Building a Persona-Driven Survey Engine Using AI for Synthetic User Modeling and Decision Simulation

Debabrata Pruseth
AI Architect & Applied AI Researcher
Singapore

Author Note

This article is a research-style companion version of the blog post **"Building a Persona-Driven Survey Engine Using AI"** and the associated GitHub implementation Persona-driven-AI-survey-engine. The blog describes a practical AI system for testing early-stage product ideas using synthetic personas, AI-generated surveys, simulated responses, and insight extraction. The GitHub repository provides the notebook-based implementation, pipeline design, architecture, and artifact structure for reproducing the experiment.

The purpose of this paper is to formalize the project as an applied AI framework for **synthetic user modeling**, **early-stage product discovery**, and **decision simulation**. The system is demonstrated through a hypothetical financial product case study, **FlexiSave Kids Plan**, a hybrid savings and micro-investment product for parents of children aged 4–10. The paper emphasizes that synthetic personas should not be treated as a replacement for real user research. Instead, they should be used as a transparent and auditable method for hypothesis generation, research preparation, segment exploration, and early product decision support.

Abstract

Early-stage product teams often require user insight before real customer data, survey panels, or interview participants are available. This creates a decision gap between product ideation and validated user understanding. Traditional research methods such as interviews, surveys, focus groups, and market studies remain essential, but they may be slow, expensive, or unavailable during concept exploration. This paper presents a persona-driven AI survey engine for synthetic user modeling and decision simulation. The proposed system combines large-scale synthetic persona data, heuristic audience filtering, behavioral segmentation, large language model-assisted persona construction, persona-aware survey generation, simulated response generation, and structured insight extraction. The framework is demonstrated through the **FlexiSave Kids Plan**, a hypothetical financial product for parents of children aged 4–10. Implementation artifacts are provided through a public GitHub repository, including a notebook workflow and JSON artifact structure. The results show how synthetic personas can surface segment-specific needs, objections, trust concerns, and feature priorities before real-world validation. The paper argues that synthetic personas should be used as decision accelerators, not empirical substitutes for real users.

Keywords: Synthetic Personas, AI Survey Engine, Persona-Driven AI, Synthetic User Modeling, Decision Simulation, Large Language Models, Product Discovery, Market Research Automation, User Research, AI Personas, Survey Simulation, Generative AI, Synthetic Data, Product Validation, Human-in-the-Loop Research

---

1. Introduction

Product development frequently begins under uncertainty. Teams must make decisions about target users, value propositions, features, pricing, trust signals, product messaging, and go-to-market strategy before sufficient real user data exists. This is especially challenging in early-stage product development, where teams may not yet have customers, research participants, usage analytics, or validated market signals.

Traditional user research methods remain the gold standard for understanding real people. Interviews, surveys, focus groups, ethnographic studies, and usability tests provide direct insight into human behavior. However, these methods may be difficult to execute at the earliest stage of product exploration. Recruiting participants can take time, survey audiences may not yet exist, and market research can be expensive. As a result, early product decisions are often shaped by assumptions, expert judgment, competitor observation, or limited anecdotal feedback.

This paper proposes a persona-driven AI survey engine as a structured method for exploring early product hypotheses before real validation. The system uses synthetic personas and large language models to simulate a small focus group. It filters a large persona dataset, segments likely target users into decision-relevant behavioral categories, transforms raw records into structured personas, generates survey questions, simulates responses, and extracts product insights.

The system is not intended to replace real user research. Instead, it provides a structured pre-research mechanism. It helps teams ask better questions before interviews, identify likely objections before launch, compare segment-level reactions before recruitment, and refine product positioning before investing in full validation.

The implementation described in the blog and GitHub repository follows an end-to-end pipeline: **raw dataset → filtered personas → segmentation → survey generation → simulated responses → insights**. The GitHub repository describes this pipeline as a synthetic focus group built using large-scale persona data and LLMs to simulate user feedback before real user testing.

---

2. Research Problem and Objectives

2.1 Research Problem

The central research problem is:

How can AI-generated or AI-structured personas be used to simulate early-stage survey feedback in a way that supports product discovery while preserving transparency about limitations and the need for real user validation?

This problem is important because early product decisions often require user insight before sufficient evidence exists. A persona-driven simulation system can support decision-making, but it must be designed carefully to avoid false validation. Synthetic responses are not equivalent to real user responses. Therefore, the system must be framed as a hypothesis-generation and decision-simulation tool rather than a replacement for empirical research.

2.2 Research Objectives

The objectives of this paper are to:

1. Formalize a persona-driven AI survey engine as an applied AI system.
2. Describe an end-to-end pipeline for filtering, segmenting, and structuring synthetic personas.
3. Demonstrate how large language models can generate persona-aware surveys and simulated responses.
4. Show how synthetic responses can be converted into product insights.
5. Evaluate the usefulness and limitations of the approach for early-stage product discovery.
6. Define governance principles for responsible use of synthetic personas.
7. Provide a reproducible implementation reference through the associated GitHub repository.

---

3. Related Work and Conceptual Background

3.1 Persona-Based Design

Personas are widely used in product design, marketing, user experience research, and service design. They help teams reason about different types of users by representing goals, behaviors, frustrations, preferences, and decision patterns. Traditional personas are usually built from real user research. However, in early-stage product exploration, real user data may not yet be available.

Synthetic personas extend the concept by using generated or simulated profiles. Their value lies in structured imagination: they allow teams to explore how different types of users might respond to a product concept. However, synthetic personas must be used carefully because they can encode assumptions, stereotypes, or model-generated artifacts.

3.2 Synthetic Personas and Large Language Models

Large language models make it possible to generate structured persona profiles, simulate user responses, and summarize patterns across synthetic participants. The GitHub implementation uses LLMs for persona generation, survey generation, response simulation, and insight extraction.

The key design choice is that the LLM is not used to invent the entire user population from nothing. Instead, it is used to transform filtered and segmented persona records into structured research artifacts. This is important because grounding the generation in dataset rows reduces unconstrained hallucination and improves traceability.

3.3 Synthetic User Research as Decision Simulation

The system should be understood as **decision simulation**, not market validation. It asks:

How might different types of users think, decide, object, and respond?

This is different from asking:

What will real users actually do?

The first question can be supported through structured synthetic personas. The second requires real user research, behavioral data, and market validation.

---

4. Proposed Framework

The proposed persona-driven survey engine consists of six core modules:

1. Persona dataset ingestion
2. Target audience filtering
3. Behavioral segmentation
4. Persona panel construction
5. LLM-based survey and response simulation
6. Insight extraction and decision synthesis

The GitHub repository describes the architecture as a pipeline moving from a persona dataset through filtering, scoring, segmentation, LLM-based persona building, survey generation, response simulation, and insight extraction.

4.1 Framework Architecture

Layer	Function	Output
Data Ingestion	Load large-scale synthetic persona dataset	Raw persona records
Filtering and Scoring	Identify target audience candidates	Filtered persona subset
Segmentation	Classify personas into decision-relevant groups	Behavioral segments

Layer	Function	Output
Persona Construction	Convert raw rows into structured profiles	Synthetic focus group
Survey Simulation	Generate questions and simulate responses	Persona-level responses
Insight Extraction	Aggregate patterns, objections, and opportunities	Product insights

4.2 System Logic

The system begins with a broad synthetic population and progressively narrows it into a structured decision panel. This design is important because raw persona datasets are too large and unstructured for direct product testing. Filtering identifies likely target users. Segmentation converts demographic and text signals into behavioral categories. Persona construction creates interpretable profiles. Survey simulation generates segment-specific responses. Insight extraction converts those responses into product decisions.

---

5. Dataset and Target Audience

5.1 Dataset

The implementation uses the **Nemotron Personas USA** dataset, described in the blog as containing more than one million synthetic personas based on U.S. demographic data. The GitHub repository describes the dataset as containing approximately one million synthetic personas with fields such as demographics, location, marital status, occupation, hobbies, skills, and free-text persona descriptions.

5.2 Target Audience

The example case study targets:

Parents of children aged 4–10 in the United States.

This target group was selected because the product concept, **FlexiSave Kids Plan**, is designed as a hybrid savings and micro-investment tool for parents of young children. The blog describes the product as a financial concept for parents of children aged 4–10.

5.3 Weak-Labeling Challenge

The dataset does not necessarily provide a direct label such as “parent of a child aged 4–10.” Therefore, the system uses proxy signals to infer target relevance. These signals include age, marital status, family-related keywords, lifestyle indicators, and persona text. The GitHub

repository describes the filtering approach as heuristic scoring using keywords, age band, and marital status.

This makes the filtering process a weak-labeling problem rather than a deterministic classification task. The goal is not to prove that each selected persona is a real parent, but to identify high-probability persona records suitable for early simulation.

---

6. Methodology

6.1 Data Preparation

The pipeline begins by loading a subset of the persona dataset. The GitHub repository describes a workflow in which a subset of 10,000 to 50,000 rows can be loaded, null values cleaned, and a combined text representation created for natural-language filtering.

The combined text field is important because many target signals are embedded in free-text persona descriptions rather than structured columns. A person's occupation, interests, family context, hobbies, or lifestyle indicators may provide clues about their suitability for the product test.

6.2 Filtering and Scoring

The filtering module assigns a **parent relevance score** based on demographic and textual indicators. The blog describes this step as using keyword detection, contextual signals, and demographic hints to retain high-confidence target personas.

The scoring logic can be represented as:

Signal Type	Example Indicators	Purpose
Demographic	Age band 25–50	Identify likely parent age range
Marital / household	Married, divorced, family-related context	Infer family relevance
Text keywords	kids, child, school, family, parenting	Detect parental context
Occupation	teacher, caretaker, family-related roles	Identify child-related orientation
Lifestyle	education, community, family activities	Infer value alignment

This stage produces a filtered set of high-probability target personas.

6.3 Behavioral Segmentation

After filtering, the system classifies personas into financial behavior segments. The GitHub repository identifies the following segment categories: security-first, growth-oriented, busy, budget-conscious, values-driven, and skeptical.

Segmentation is necessary because product adoption depends not only on demographic characteristics but also on behavioral orientation. For a financial product aimed at parents, relevant decision variables include risk appetite, convenience preference, budget sensitivity, trust level, and education orientation.

The segmentation categories used in the experiment are:

Segment	Behavioral Interpretation
Security-First Parent	Prioritizes safety, trust, and capital protection
Growth-Oriented Parent	Interested in investment growth and long-term returns
Busy Convenience Parent	Values simplicity, automation, and time savings
Budget-Conscious Parent	Sensitive to cost, fees, and affordability
Values-Driven Parent	Prioritizes education, responsibility, and child development
Skeptical Parent	Low initial trust; concerned about risk, complexity, or claims
General Parent	Relevant but without a strong dominant behavioral signal

6.4 Persona Panel Selection

The pipeline then selects a smaller set of representative personas to form a synthetic focus group. The GitHub repository describes this step as deduplicating personas, ranking by relevance score, and selecting 10–12 representative samples.

A small panel is useful because it supports interpretability. Large-scale synthetic responses may produce volume, but product teams often need a manageable set of differentiated personas that can be examined qualitatively.

6.5 LLM-Based Persona Construction

The LLM transforms selected raw records into structured personas. The GitHub repository describes a JSON structure containing name, segment, financial mindset, pain points, purchase triggers, and objections.

The persona generation step should follow three principles:

1. **Grounding:** Use available data from the raw persona record.
2. **Traceability:** Preserve segment and score information.
3. **Constraint:** Avoid inventing unsupported claims.

The output is a structured persona suitable for survey simulation.

6.6 Survey Generation

The LLM then generates persona-aware survey questions. The GitHub repository describes survey sections such as awareness, trust, pricing, usability, and decision triggers.

The survey is designed to test:

- initial product appeal
- feature prioritization
- trust concerns
- risk tolerance
- willingness to use automated savings
- perceived value of educational features
- likelihood of recommendation
- objections and rejection reasons

6.7 Response Simulation

Each persona responds to the survey using its profile and segment orientation. The GitHub repository describes this step as having each persona answer all survey questions, provide reasoning, and reflect segment-specific behavior.

The purpose is to simulate differentiated user reasoning rather than generic market feedback.

6.8 Insight Extraction

Finally, the LLM aggregates simulated responses to identify recurring needs, objections, segment preferences, rejection drivers, and product opportunities. The GitHub repository describes output artifacts including `clean_financial_parent_personas.json`, `financial_product_survey.json`, `simulated_survey_responses.json`, and `insights.json`.

---

7. Implementation

7.1 Technical Environment

The implementation uses a notebook-based workflow. The GitHub repository identifies the tech stack as Python in Colab/Jupyter, Hugging Face Datasets for data ingestion, Pandas for processing, optional scikit-learn for clustering, optional Sentence Transformers for embeddings, the OpenAI API for persona generation and simulation, and JSON pipelines for intermediate artifacts.

Component	Role
Python / Colab / Jupyter	Experiment execution
Hugging Face Datasets	Dataset ingestion
Pandas	Data cleaning, filtering, and transformation
scikit-learn	Optional clustering
Sentence Transformers	Optional embeddings and similarity analysis
OpenAI API	Persona construction, survey generation, response simulation, insight extraction
JSON artifacts	Reproducible intermediate and final outputs

7.2 Output Artifacts

Artifact	Purpose
`clean_financial_parent_personas.json`	Structured persona panel
`financial_product_survey.json`	Generated master survey
`simulated_survey_responses.json`	Persona-level simulated responses
`insights.json`	Aggregated needs, pain points, objections, and recommendations

These artifacts make the workflow more reproducible and auditable than a purely conversational LLM interaction.

---

8. Case Study: FlexiSave Kids Plan

8.1 Product Concept

The system is demonstrated using **FlexiSave Kids Plan**, a hypothetical financial product combining goal-based savings, micro-investments, and financial learning for children. The GitHub repository describes the tested product as including goal-based savings, micro-investments, and financial learning for kids.

8.2 Case Study Objective

The case study aims to explore whether a persona-driven AI survey engine can generate useful early insight into:

- which parent segments may trust the product;
- which objections may arise;
- which features may matter most;
- how product messaging could be improved;
- whether the idea should be refined before real user testing.

8.3 Research Questions

RQ1: Which synthetic parent segments show the strongest interest in the product concept?
RQ2: What objections emerge across different behavioral segments?
RQ3: Which product features are most consistently valued?
RQ4: How can simulated responses improve product positioning before real user research?
RQ5: What governance controls are required to prevent misuse of synthetic persona insights?

---

9. Results

The results below are drawn from the code-output examples reported in the blog. These outputs should be interpreted as experimental artifacts from a synthetic simulation pipeline, not empirical findings about real parents.

9.1 Filtering Output

The filtering stage produced high-scoring candidate personas based on parent-relevance signals. The top candidates included occupations such as chemical engineer, preschool or kindergarten teacher, construction inspector, elementary or middle school teacher, customer service representative, manager, and housekeeper. The reported parent scores ranged from 14 to 16 among the top preview examples.

Example Candidate Attribute	Observed Output Pattern
Age range	25–46 among top preview examples
Marital status	Divorced, never married, married present
Occupations	Teachers, engineers, managers, service workers
Parent relevance score	14–16 in top candidate preview

This output suggests that the heuristic scoring method was able to identify a diverse set of candidate personas rather than a single narrow demographic profile.

9.2 Segment Distribution

The behavioral segmentation step produced the following distribution:

Segment	Count
Values-Driven Parent	2,119
Busy Convenience Parent	1,894
Security-First Parent	964
Growth-Oriented Parent	265
Budget-Conscious Parent	190
General Parent	68
Skeptical Parent	53

This distribution indicates that values-driven and convenience-oriented personas dominated the filtered persona pool. For the FlexiSave Kids Plan case, this suggests that product positioning may need to emphasize child development, educational value, simplicity, and automated savings before emphasizing investment sophistication.

9.3 Persona Panel Output

The final panel contained 12 personas selected across segments. The reported panel included busy convenience, values-driven, growth-oriented, general, security-first, skeptical, and budget-conscious parent segments.

Segment	Age	Occupation	State	Parent Score
Busy Convenience Parent	31	Chemical engineer	FL	16
Values-Driven Parent	25	Preschool/kindergarten teacher	NY	16
Growth-Oriented Parent	25	Customer service representative	DC	15
Values-Driven Parent	33	Elementary/middle school teacher	NY	15
Busy Convenience Parent	46	Not in workforce	FL	14

Segment	Age	Occupation	State	Parent Score
Growth-Oriented Parent	42	No occupation	TX	14
General Parent	48	Cabinetmaker / bench carpenter	WV	14
Security-First Parent	37	Painting worker	MA	14
Security-First Parent	48	Heavy vehicle / mobile equipment technician	NY	14
Skeptical Parent	38	Animal caretaker	TX	14
Budget-Conscious Parent	25	Cost estimator	MO	13
Budget-Conscious Parent	30	Construction laborer	CA	13

The panel design provides coverage across behavioral segments while keeping the synthetic focus group small enough for qualitative interpretation.

9.4 Structured Persona Example

The LLM-based persona construction step transformed raw persona records into structured personas. One reported example was **Emma — Ambitious Parent**, classified as a Busy Convenience Parent. The generated persona described Emma as a 31-year-old chemical engineer balancing a demanding career with interests in art and community. Her financial mindset was cautious but open to innovative financial solutions aligned with sustainability and education. Her quote emphasized the need for a financial solution that fits into a busy life.

Interpretation:
This persona is useful for testing whether automated contributions, clear dashboards, and educational features appeal to busy professional parents who are open to innovation but require trust and convenience.

A second reported example was **Vivian — Values-Driven Parent Educator**, a 25-year-old early-childhood educator who values curiosity, hands-on learning, responsible financial decisions, and educational experiences.

Interpretation:
This persona is useful for testing whether educational features and child money-learning nudges strengthen product appeal.

9.5 Survey Generation Output

The survey-generation step produced five sections and 15 questions. Reported sections included overall interest and appeal, savings and investment features, and educational and engagement features.

Survey Dimension	Example Question	Research Purpose
Overall Interest	How interested are you in using a financial product like FlexiSave Kids Plan for your child?	Measures initial adoption intent
Feature Appeal	Which feature appeals most?	Identifies product priorities
Investment Comfort	How important is low-risk micro-investing?	Tests investment acceptance
Automation	Are monthly automatic contributions helpful?	Tests convenience value
Education	How valuable are money-learning nudges for children?	Tests educational positioning

The survey design is directly linked to product decision questions rather than generic satisfaction measurement.

9.6 Simulated Response Output

The reported simulated response for Emma showed that she was **somewhat interested** in the product. Her reasoning emphasized structured savings and early financial learning, while also expressing caution about new financial products and the need to fit her busy schedule. She identified goal-based savings for children's future needs as the most appealing feature.

This response is internally consistent with the persona's segment classification. As a Busy Convenience Parent, Emma values structure and convenience but remains cautious about adoption friction and product trust.

9.7 Insight Extraction Output

The final insight extraction stage identified needs, pain points, objections, interested segments, less interested segments, rejection reasons, key features, product improvements, and positioning.

Insight Category	Extracted Findings
Needs	Goal-based savings, monthly automatic contributions, customization, educational features, fee and risk transparency

Insight Category	Extracted Findings
Pain Points	Hidden fees, investment complexity, time constraints, uncertainty about educational nudges, difficulty maintaining contributions
Objections	Safety of micro-investing, caution about new financial products, need for proof of benefits, complexity, hidden costs
Key Features	Goal-based savings, automatic contributions, parent dashboard, money-learning nudges, low-risk micro-investing
Product Improvements	Improve transparency, simplify investment terminology, strengthen educational content
Positioning	Structured, transparent, education-oriented savings product for parents

This output demonstrates the core utility of the system: it transforms simulated persona responses into product design and positioning hypotheses.

---

10. Evaluation

10.1 Evaluation Criteria

Because the system uses synthetic personas, evaluation should not be based on whether the outputs represent real users. Instead, the appropriate evaluation criteria are internal consistency, reproducibility, usefulness for hypothesis generation, and readiness for real validation.

Criterion	Evaluation Question
Persona relevance	Do selected personas plausibly match the target audience?
Segment diversity	Does the panel cover multiple behavioral orientations?
Response consistency	Do responses align with persona profiles and segments?
Insight usefulness	Do outputs identify actionable needs, objections, and positioning ideas?
Reproducibility	Are code, pipeline, and artifacts documented?
Governance	Are limitations and validation requirements disclosed?

10.2 Qualitative Evaluation

The case study suggests that the system performs well as a pre-validation tool. It generates differentiated personas, segment-aware survey questions, and structured insights. However, its outputs remain synthetic and should not be treated as evidence of market demand.

10.3 Reproducibility Evaluation

The GitHub repository strengthens reproducibility because it contains the notebook, architecture, pipeline description, tech stack, dataset description, and artifact list.

---

11. Discussion

The persona-driven survey engine demonstrates how LLMs can support early-stage product discovery when used within a structured pipeline. The value does not come from asking a model for generic product feedback. It comes from combining data-grounded personas, segment classification, persona-aware survey generation, simulated responses, and structured insight extraction.

The most important contribution is the conversion of persona data into **decision intelligence**. A raw persona record is difficult to use directly. A segmented, structured persona can support targeted product questions. A panel of such personas can reveal likely variation across user mindsets.

For the FlexiSave Kids Plan case, the strongest simulated themes were trust, simplicity, goal-based savings, child education, automatic contributions, and fee transparency. These insights are plausible for an early financial product targeted at parents. However, plausibility is not proof. The outputs should be used to design better real-world surveys and interviews.

This distinction is central. The system helps answer:

What should we test with real users?

It does not answer:

What do real users definitely want?

---

12. Governance and Ethical Considerations

12.1 Transparency

Every output should clearly state that the responses are synthetic and AI-generated. This prevents decision-makers from confusing simulation with evidence.

12.2 Human Review

Synthetic insights should be reviewed by product, research, design, compliance, and domain experts before influencing product decisions.

12.3 Bias and Representation

Bias can enter through the dataset, filtering rules, segmentation logic, prompt wording, and model outputs. Teams should examine whether certain demographics, occupations, regions, or family structures are overrepresented or stereotyped.

12.4 Auditability

A responsible system should preserve:

- dataset source
- filtering rules
- scoring logic
- segment definitions
- prompts
- model version
- generated personas
- generated survey questions
- simulated responses
- extracted insights

The JSON artifact structure described in the GitHub repository supports this auditability.

12.5 Real User Validation

Synthetic simulation should always be followed by real validation. This is especially important for financial products, where trust, suitability, risk disclosure, regulatory review, and consumer protection are critical.

---

13. Limitations

This study has several limitations.

First, no real users participated in the simulation. Therefore, the results cannot be interpreted as empirical evidence of user demand.

Second, the parent target label was inferred using heuristic scoring rather than confirmed ground truth. This means some selected personas may not accurately represent real parents of children aged 4–10.

Third, behavioral segments were created using rule-based classification. While interpretable, such classification may reflect designer assumptions.

Fourth, LLM-generated personas and responses may contain bias, stereotypes, overgeneralization, or prompt-induced artifacts.

Fifth, the simulated responses may be internally coherent but behaviorally inaccurate. Real users may behave differently under actual financial, emotional, social, or regulatory constraints.

Sixth, the financial product concept would require compliance, risk, legal, and consumer testing before any production use.

Finally, the study evaluates usefulness for early product discovery, not predictive accuracy against real survey data.

---

14. Reproducibility

The implementation code is available in the GitHub repository **Persona-driven-AI-survey-engine**. The repository includes the Jupyter Notebook, README, architecture description, pipeline overview, dataset description, technology stack, output artifact names, and extension ideas.

To reproduce or extend the experiment, a reader should:

1. Load the synthetic persona dataset.
2. Clean and prepare demographic and text fields.
3. Define a target audience.
4. Apply scoring and filtering rules.
5. Classify personas into behavioral segments.
6. Select a balanced persona panel.
7. Generate structured personas using an LLM.
8. Generate persona-aware survey questions.
9. Simulate responses.
10. Extract insights.
11. Compare outputs against real user research when available.

---

15. Future Work

Future work can strengthen the framework in several ways.

First, synthetic responses can be compared against real survey responses to evaluate alignment and divergence.

Second, multiple LLMs can be tested to examine response consistency, bias, and robustness.

Third, rule-based segmentation can be replaced or complemented with clustering, embedding similarity, or hybrid human-in-the-loop classification. The GitHub repository also identifies

clustering, embedding-based persona similarity, Streamlit UI, real survey integration, and multi-product A/B simulation as possible extensions.

Fourth, the system can support multi-agent discussion, where personas debate a product concept before answering individually.

Fifth, governance features can be added, including bias reporting, confidence labels, uncertainty markers, prompt logs, and audit dashboards.

Sixth, real survey feedback can be integrated to create a hybrid system that compares synthetic predictions with empirical results.

---

16. Conclusion

This paper presented a persona-driven AI survey engine for synthetic user modeling and decision simulation. The system combines large-scale synthetic persona data, heuristic filtering, behavioral segmentation, LLM-assisted persona construction, persona-aware survey generation, simulated responses, and insight extraction.

The FlexiSave Kids Plan case study demonstrates how the system can generate early insight into segment-level needs, objections, trust barriers, feature priorities, and positioning opportunities. The implementation shows that synthetic personas can support early product discovery when they are structured, auditable, and clearly framed as exploratory tools.

The central contribution of the paper is not the claim that synthetic users can replace real users. They cannot. The contribution is a practical framework for using synthetic personas to improve the quality of early product thinking before real validation begins.

The responsible conclusion is therefore:

Synthetic personas are not evidence of market truth. They are tools for structured imagination, hypothesis generation, and early decision simulation.

References

Pruseth, D. (2026). *Building a Persona-Driven Survey Engine Using AI*. Debabrata Pruseth AI Blog.

Pruseth, D. (2026). *Persona-driven-AI-survey-engine*. GitHub repository.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). *Scaling Synthetic Data Creation with 1,000,000,000 Personas*. arXiv.

Suggested Citation

Pruseth, D. (2026). *Building a Persona-Driven Survey Engine Using AI for Synthetic User Modeling and Decision Simulation*. Debabrata Pruseth AI Blog.