word marketing intelligence narrative.

Stage 6 - Population Analyzer (T=0.7): Extrapolates panel findings over N = 1,000 synthetic consumers, projecting overall sentiment distribution, conversion rates, and top objection frequencies.

Stage 7 - Review Generator (T=0.8): Generates 10-50 sample customer reviews conditioned on the Stage 1 price-to-value assessment.

Agent 2 Pipeline Flow:

```
Marketing Asset (Ad / Copy) --> Asset Analyzer (T=0.3) --> Persona Loader
      |
      v
Cognitive Simulator (T=0.6, 5x Personas) --> Panel Evaluator (Stats)
      |
      v
Strategic Synthesizer --> Population Simulation (N=1,000) --> Customer Reviews
```

6.3 Benchmark Persona Profiles

Table 2 outlines the five benchmark personas grounded in actual Indian urban demographic profiles, representing Gen-Z students, tech leaders, traditional business owners, and creative professionals.

Persona Name	Age	City & Tier	Income Bracket	Archetype / Role
Saurabh Sharma	21	Bangalore (Tier 1)	Student (15k/mo)	Tech-Savvy Gen Z Student / Intern
Sarthak Bankar	21	Nagpur (Tier 2)	Student (15k/mo)	Budget-Conscious Tech Student / Value See
Shree Larokar	21	Pune (Tier 1)	High (350k/mo)	High-Income Busy Tech Leader / Director
Aditya Tiwari	22	Jaipur (Tier 2)	MSME (140k/mo)	Traditional MSME / Retail Business Owner
Pooja Iyer	28	Mumbai (Tier 1)	Freelance (95k/mo)	Conscious Urban Freelance Marketer

Table 2: Benchmark Consumer Persona Demographics in Agent 2.

7. EXPERIMENTAL METHODOLOGY

7.1 Agent 1 Evaluation Protocol

Agent 1 was evaluated against an authentic reference dataset of 9,240 lead conversion records across four dimensions: (1) *structural validity* (null rates, duplicate records); (2) *semantic field coverage* (domain-relevance of generated attributes); (3) *correlation preservation* (Pearson r matrix fidelity between related fields); and (4) *downstream machine learning utility*, where classification models (Logistic Regression and Random Forest) were trained on synthetic data and evaluated on authentic test sets.

7.2 Agent 2 Evaluation Protocol

Agent 2 was evaluated across three contrasting marketing assets chosen to stress-test value and credibility perception: (a) **AI Smart Lunchbox** (INR 4,999 - Fair Market Price lifestyle gadget); (b) **TrueSound X1 Earbuds** (INR 1,299 - Exceptional Deal consumer electronics); and (c) **SolarFlow PowerBank** (INR 10 - Exceptional Deal* credibility stress-test with an implausible price). Each asset underwent five independent pipeline executions to measure score stability (mean standard deviation), a macro population simulation (N= 1,000 synthetic consumers), and an ablation comparison against an unconditioned generic consumer prompt.

8. EMPIRICAL RESULTS

8.1 Agent 1 Prototype Results

Table 3 summarizes the structural and statistical results for datasets generated by Agent 1.

Dataset Domain	Target Rows	Fields	Missing Values	Duplicate Rows	Synthesis Time	Validation
Lead Scoring (Gemini)	9,240	12	0.0% (0)	0.0% (0)	12.4s PA	SSSED (100%)
Restaurant Analytics	1,000	15	0.0% (0)	0.0% (0)	1.49s PA	SSSED (100%)
Software SaaS Company	500	14	0.0% (0)	0.0% (0)	0.78s PA	SSSED (100%)

Table 3: Agent 1 Synthetic Dataset Generation Performance and Quality Validation.

In the lead scoring experiment, the relationship engine successfully propagated dependency: the correlation between lead score and conversion in the synthetic dataset was $r = 0.51$ (vs. $r = 0.49$ in the authentic corpus). When training a downstream Random Forest classifier on synthetic data, the model achieved an ROC-AUC of 0.79 on real test data, confirming high machine learning utility.

8.2 Agent 2 Persona Simulation Results

Table 4 reports Run 1 panel metrics, multi-run stability (N=5 runs), and N=1,000 population conversion.

Asset Name	Price PointAv	g. Resonance (1-1	A0)vg. Intent (%)	Avg. CredibilityS	tability (Mean SPT	do)p. Conversion
AI Smart Lunchbox	INR 4,999	5.6 / 10	40.0%	4.6 / 10	SD = 1.02	22.0%
TrueSound X1 Earbuds	INR 1,299	7.2 / 10	63.0%	5.2 / 10	SD = 0.40	21.0%
SolarFlow PowerBank*	INR 10	3.4 / 10	16.0%	1.2 / 10	SD = 0.80	4.5%

Table 4: Agent 2 Persona Panel Simulation, Stability, and Population Results.

Key Findings:

- Directional Coherence:** Earbuds achieved high resonance (7.2) and intent (63%), driven by strong Gen-Z appeal (9/10). Lunchbox revealed sharp polarisation between ROI-focused Aditya (resonance 2) and lifestyle-driven Shree (resonance 8).
- Credibility Collapse Detection:** The INR 10 PowerBank triggered an immediate credibility collapse (credibility trust 1.2/10), producing 80% skeptical/negative sentiment and only 4.5% conversion in population simulation with 85% citing 'Too good to be true'.
- Ablation Superiority:** The unconditioned generic consumer baseline gave the PowerBank a resonance of 5.0 and 30% intent, failing completely to flag the credibility collapse identified by Agent 2.

9. LITERATURE REVIEW

1. M. Nadas et al., 2025

Synthetic Data Generation Using Large Language Models: Advances in Text and Code

Surveys prompt-based tabular synthesis, retrieval-augmented generation pipelines, and self-refinement strategies. Directly informs Agent 1's multi-stage schema and distribution planning architecture.

2. Various Authors, 2024

On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey

Proposes standard seed-generate-filter-score workflows for tabular and unstructured data generation. Informs the structural validation and statistical reporting stages in Agent 1.

3. M. M. H. Nahid and S. B. Hasan, 2024

SafeSynthDP: Leveraging LLMs for Privacy-Preserving Synthetic Data Generation

Analyzes privacy preservation in LLM-generated data through differential privacy constraints. Demonstrates that description-based synthesis inherently minimizes identity re-identification risks.

4. L. J. Argyle et al., 2023

Out of One, Many: Using Language Models to Simulate Human Samples (Political Analysis)

Seminal work establishing that conditioning LLMs with demographic and ideological backstories reproduces statistically coherent human sample distributions, validating Agent 2's persona injection methodology.

5. J. S. Park et al., 2023

Generative Agents: Interactive Simulacra of Human Behavior (ACM UIST)

Demonstrates that believable agent simulations require persistent memory, first-person reflection, and structured psychographic context, directly influencing Agent 2's Cognitive Simulator design.

6. Google DeepMind, 2023

Gemini: A Family of Highly Capable Multimodal Models (arXiv:2312.11805)

Details cross-modal reasoning and high parameter efficiency in the Gemini architecture. Establishes the viability of using flash-tier models for high-throughput, low-latency agent pipelines.

7. S. Agrawal and P. Mittal, 2022

Consumer Behavior in Indian E-commerce: Price Sensitivity and Trust Dynamics (JRCS)

Empirically documents price sensitivity, value perception, and trust verification mechanisms among Indian e-commerce consumers, providing empirical grounding for Agent 2's benchmark persona parameters.

10. LIMITATIONS

1. **Agent 1 Prior Calibration:** When prompting Agent 1 without explicit parameter guidelines, uncalibrated default parameter assumptions can cause distribution skewness relative to real-world datasets.
2. **Demographic Sample Breadth:** The five benchmark personas in Agent 2 represent urban, educated student and professional archetypes, under-representing rural, elderly, and lower-literacy consumer demographics.
3. **Token Limits in Review Synthesis:** Generating 50 detailed customer reviews with rich structured JSON occasionally exceeds token output windows, necessitating deterministic assembly fallback.
4. **Text-Only Asset Input:** The current implementation processes textual ad copy, product pitches, and taglines; multimodal image and video processing remains an ongoing architectural development.

11. FUTURE WORK

1. **Automated Census Grounding:** Dynamically generate and weight thousands of regional personas derived from Indian census and consumer expenditure surveys to enable hyper-local market simulations.
2. **Multimodal Creative Evaluation:** Extend Agent 2 to accept visual creatives, packaging designs, and video storyboards using Gemini Vision capabilities.
3. **Closed-Loop Dual-Agent Optimization:** Couple both agents into a self-optimizing feedback loop where Agent 2 iteratively evaluates marketing copy generated by Agent 1, optimizing conversion before campaign launch.
4. **Real-World Survey Calibration:** Conduct live consumer panel validation comparing simulated persona scores against actual consumer survey responses to establish rigorous empirical confidence bounds.

12. CONCLUSION

This paper has presented the architecture, working prototype, and empirical evaluation of an adaptive **Two-Agent AI Marketing Intelligence System** combining Synthetic Tabular Data Generation (Agent 1) and Consumer Persona Simulation (Agent 2). Built upon the Gemini API backend and orchestrated via a responsive single-page web interface, the system eliminates traditional trade-offs between cost, speed, and data privacy in product marketing and machine learning workflows.

Empirical validation confirms that Agent 1 successfully synthesizes structured tabular datasets with complex relational dependencies and robust downstream classification utility (ROC-AUC = 0.79), while Agent 2 produces demographically coherent consumer intelligence, reliable multi-run stability, and high-fidelity objection analysis. By uniting zero-shot data generation with pre-launch market simulation, the system provides a scalable, commercially viable, and academically grounded foundation for next-generation AI-driven marketing intelligence.

REFERENCES

1. M. Nadas et al., 'Synthetic Data Generation Using Large Language Models: Advances in Text and Code,' arXiv preprint, 2025.
2. Various Authors, 'On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey,' arXiv preprint, 2024.
3. M. M. H. Nahid and S. B. Hasan, 'SafeSynthDP: Leveraging LLMs for Privacy-Preserving Synthetic Data Generation Using Differential Privacy,' arXiv preprint, 2024.
4. L. J. Argyle et al., 'Out of One, Many: Using Language Models to Simulate Human Samples,' Political Analysis, vol. 31, no. 3, pp. 337-351, 2023.
5. J. S. Park et al., 'Generative Agents: Interactive Simulacra of Human Behavior,' in Proc. ACM UIST, San Francisco, CA, 2023, pp. 1-22.
6. Google DeepMind, 'Gemini: A Family of Highly Capable Multimodal Models,' arXiv:2312.11805, Dec. 2023.
7. S. Agrawal and P. Mittal, 'Consumer Behavior in Indian E-commerce: Price Sensitivity and Trust Dynamics,' Journal of Retailing and Consumer Services, vol. 67, 2022.

This paper was prepared as a project review and technical report for the final year project in Artificial Intelligence and Data Science, Priyadarshini College of Engineering, Nagpur - 2025-2026. Under the mentorship of Prof. U.A.S. Gani. All reported experimental metrics are actual outputs from the operational Agent 1 and Agent 2 prototypes and have not been fabricated.