AI System Architecture & Model Usage

Last Updated: March 11, 2025

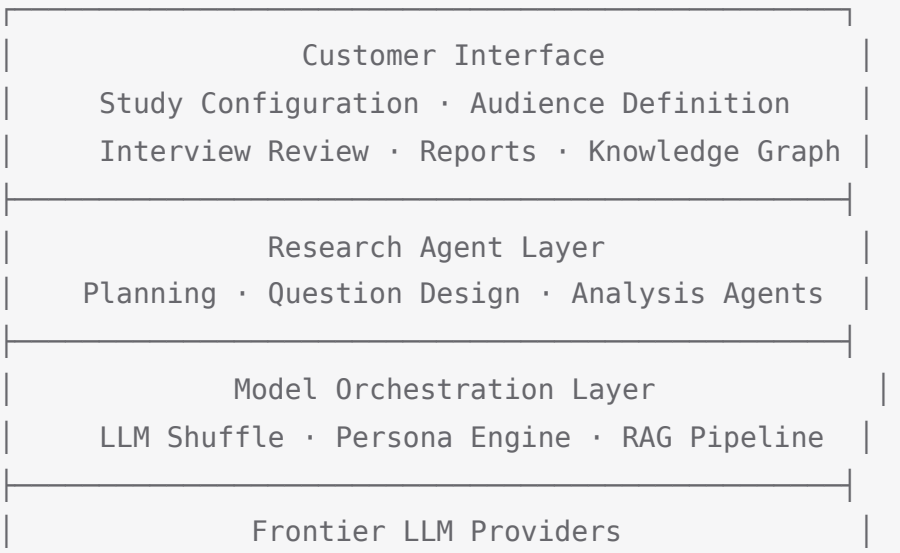
1. Introduction

Synthetic Users is built on a multi-model AI architecture designed to generate realistic, diverse, and bias-resistant synthetic interview participants. This document provides transparency into how our platform orchestrates frontier LLMs, processes customer inputs, and produces research outputs.

This is a high-level architectural overview intended for customers, partners, and compliance reviewers. It does not disclose proprietary implementation details, prompt structures, or model fine-tuning configurations.

2. Architecture Overview

The Synthetic Users platform consists of four primary layers:



```
│    OpenAI · Anthropic · Google · Meta · Mistral  │
└──────────────────────────────────────────────────┘
```

Customer Interface Layer

The web application through which researchers configure studies, define audiences, review interviews, and analyze findings. This layer handles:

- Study creation and configuration (interview type, audience, goals, problems, solutions)
- Document upload for RAG enrichment
- Interview transcript review and follow-up questioning
- Report generation and interactive follow-up
- Knowledge Graph visualization
- PRISMA™ multi-study orchestration

Research Agent Layer

Dedicated AI agents that assist researchers with study design and analysis. These agents operate between the customer and the model orchestration layer:

- **Planning Agent** — Recommends methodologies, sampling strategies, and interview structures based on the research objective.
- **Question Design Agent** — Helps formulate neutral, well-structured interview questions and flags potentially leading or biased prompts.
- **Analysis Agent** — Summarizes findings, identifies themes and patterns, and highlights areas of consensus or divergence across participants.

Research Agents provide guidance and recommendations; final decisions always rest with the human researcher.

Model Orchestration Layer

The core engine that transforms study configurations into synthetic interview conversations. This layer manages:

- **LLM Shuffle** — Distribution of generation tasks across multiple frontier models.
- **Persona Engine** — Construction of psychologically grounded synthetic participants using OCEAN personality architectures and customer-defined audience parameters.
- **RAG Pipeline** — Retrieval and integration of customer-uploaded documents into the generation context.
- **Prompt Construction** — Assembly of structured prompts that combine persona definitions, interview questions, RAG context, and conversation history.
- **Response Processing** — Post-processing of model outputs for quality, consistency, and content safety.

Frontier LLM Provider Layer

The external AI models that generate the underlying text outputs. Synthetic Users currently integrates with:

Provider	Models Used	Role
OpenAI	GPT-4 class models	General-purpose generation, concept testing, analytical reasoning
Anthropic	Claude model family	Nuanced conversation, safety-aligned generation, detailed analysis
Google	Gemini model family	Multilingual generation, broad knowledge, diverse perspectives
Meta	Llama model family	Open-weight model diversity, alternative reasoning patterns
Mistral	Mistral model family	European-perspective generation, efficient high-quality outputs

Model selection and version management are handled by the orchestration layer and are not configurable by end users.

---

3. LLM Shuffle: Multi-Model Generation

Purpose

No single LLM is free from bias. Each model reflects the biases of its training data, alignment procedures, and architectural choices. By distributing generation across multiple models, we reduce the influence of any single model's systematic biases on research outcomes.

How It Works

1. **Study initialization:** When a study is created, the orchestration layer determines the model allocation strategy based on the number of participants, interview type, and available models.
2. **Participant assignment:** Each synthetic participant is assigned to a model (or combination of models) such that the overall study draws from a diverse set of providers.
3. **Generation:** Interview responses are generated by the assigned model, with the full persona definition, conversation history, and any RAG context included in the prompt.
4. **Balancing:** The system ensures that no single model generates a disproportionate share of participants within a study, preventing any one model's characteristics from dominating findings.

Benefits

- **Bias dilution:** Systematic biases from individual models are offset by the different biases of other models.
- **Output diversity:** Different models produce different linguistic styles, reasoning patterns, and perspectives, more closely approximating the natural variability of human responses.
- **Resilience:** If a model experiences degraded performance, rate limiting, or an outage, the platform can redistribute generation to other providers without interrupting the study.
- **No model lock-in:** Customers are not dependent on any single AI provider's capabilities or policies.

4. Persona Engine

OCEAN Personality Architecture

Each synthetic participant is constructed with a personality profile based on the Big Five (OCEAN) personality traits:

- **Openness** — Influences creativity, curiosity, and receptivity to new ideas.
- **Conscientiousness** — Affects organization, reliability, and attention to detail.
- **Extraversion** — Shapes sociability, energy, and communication style.
- **Agreeableness** — Governs cooperation, empathy, and interpersonal warmth.
- **Neuroticism** — Determines emotional sensitivity, stress response, and anxiety levels.

These traits are combined with customer-defined audience parameters (demographics, profession, context, goals, pain points) to produce a complete participant profile that guides the model's generation behavior throughout the interview.

Why OCEAN

- **Empirically validated:** The Big Five model is the most widely accepted and researched personality framework in psychology.
- **Culturally studied:** OCEAN traits have been studied across cultures, providing a foundation for generating culturally aware personas.
- **Behaviorally predictive:** Trait combinations produce consistent behavioral patterns, making synthetic participants more realistic and internally coherent.
- **Bias-resistant:** Explicit personality parameters prevent models from filling gaps with stereotypical defaults.

5. RAG Pipeline (Retrieval-Augmented Generation)

Purpose

Customers can upload their own documents to enrich synthetic interviews with domain-specific context. This allows synthetic participants to reference proprietary information, previous research findings, or specialized knowledge that would not be present in the LLM's training data.

How It Works

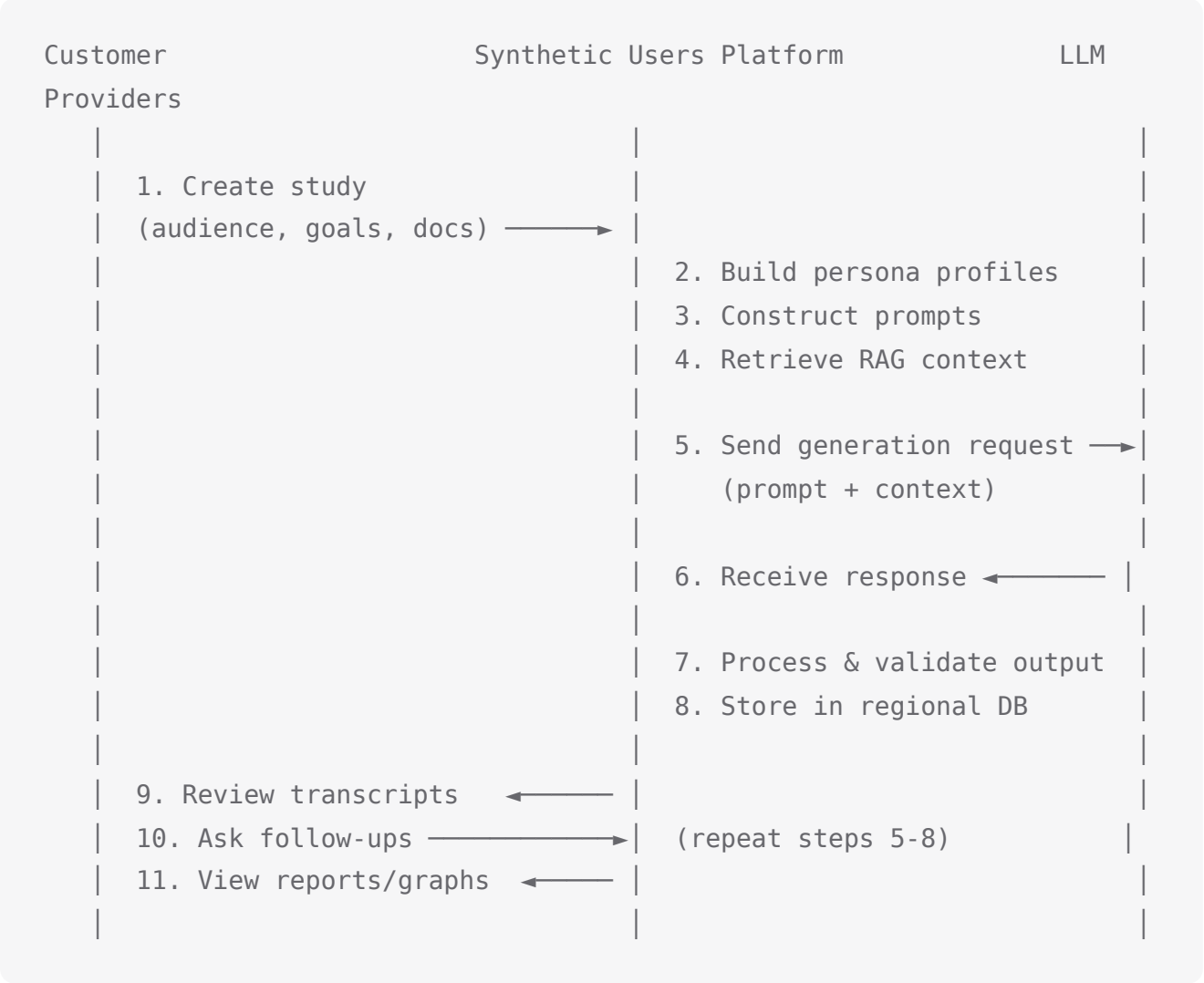
1. **Document upload:** Customers upload PDF documents (up to 100MB or 100 pages per document, up to 3 documents per study).
2. **Ingestion and indexing:** Documents are parsed, chunked, and indexed for efficient retrieval.
3. **Contextual retrieval:** During interview generation, relevant document segments are retrieved based on the current question and conversation context.
4. **Augmented generation:** Retrieved content is included in the model prompt alongside the persona definition and conversation history, grounding the participant's responses in the customer's own data.

Data Protection

- Uploaded documents are stored in the customer's regional data center.
- Documents are encrypted at rest (AES-256) and in transit (TLS 1.2+).
- Document content is transmitted to AI providers only during active generation and is subject to zero-training agreements with all providers.
- Documents are deleted when the customer removes them or when the study is deleted, in accordance with our User Data Deletion and Retention Policy.

6. Data Flow

The following describes the data flow for a typical study:



What Is Sent to LLM Providers

- Constructed prompts containing: persona definitions, interview questions, conversation history, and relevant RAG content.
- No raw customer account data, credentials, or PII beyond what is included in the study configuration.

What Is NOT Sent to LLM Providers

- Customer account information (email, name, billing details).

- Data from other customers or other studies.
- Raw uploaded documents in their entirety (only retrieved relevant segments).

Provider Data Commitments

All LLM providers used by Synthetic Users operate under agreements that include:

- **Zero-training guarantees:** Customer data sent via API is not used to train, fine-tune, or improve the provider's models.
- **No data retention:** Providers do not retain prompt or completion data beyond the duration needed to fulfill the API request (subject to provider-specific retention policies, which are evaluated during vendor assessment).
- **Data Processing Addendums:** Formal agreements governing data handling, security, and breach notification.

For a complete list of subprocessors and data flow details, see Subprocessors and Data Flow.

---

7. Preventing Model Lock-In

Our multi-model architecture is intentionally designed to prevent dependence on any single AI provider:

- **Provider-agnostic orchestration:** The orchestration layer abstracts model-specific APIs behind a unified interface. New models can be added or existing models replaced without changes to the rest of the platform.
- **No provider-specific fine-tuning:** We do not fine-tune provider models, which would create switching costs and dependency on a specific provider's infrastructure.
- **Continuous evaluation:** We regularly evaluate new models and providers against our quality, safety, and bias criteria. Models that no longer meet our standards are phased out.
- **Graceful degradation:** If a provider experiences an outage or degraded performance, generation is automatically redistributed to other available models.

---

8. Preventing Bias Amplification

Multi-model generation reduces bias, but without careful design it could also amplify shared biases. We address this through:

- **Diverse model selection:** We intentionally select models from different providers with different training approaches, data sources, and alignment methodologies.
- **Output monitoring:** We monitor for convergence patterns where multiple models produce the same biased output, which may indicate a shared bias that requires additional mitigation.
- **Persona-driven generation:** Structured persona definitions (OCEAN traits + audience parameters) constrain generation to the defined participant profile, reducing the model's tendency to default to generic or majority-culture responses.
- **Research Agent guidance:** Question Design Agents help researchers avoid prompts that could trigger shared model biases (e.g., leading questions, culturally loaded framing).

For detailed bias mitigation controls, see AI Safety & Bias Mitigation Controls.

9. Infrastructure and Security

Hosting

- Platform infrastructure runs on **AWS**, **Render**, and **Vercel**.
- Customer data is stored in regional AWS data centers (US, EU, UK, Canada) based on the customer's location.
- Database management uses PostgreSQL with encrypted storage.

Security Controls

- All data encrypted in transit (TLS 1.2+) and at rest (AES-256).
- Least-privilege access controls with MFA and SSO enforcement.
- Centralized logging with anomaly detection and alerting.

- Regular penetration testing by third-party security firms.
- SOC 2 Type II certified.

For full security details, see our Security Policy Document.

Sandboxed Execution

- All dynamically generated code and AI processing runs in fully isolated environments using containerization.
- Strict CPU, memory, and execution time limits prevent resource abuse.
- Runtime monitoring tracks execution behavior and triggers alerts for anomalies.

---

10. Model Update and Change Management

When LLM providers release new model versions or we integrate new providers:

1. **Evaluation:** New models undergo quality, bias, and safety evaluation before production deployment.
2. **Staged rollout:** Model changes are rolled out incrementally with monitoring for output quality regressions.
3. **Change documentation:** Significant model changes are documented and, where relevant, communicated to customers.
4. **Rollback capability:** The orchestration layer supports rapid rollback to previous model configurations if issues are detected.

All model changes follow our Change Management Policy.

---

11. Related Documents

- Responsible AI & Risk Management Overview
- AI Safety & Bias Mitigation Controls
- Subprocessors and Data Flow

- Security Policy Document
- Change Management Policy
- User Data Deletion and Retention Policy
- Product Features

---

12. Contact

For questions about our AI system architecture, contact us at support@syntheticusers.com.

Synthetic Users, Inc. 4223 Glencoe Ave, Suite C215-523, Marina del Rey CA 90292