AI Cost Analysis

AI Whiteboard — Collaborative Drawing Application

1. Development & Testing Costs

All development was performed using Claude Code (Anthropic's agentic CLI) powered by Claude Opus 4.6. The production AI endpoint uses Claude Sonnet 4.5 via the Anthropic Messages API with multi-turn tool use.

1.1 LLM API Costs

Cost Category	Model	Cost
Claude Code (development agent)	Opus 4.6 ($5 / $25 per MTok)	$52.50
Production AI endpoint (testing)	Sonnet 4.5 ($3 / $15 per MTok)	$2.07
Total LLM costs		$54.57

1.2 Total Tokens Consumed

Context	Input Tokens	Output Tokens	Total Tokens
Claude Code (Opus 4.6)	7,000,000	700,000	7,700,000
Production API testing (Sonnet 4.5)	375,000	62,500	437,500
Grand Total	7,375,000	762,500	8,137,500

Claude Code sessions accumulate context as files are read, code is generated, and tool results are processed. Later turns in a session consume significantly more input tokens due to conversation history (up to 30K+ tokens per turn). Subagent calls (Explore, Plan) account for approximately 15% of total development tokens.

1.3 Number of API Calls Made

Context	API Calls	Avg Input/Call	Avg Output/Call
Claude Code — main conversation	~400 turns	~12,000 tok	~1,500 tok
Claude Code — subagents (Explore, Plan)	~100 turns	~10,000 tok	~1,000 tok
Production AI endpoint (testing)	~75 calls	~5,000 tok	~830 tok
Total	~575 calls		

Each production AI command triggers 2–4 Anthropic API calls (multi-turn tool use). Approximately 25 AI commands were executed during testing, generating ~75 total API calls to the Sonnet 4.5 model.

1.4 Other AI-Related Costs

Service	Tier	Monthly Cost	Notes
Supabase (DB, Auth, Realtime)	Free	$0.00	PostgreSQL, RLS, Realtime Broadcast/Presence
Vercel (Hosting, Serverless)	Hobby	$0.00	Frontend SPA + /api/ai serverless function
Embeddings / Vector DB	N/A	$0.00	Not used in current architecture
Domain / CDN	N/A	$0.00	Using default .vercel.app subdomain
Total infrastructure		$0.00	

Total development & testing cost: $54.57 (LLM only; all infrastructure on free tiers).

2. Production Cost Projections

The following projections estimate monthly operating costs at four user scales. All LLM costs assume Claude Sonnet 4.5 ($3/M input, $15/M output tokens) for the production AI endpoint.

Assumptions

Parameter	Value	Rationale
AI adoption rate	40%	Not all users use the AI assistant each month
AI commands per session	5	Typical: 2–3 layout commands + 1–2 modifications
Sessions per user per month	8	~2 sessions/week for active collaborative users
Commands per active user/month	40	5 commands × 8 sessions

Token Counts per Command Type

Command Type	Example	Tool Calls	Input Tok	Output Tok	Cost/Cmd
Simple creation	"Draw a blue circle"	1–2	~8,000	~1,200	$0.042
Complex layout	"Create a flowchart, 5 steps"	5–8	~28,000	~4,000	$0.144
Query + modify	"Make all rectangles red"	2–3	~16,000	~2,000	$0.078
Content generation	"Add agenda sticky notes"	3–5	~20,000	~3,000	$0.105
Weighted average	30/25/25/20% mix	~3	~17,000	~2,400	$0.089

Input tokens grow with each tool-use turn because the full message history is re-sent. Complex commands with 5–8 tool calls accumulate 3–4× the input of a simple single-tool command.

Monthly Cost by User Scale

	100 Users	1,000 Users	10,000 Users	100,000 Users
Active AI users (40%)	40	400	4,000	40,000
AI commands/month	1,600	16,000	160,000	1,600,000
LLM cost (Sonnet 4.5)	$142	$1,424	$14,240	$142,400
Supabase	$0 (Free)	$25 (Pro)	$175 (Pro+)	$599 (Team)
Vercel	$20 (Pro)	$50 (Pro+)	$200 (Pro+)	$550 (Enterprise)
Realtime / bandwidth	$0	$25	$150	$800
Total estimated/month	$162	$1,524	$14,765	$144,349
Cost per user/month	$1.62	$1.52	$1.48	$1.44

LLM API costs represent 88–99% of total operating expenses at every scale. Infrastructure costs (Supabase, Vercel) are comparatively negligible.

Cost Optimization Strategies

- **Prompt caching:** The system prompt and 10 tool schemas (~1,500 tokens) are identical across all requests. Supabase prompt caching at 0.1× read cost could reduce input costs by 15–25%.
- **Model routing:** Route simple commands (single-tool, <2K output) to Haiku 4.5 ($1/$5 per MTok). Estimated 30% of commands qualify, saving ~25% on LLM costs.
- **Batch processing:** Non-interactive AI tasks (e.g., auto-layout, bulk content) via Anthropic Batch API at 50% discount.
- **Response streaming:** Stream partial results to reduce perceived latency without affecting token costs.
- **Board state compression:** Send only relevant objects (within viewport or referenced) instead of full board state, reducing input tokens by 30–50% for large boards.

With all optimizations applied, estimated LLM costs could be reduced by 40–60%, bringing the per-user cost to approximately $0.60–$0.90/month at scale.