Gemini Pricing Plan – RAG Assistant

1. Final Model Combo

Embedding model: gemini-embedding-001
Chat model: gemini-2.5-flash

This combo is optimized for:

- - High-quality multilingual semantic search over ~1.2M words of staff, books, tabs, pages and events text.
- - Fast, low-latency answers for 1,000+ concurrent users.
- - Lowest cost within Gemini for a "serious" production workload (RAG) where quality still matters.

2. One-time Embedding Cost (gemini-embedding-001)

Corpus size assumptions:

- - Total words: ~1,194,489 words
- - Total characters: ~31,245,221 characters
- - Approx. tokens: ~8M tokens (conservative)

Pricing: gemini-embedding-001 – $0.15 per 1M tokens (input).

Estimated cost:

- - 8M tokens × $0.15 / 1M = $1.20 (one-time)
- - Even if real tokens are 10M → about $1.50.

Conclusion: Embedding cost is negligible compared to monthly chat usage.

3. Per-question Token Assumptions

For each user query, you are doing RAG: retrieve several chunks, build a context prompt, and generate a short but informative answer.

- - Input tokens (question + system prompt + top-N retrieved chunks): ≈ 2,000 tokens
- - Output tokens (answer): ≈ 500 tokens
- - Total per question: ≈ 2,500 tokens

This is a realistic number for an academic directory assistant with 3–8 retrieved chunks per question.

4. Google Pricing – Models Used

From Google's official Gemini Developer API pricing (Standard tier):

- - Embedding – gemini-embedding-001: $0.15 per 1M tokens (input).
- - Chat – gemini-2.5-flash input: $0.30 per 1M tokens.
- - Chat – gemini-2.5-flash output: $2.50 per 1M tokens (text output).

5. Base Scenario – 1,000 Users × 20 Questions/Day

We start from the base assumption you already liked:

- - Concurrent users (peak): 1,000+
- - Active users for pricing model: 1,000
- - Questions per user per day: 20
- - Days per month: 30

So:

- - Questions per day = 1,000 × 20 = 20,000
- - Questions per month ≈ 20,000 × 30 = 600,000

Tokens per month:

- - Input tokens = 600,000 × 2,000 = 1,200,000,000 (1.2B tokens)
- - Output tokens = 600,000 × 500 = 300,000,000 (0.3B tokens)

Cost:

- - Input cost = 1,200,000,000 ÷ 1,000,000 × $0.30 = $360
- - Output cost = 300,000,000 ÷ 1,000,000 × $2.50 = $750

Total monthly cost ≈ $1,110

6. Traffic Tiers – 500, 1,000, 5,000 Users

We keep the same pattern: 20 questions/day per active user, 30 days/month, and 2,000 input + 500 output tokens per question.

General formulas:

- - Questions/month = Users × 20 × 30 = 600 × Users
- - Input tokens/month = Questions × 2,000 = 1,200,000 × Users
- - Output tokens/month = Questions × 500 = 300,000 × Users
- - Input cost = 1,200,000 × Users ÷ 1,000,000 × $0.30 = $0.36 × Users
- - Output cost = 300,000 × Users ÷ 1,000,000 × $2.50 = $0.75 × Users
- - Total cost ≈ ($0.36 + $0.75) × Users = $1.11 × Users

Concrete tiers:

Users	Questions / Month	Input Tokens / Month	Output Tokens / Month	Estimated Monthly Cost
500	300,000	600,000,000	150,000,000	$555
1,000	600,000	1,200,000,000	300,000,000	$1,110
5,000	3,000,000	6,000,000,000	1,500,000,000	$5,550

7. Token Usage Bands – Budget Planning

Same models (gemini-embedding-001 + gemini-2.5-flash), but grouped by monthly question volume instead of user count.

Assumptions:

- - 2,000 input + 500 output tokens per question = 2,500 total.

Band	Questions / Month	Input Tokens	Output Tokens	Estimated Cost
Low	100,000	200,000,000	50,000,000	$185
Medium	300,000	600,000,000	150,000,000	$555
High	600,000	1,200,000,000	300,000,000	$1,110
Very High	1,000,000	2,000,000,000	500,000,000	$1,850
Extreme	2,000,000	4,000,000,000	1,000,000,000	$3,700

8. Original Detailed Pricing Explanation (Verbatim Content)

Here's a detailed pricing calculation for your decided combo: gemini-embedding-001 + gemini-2.5-flash, based on your workload assumptions.

Gemini Pricing Report v2🔍 Given Assumptions

- **Database Size:** You have approximately **1.19 million words** (or ~31.2 million characters) of text in your database (for embedding once).
- **Usage Scenario:** Many concurrent users (1,000+). Usable approximation for monthly volume:
  - Each user asks **20 questions/day** (adjustable).
  - For 1,000 users: **~20,000 questions/day**.
  - Over 30 days: **~600,000 questions/month**.
- **Average Token Usage per Question:**
  - Input (question + retrieved context) ≈ **2,000 tokens**
  - Output (answer) ≈ **500 tokens**
  - **Total ≈ 2,500 tokens/question**

✅ Pricing Rates from Google (for these models)

(From Google's published pricing)

- **Embedding Model:** `gemini-embedding-001` (paid tier)
  - Rate: **$0.15 per 1 Million tokens** (input only).
- **Chat Model:** `gemini-2.5-flash` (standard tier)
  - Input: **$0.30 per 1 Million tokens** (text/image/video).

- Output: **$2.50 per 1 Million tokens** (text output).

🧮 Embedding Cost (one-time)

Your full text corpus is estimated at **~8 million tokens**.

- **Calculation:** (8 M / 1 M) * $0.15 = **$1.20**
- *Note: Even if you embed more (say 10 M tokens) → ~$1.50. Embedding cost is negligible.*

📊 Monthly Chat Usage Cost Estimate

For **~600,000 questions/month:**

- **Total Input Tokens:** 600,000 × 2,000 = **1,200,000,000 tokens** (1.2 billion)
- **Total Output Tokens:** 600,000 × 500 = **300,000,000 tokens** (300 million)

Cost Calculation:

Component	Tokens (Total)	Rate (per 1M)	Cost Calculation	Monthly Cost
Input	1,200,000,000	$0.30	(1,200 / 1) * $0.30	**$360**
Output	300,000,000	$2.50	(300 / 1) * $2.50	**$750**
Total				**~$1,110**

🔮 Adjustments & Sensitivities

- **Lower Input Tokens (e.g., 1,000 instead of 2,000):** Cost halves.
- **Lower Output Tokens (e.g., 300 instead of 500):** Cost falls accordingly.
- **Fewer Questions (e.g., 10 instead of 20/day):** Monthly cost **~ $555**.
- **Scaling Up (1,000 users × 30 questions/day):** 900,000/month → Cost **~ $1,665**.

✅ Summary Table

Scenario	Questions/month	Estimated Monthly Cost
1,000 users × 20 q/day	~600,000	≈ $1,110
1,000 users × 10 q/day	~300,000	≈ $555
1,000 users × 30 q/day	~900,000	≈ $1,665