THUTO LMS AI TUTOR DEMO: FORMAL COST ANALYSIS

Prepared For: Thuto LMS Development Team
Prepared By: Tshemollo Rapolai
Date: July 2025

1. Overview

This document outlines the estimated monthly costs associated with running a demo of the AI-powered Tutor Assistant feature within the Thuto LMS platform. The system architecture includes the LMS core application hosted on Render, and the AI Tutor service hosted on Google Cloud Run using Gemini 1.5 Pro. MongoDB Atlas is used for storing learner interactions and details.

2. System Stack & Services

Component	Technology	Provider	Hosting Tier
Frontend	Static HTML/CSS/JS	Render	Free Tier
Backend API	Node.js	Render	Starter Plan
AI Assistant	Python (Flask) + Gemini 1.5 Pro	Google Cloud Run	Pay-as-you-go
Database	MongoDB	MongoDB Atlas	M0 Shared Cluster

3. Estimated Monthly Costs (in South African Rands)

A. Render Hosting

- **Node.js Web Service (Starter)**: R130
- **Static Frontend**: Free

Subtotal (Render): R130

B. Google Gemini API (1.5 Pro via Cloud Run)

- **Assumptions**:
    - 100 learners
    - 15 questions per learner/day
    - 30 input tokens & 120 output tokens per question
    - 30 days/month
- **Total Tokens**:
    - Input: 1.35 million tokens
    - Output: 5.4 million tokens

- **Cost Calculation**:
    - Input: $0.75/M = $1.01 ≈ R18.70
    - Output: $2.25/M = $12.15 ≈ R224.85

Subtotal (Gemini Pro): R243.55

C. Google Cloud Run (AI Flask API Hosting)

- **Estimated API Hosting Cost**: ~R25/month (based on pay-per-use pricing and traffic)

Subtotal (Cloud Run): R25

D. MongoDB Atlas (M0 Free Tier)

- **Storage Estimate**: ~90MB/month
- **Performance**: Suitable for MVP and pilot testing

Subtotal (MongoDB): R0

4. Total Monthly Estimate

Service	**Cost (ZAR)**
Render Hosting	R130.00
Gemini API (Pro)	R243.55
Gemini Recommender	R50
Google Cloud Run	R25.00
MongoDB Atlas	R0.00
Total	**R448.55**

5. Notes & Considerations

- The current setup is appropriate for **100 active learners**.
- The MongoDB M0 tier will suffice for approximately 3–4 months of data retention before approaching capacity.
- Gemini 1.5 Flash can be considered later for a cheaper alternative if less advanced responses are acceptable.
- Google Cloud Run allows pay-per-use billing, preventing unnecessary fixed monthly charges.
- Monthly API usage should be monitored via Google Cloud billing dashboard.