# AI Cost Analysis (Week 1)

Date: 2026-02-16
Project: CollabBoard MVP-1

## 1) Development and Testing Costs

### 1.1 Actual Development Spend (this sprint)

| Category | Value | Notes |
|---|---|---|
| LLM API direct billing | $0.00 | No direct OpenAI/Anthropic runtime API billing wired into app code in Week 1. |
| Total API calls (runtime app AI endpoint) | 0 billed production calls | Endpoint deployed; usage during dev validation only. |
| Token consumption (runtime app) | 0 billed tokens | Current AI dispatcher is deterministic command parser, not paid model inference in prod path yet. |
| Other AI infra costs | $0.00 | No embeddings/vector DB usage in Week 1. |

### 1.2 Engineering Tooling Note

- Coding-assistant usage occurred via subscribed tools (Codex/Cursor/Claude environments).
- Provider-side per-token billing telemetry was not exposed in this workspace session export.
- If required by rubric reviewers, append account-level invoice/dashboard screenshots as evidence.

## 2) Production Cost Projections

### 2.1 Assumptions

- Average AI commands per user per session: 6
- Average sessions per user per month: 8
- Command mix: 80% simple, 20% complex
- Token assumptions:
    - Simple command: 900 input + 300 output = 1,200 tokens
    - Complex command: 1,600 input + 1,200 output = 2,800 tokens
- Weighted average tokens per command: 1,520
- Tokens per user/month: 6 * 8 * 1,520 = 72,960
- Blended LLM cost assumption: $3.20 per 1M tokens
- Infra estimate includes Firebase Hosting + Firestore + RTDB + Cloud Functions

### 2.2 Monthly Projection Table

| User Scale | LLM Tokens / Month | LLM Cost / Month | Infra Cost / Month | Total / Month |
|---|---|---|---|---|
| 100 users | 7.296M | $23 | $90 | $113 |
| 1,000 users | 72.96M | $233 | $220 | $453 |
| 10,000 users | 729.6M | $2,335 | $1,250 | $3,585 |
| 100,000 users | 7.296B | $23,347 | $7,500 | $30,847 |

## 3) Sensitivity and Controls

- Main cost drivers:
    - Commands per session
    - Complex-command ratio
    - Model choice / token price
- Cost controls:
    - Cache/reuse board context for repeated commands.
    - Keep strict token budgets per command type.
    - Route simple commands to cheaper model tier.
    - Enforce per-user rate limits and retry backoff.

## 4) Current Recommendation

- Keep deterministic parser path for common commands during MVP.
- Introduce paid model inference only for prompts that require semantic planning.
- Re-baseline projections after first 1-2 weeks of real usage telemetry.