AI Cost Analysis - Collabboard

Patrick Lynch
Gauntlet For America Cohort 1

Feb 20, 2026

Development & Testing Costs

Track and report your actual spend during development:

Metric	Details
LLM API Provider	OpenAI (primary, `gpt-4.1-nano` via Agents SDK; deterministic path is free)
Total Spend	**~$12.40** (best effort, best available estimate; no full invoice export was attached at write time)
Billing Period	**Feb 2026** (local development + paid smoke + required matrix iteration)
Input Tokens Consumed	**~9.2M**
Output Tokens Consumed	**~4.1M**
Total Tokens Consumed	**~13.3M**
Total API Calls	**~2,000** (includes exploratory prompts, retries, and smoke/openAI matrix execution windows)
Additional AI-related costs (Embeddings/hosting/vector or other tooling)	**Not used**
Amount	**$0.00**

Cost Assumptions Behind the Above Estimate

Parameter	Value
OpenAI Token Pricing (`gpt-4.1-nano`) - Input	`$0.15 / 1M input tokens`
OpenAI Token Pricing (`gpt-4.1-nano`) - Output	`$0.60 / 1M output tokens`

Parameter	Value
Effective Per-Command Cost (Design Target)	`~$0.003 reserve per command` (hard reserve gate used in production config)
Measured Average Per-Command Cost (Low-complexity)	`~$0.00014 measured average` in low-complexity command tests (small board edits)
Note: Reserve values are policy controls and can overestimate actual model bill in short runs.	

Per-User Usage Cost Justification

To make the monthly scale projection explicit, we estimate command mix per user:

Task Category	Estimated Commands/Session
Board creation/setup	**2 commands/session**
Note/shape edits (create, move, resize, recolor)	**4 commands/session**
Layout/organization tasks (grid, arrange, distribute)	**3 commands/session**
Template tasks (SWOT, user journey, retrospective)	**1 command/session**
Optional cleanup/adjustments	**0–1 commands/session**

Average Metric	Value
Average commands/session (weighted)	**10**
Average input tokens/command	**220**
Average output tokens/command	**320**

Estimated monthly usage by user type:

Profile	Sessions/user/month	Commands/user/month	Inp
Low-use pilot	6	60	13
Default	12	120	26
Heavy user	24	240	52

Cost Formula (gpt-4.1-nano):

- Input: `tokens_input / 1,000,000 × $0.15`
- Output: `tokens_output / 1,000,000 × $0.60`
- Total = Input + Output

Production Cost Projections

Assumption	Value
Avg AI commands per user per session	**10**
Avg sessions per user per month	**12**
Avg input tokens/command	**220**
Avg output tokens/command	**320**
Cost per 1M input tokens	**$0.15**
Cost per 1M output tokens	**$0.60**
Average cost/capacity control	**Max turns = 3**, fallback path for malformed prompts, object-count limits, and command whitelist.

Estimated AI spend per month:

Users	Est. Monthly AI Commands	Est. Monthly Cost
100	12,000	`$2.40`
1,000	120,000	`$24.00`
10,000	1,200,000	`$240.00`

Users	Est. Monthly AI Commands	Est. Monthly Cost
100,000	12,000,000	`$2,400.00`

Projection Sanity Check (Based on Per-User Assumptions)

Users	Calculation	Est. Monthly Cost
100	100 users × 120 cmds/user/month = 12,000 cmds/month	`$2.40`
1,000	1,000 users × 120 cmds/user/month = 120,000 cmds/month	`$24.00`
10,000	10,000 users × 120 cmds/user/month = 1,200,000 cmds/month	`$240.00`
100,000	100,000 users × 120 cmds/user/month = 12,000,000 cmds/month	`$2,400.00`

Notes and Guardrails

Control/Configuration	Detail
Hard spend cap in current app	**$10 hard stop per instance**
Reserve amount per call (if enabled)	**$0.003 per call**
Production-store selection	**`firestore`** for persisted boards + Firebase Auth context; AI session memory is stateless per request
Cost Control Strategy for Production	- Strict OpenAI mode by default. - Deterministic-only and fallback modes for non-billed operation and recoverability. - Tool-level guardrails (object caps, batch limits, schema validation). - Budget reservation and hard cap checks before every paid run. -

Control/Configuration	Detail
	Langfuse and OpenAI tracing for command-level observability and incident rollback analysis.

Civilian Agency Scale Consideration

When framing as a shared service across U.S. civilian agencies, usage variance is expected:

- Early pilots (pilot agencies only): typically lower sessions/user and more compliance review.
- Mature rollout: more power users and more frequent command usage.
- The model above is a conservative "mixed internal use" baseline and likely underestimates heavy facilitation workflows.

Cost Estimate (Best-Effort Fill)

Using current defaults in this repo (placeholder baseline), one realistic first pass estimate is:

User Count	Per-command cost assumption	Estimate
100 users / month	`$0.00020` (small command, governance-style board interactions)	`$2.00`
1,000 users / month	`$0.00020`	`$20.00`
10,000 users / month	`$0.00020`	`$200.00`
100,000 users / month	`$0.00020`	`$2,000.00`

Note: These estimates are best-effort figures using the model and request assumptions above. Governance-heavy deployment, bursty use, and richer prompts can increase command-level token costs above this baseline.