



# Technical Decision Document

## Triip: AI Model & System Architecture

### Document Information

Field	Value
Product Name	Triip
Document Owner	Arnab Das
Stakeholders	[Engineering Lead, Data Science Lead, Product Manager]
Last Updated	September 25, 2025
Status	Implemented (v3)
Version	1.2 (Final)

### 1. Executive Summary

This document outlines the core technical decisions for the Triip application, a generative AI-powered travel planner for Gen Z. The key decisions, which have been implemented in the current version, are:

- Fine-Tuning Approach:** We have successfully utilized **Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA)** on a **Gemma** foundation model. The resulting specialized model, **astroani/travel-advisor-gemma**, is now hosted on Hugging Face.
- Data Strategy:** The model was trained on a custom, curated dataset, **astroani/travel\_itineraries\_dataset\_India**, which focuses on Indian travel itineraries to ensure high-quality, regionally-specific outputs for our initial target market.
- System Architecture:** The MVP and interactive demo are successfully hosted within the **Hugging Face ecosystem**, proving the viability of our core technology.

These implemented decisions directly validate the core requirements outlined in the PRD, proving our ability to deliver fast, reliable, and personalized travel itineraries.

### 2. Background and Problem Context

The PRD for Triip identifies a clear need for a travel planning tool that provides highly personalized and trustworthy recommendations. General-purpose Large Language Models (LLMs), while powerful, exhibit several limitations for this use case, including high hallucination rates and a lack of domain nuance. Our initial research confirmed that a base model accessed via simple prompt engineering is insufficient. We require a specialized model that is an expert in Gen Z travel.

### 3. Core Technical Decision: Fine-Tuning Approach

### 3.1. Justification for Fine-Tuning

Fine-tuning a foundation model was necessary to move from a generalist AI to a specialist "Travel Copilot." The primary benefits realized are:

1. **Reduced Hallucinations:** By training the model on our curated travel\_itineraries\_dataset\_India, we have significantly decreased the likelihood of fabricated or incorrect outputs, directly addressing our **Hallucination Rate guardrail ( $\leq 2\%$ )**.
2. **Improved Accuracy & Relevance:** The model has learned the patterns, locations, and terminology specific to the Indian travel domain, enabling it to provide highly accurate and relevant suggestions.
3. **Enhanced Controllability:** We have successfully instilled a specific persona and set of safety constraints into the model's behavior, ensuring a consistent user experience.

### 3.2. Implemented Approach: PEFT with LoRA on Gemma

We have successfully fine-tuned the **Gemma** foundation model using **Parameter-Efficient Fine-Tuning (PEFT) with the Low-Rank Adaptation (LoRA)** technique.

**Why LoRA was the right choice for Triip:**

- **Cost-Effectiveness:** It dramatically reduced training costs and time, allowing us to iterate and deploy a custom model rapidly.
- **Performance:** It provided performance comparable to full fine-tuning for our domain specialization task, validated through initial tests on the Hugging Face Space.
- **Agility:** The smaller size of the trained LoRA adapters made it simple to upload and manage on the Hugging Face Hub. Our final model is [astroani/travel-advisor-gemma](#).

## 4. Data Strategy and Pipeline

The quality of our fine-tuned model is entirely dependent on our training data.

Phase 1: Foundational Dataset Curation (Implemented)

The result of our initial data strategy is the astroani/travel\_itineraries\_dataset\_India dataset, now publicly available on Hugging Face. This was created by aggregating, cleaning, and structuring high-quality public data into an instruction-response format. While initially focused on India, this provides a strong template for future expansion.

Phase 2: Continuous Improvement via User Feedback (Planned)

The thumbs-up/thumbs-down feature is a critical component of our long-term strategy. User feedback on generated itineraries will be collected, anonymized, and used to create new data points for periodic re-tuning of the model, ensuring it continuously improves.

## 5. Planned Production Architecture

While the current MVP is successfully hosted on the Hugging Face ecosystem for rapid prototyping, the full production version of Triip will be built on this scalable, three-tier architecture to ensure high availability and separation of concerns as the user base grows.

1. **Client (Mobile App):** The user-facing application built with a modern framework (e.g., React Native).
2. **Backend Service (API Layer):** A service (e.g., Python/FastAPI) responsible for user management, business logic, and orchestrating calls to the AI model.
3. **AI Model Endpoint:** The fine-tuned LoRA model will be hosted on a scalable, serverless platform (e.g., Google Vertex AI, AWS SageMaker) to handle variable loads and ensure low latency (<1.5s).

## 6. Risks and Mitigation

Risk	Impact	Probability	Mitigation Strategy
<b>Dataset Geographic Limitation</b>	High	High	The current model is an expert on India only. Clearly communicate this limitation in the MVP. Plan for sourcing and integrating new datasets for other regions as the next strategic step.
<b>Scalability of MVP Architecture</b>	Medium	High	The Hugging Face Spaces free tier has performance and uptime limitations. The planned production architecture (Section 5) is the explicit mitigation strategy for long-term growth.

<b>Data Poisoning (User Feedback)</b>	High	Low	For the next phase, implement strict validation and anomaly detection in the data ingestion pipeline. Manually review samples of user-submitted data before using it for retraining.
<b>Production Vendor Lock-in</b>	Medium	Medium	The planned production architecture will use a modular interface for the AI model, allowing us to swap the underlying hosting provider (e.g., from GCP to AWS) with minimal code changes if necessary.

## 7. Approval & Change Log

### Approval Signatures:

Role	Name	Signature	Date
Engineering Lead			
Data Science Lead			
Product Manager	Arnab Das		

### Change Log:

Version	Date	Changes	Author
1.0	Sep 23, 2025	Initial Draft	Arnab Das
1.1	Sep 24, 2025	Updated to reflect the implemented model, dataset, and app on Hugging Face.	Arnab Das
1.2	Sep 25, 2025	Clarified distinction between MVP and Production architecture. Refined sections and updated risks.	Arnab Das