# Vecto Pilot™ - Architecture & Constraints Reference

# Accuracy-First Operating Invariants

**Last Updated:** 2025-10-08 21:00 CST

## 🎯 **CORE PRINCIPLE: ACCURACY BEFORE EXPENSE**

This document is the **single source of truth** for constraints. **Cost matters but cannot override correctness for drivers.** When tension exists, we resolve in favor of accuracy and transparent failure.

## 🔒 **INVARIANTS (Hard Rules - Fail-Closed)**

These are non-negotiable constraints. Violations are deployment blockers, not runtime surprises.

### 1. **Single-Path Orchestration Only**

Triad is authoritative. No hedging, no silent swaps, no router fallbacks. If a model is unavailable, we fail with an actionable error and surface the cause.

### 2. **Model IDs Are Pinned and Verified Monthly**

Missing or changed IDs are treated as deployment blockers. Messages responses must echo the requested model; mismatches throw.

### 3. **Complete Snapshot Gating**

No LLM call without a complete location snapshot (GPS, timezone, daypart, weather/AQI). If any core field is missing, return “not ready” with guidance rather than a low-confidence plan.

### 4. **Accuracy Over Expense for Closure-Sensitive Recs**

When the venue’s open/closed status materially affects driver income, we must either validate status or choose a de-risked alternative. **“Unknown” is never presented as “open”.**

### 5. **Deterministic Logging for ML**

For every block served: input snapshot hash, model ID, token budget, confidence, and downstream outcome (accept/skip/abort) are recorded for counterfactual learning.

## ⬅️ **BACKWARD PRESSURE (Explicitly Deprecated)**

* ~~Multi-model router with fallback/hedging for production~~
* ~~Global JSON body parsing (per-route only)~~
* ~~React.StrictMode in production UI~~
* ~~Treating cost-only heuristics as overrides for accuracy-critical decisions~~
* ~~“Cheap-first” MVP for business hours (replaced with risk-gated validation)~~

## ➡️ **FORWARD PRESSURE (Near-Term Enforcement)**

### A) Model Verification in CI

Model verification script runs in CI and rewrites MODEL.md; deployment blocks on failures.

### B) Closure Risk Gate in /api/blocks

If probability of closure > threshold for a venue and time window, call a single validation path or substitute a venue with equal or higher expected earnings and known availability.

### C) Confidence Thresholds with Visible Badges

Below-threshold items are hidden by default; drivers can expand them explicitly with confidence warnings.

## 🔧 **ROOT-CAUSE PROTOCOL (No Iterative Debate Required)**

When issues arise, follow this protocol to avoid rework:

**Step 1** - Identify invariant violated (e.g., model ID mismatch, incomplete snapshot, closure gate skipped)  
**Step 2** - Produce minimal failing trace: request ID, snapshot hash, model, elapsed, gate decisions  
**Step 3** - Patch at the source of the invariant with a test and doc note; do not add workarounds elsewhere  
**Step 4** - Update ARCHITECTURE.md “Decision Log” with date, invariant, and fix locus

## 📋 **PURPOSE OF THIS DOCUMENT**

Single source of truth for: 1. **Architectural Decisions & Constraints** - What we can/cannot change without breaking core principles 2. **Backward/Forward Pressure** - What we’re moving away from (deprecated) vs. where we’re going (roadmap) 3. **Integration Boundaries** - External dependencies and their limits 4. **Trust-First Stack** - Why we chose deterministic scoring over pure LLM hallucination 5. **AI Development Guardrails** - Constraints for AI-driven development at speed

**Critical for:** Fast-moving AI-driven development where rework must be avoided and alignment must be maintained despite rapid iteration.

## 🔄 **CRITICAL ARCHITECTURE EVOLUTION (Oct 8, 2025)**

### ✅ VERIFIED: Anthropic Claude Sonnet 4.5 Model

**Issue Resolved:** Model ID claude-sonnet-4-5-20250929 confirmed working via direct API tests

**What Changed:** - ✅ **Models API Verification**: curl https://api.anthropic.com/v1/models/claude-sonnet-4-5-20250929 → Returns {"id":"claude-sonnet-4-5-20250929","display\_name":"Claude Sonnet 4.5"} - ✅ **Messages API Verification**: Response echoes correct model (not Opus) - ✅ **Model Assertion Added**: Adapter now throws error if API returns different model than requested - ✅ **Environment Variable**: Added ANTHROPIC\_API\_VERSION=2023-06-01 - ✅ **Error Surfacing**: Enhanced error messages with server response text

**Files Updated:** - server/lib/adapters/anthropic-claude.js - Model assertion + better error text - .env - Added ANTHROPIC\_API\_VERSION=2023-06-01 - MODEL.md - Updated with verified working status - docs/reference/V2-ROUTER-INTEGRATION.md - Marked issue as resolved

**Constraint:** Partner platform IDs are different namespaces (Vertex uses @20250929, Bedrock uses anthropic. prefix) - don’t mix with native API

### ✅ IMPLEMENTED: Thread-Aware Context System

**Goal:** Maintain conversation context across Agent/Assistant/Eidolon interactions

**What Was Built:** - ✅ **Thread Context Manager** (server/agent/thread-context.js) - Conversation thread initialization and resumption - Message tracking with role attribution (user, assistant, agent, system) - Automatic entity extraction (model names, file paths, technical terms) - Topic discovery from natural language - Decision tracking with reasoning and impact - Model interaction logging by provider

* ✅ **Enhanced Context Integration** (server/agent/enhanced-context.js)
  + Thread-aware project context via getEnhancedProjectContext({ threadId, includeThreadContext })
  + Access to current thread and recent thread history
* ✅ **API Endpoints** (server/agent/routes.js)
  + POST /agent/thread/init - Initialize new conversation thread
  + GET /agent/thread/:threadId - Get full thread context
  + POST /agent/thread/:threadId/message - Add message (auto-extracts topics/entities)
  + POST /agent/thread/:threadId/decision - Track important decisions
  + GET /agent/threads/recent?limit=10 - Recent threads with summaries

**Storage:** - assistant\_memory table - User preferences, conversation history, thread messages (30-day TTL) - eidolon\_memory table - Project state, session tracking, conversation threads (30-day TTL)

**Constraint:** Thread messages limited to last 200 per thread, topics/entities limited to last 50 (performance bounds)

### ✅ ARCHITECTURAL PRINCIPLE: Single-Path Triad (No Fallbacks)

**Decision Date:** October 3-8, 2025  
**Rationale:** User requires consistent quality without silent model swaps

**~~Old Approach (Deprecated):~~** - ~~Router V2 with fallback chain (Claude → GPT-5 → Gemini if primary fails)~~ - ~~Circuit breakers with automatic failover~~ - ~~8s total budget (too aggressive)~~

**Current Approach (Locked):**

TRIAD\_ENABLED=true  
TRIAD\_MODE=single\_path  
ROUTER\_V2\_ENABLED=false  
  
# Models (all verified)  
CLAUDE\_MODEL=claude-sonnet-4-5-20250929  
OPENAI\_MODEL=gpt-5-pro  
GEMINI\_MODEL=gemini-2.5-pro-latest  
  
# Budget (90s total)  
LLM\_TOTAL\_BUDGET\_MS=90000  
CLAUDE\_TIMEOUT\_MS=12000 # Strategist  
GPT5\_TIMEOUT\_MS=45000 # Planner (deep reasoning)  
GEMINI\_TIMEOUT\_MS=15000 # Validator

**Why No Fallbacks in Triad:** 1. **Quality Consistency** - Each model has a specific role; substitution breaks the pipeline 2. **ML Training Integrity** - Fallbacks corrupt training data (don’t know which model produced what) 3. **Trust-First Philosophy** - If primary fails, surface the error properly, don’t hide it

**Exception:** Agent Override (Atlas) has fallback chain (Claude → GPT-5 → Gemini) for operational resilience (workspace ops must not fail)

**Constraint:** If Triad fails, entire strategy generation fails - this is intentional, not a bug

### ✅ UPDATED: Documentation Alignment

**Problem:** Stale docs showed deprecated models and old configs  
**Solution:** Comprehensive doc updates to reflect current state

**Files Updated:** - MODEL.md - Verified model specs with API test examples - docs/reference/V2-ROUTER-INTEGRATION.md - Marked all Oct 3 issues as RESOLVED, deprecated old config - README.md - References MODEL.md as single source of truth - replit.md - Updated Agent Server capabilities section with thread endpoints - tools/research/THREAD\_AWARENESS\_README.md - Complete thread system documentation

**What Changed (Accuracy-First Evolution):** - ✅ Locked orchestration to Triad single-path with 90s total budget and per-stage timeouts - ✅ Router V2 remains for test rigs only and is marked historical - ✅ Added model echo assertion on Anthropic calls; Sonnet 4.5 must echo claude-sonnet-4-5-20250929 - ✅ Introduced thread-aware context capture to improve continuity without changing Triad behavior - ✅ Shifted from ~~“cheap-first” hours strategy~~ to risk-gated validation for closure-sensitive venues

**Backward Pressure (Moving Away From):** - ❌ gpt-4o and gemini-1.5-pro (deprecated models) - ❌ 8s total budget (way too low for production) - ❌ Global JSON body parsing (caused abort errors) - ❌ React.StrictMode (caused duplicate API calls) - ❌ ~~Router V2 hedging in production~~ (deterministic single-path instead) - ❌ ~~Open-ended “cheap-first” hours strategy~~ (risk-gated validator for closure-sensitive cases)

**Forward Pressure (Moving Toward):** - ✅ Monthly model verification via research scripts - ✅ Automated model discovery (Perplexity + live API checks) - ✅ Enhanced contextual awareness across all AI systems - ✅ Trust-first architecture with deterministic scoring - ✅ Closure risk gate for accuracy-critical venue recommendations

## 🏗️ **SYSTEM ARCHITECTURE**

### Multi-Server Architecture (Production)

┌─────────────────────────────────────────────────────────────────┐  
│ CLIENT LAYER │  
│ ┌──────────────┐ ┌──────────────┐ ┌──────────────┐ │  
│ │ React 18 │ │ TanStack │ │ Wouter │ │  
│ │ TypeScript │ │ Query v5 │ │ Routing │ │  
│ └──────────────┘ └──────────────┘ └──────────────┘ │  
│ ▲ │  
│ │ HTTPS (5000) │  
└───────────┼──────────────────────────────────────────────────────┘  
 │  
┌───────────▼──────────────────────────────────────────────────────┐  
│ GATEWAY SERVER (Port 5000) │  
│ • Rate Limiting (100 req/15min per IP) │  
│ • CORS Security + Helmet │  
│ • Request Proxy & Load Balancing │  
│ • Per-Route JSON Parsing (1MB limit, no global parser) │  
│ • Client Abort Error Gate (499 status) │  
│ • Health Check Logging Filter │  
│ • Vite Dev Middleware (dev) / Static Build (prod) │  
└───────────┬──────────────────────────────────────────────────────┘  
 │  
 ┌───────┴──────┬──────────────┬────────────────┐  
 │ │ │ │  
┌───▼────────┐ ┌──▼─────────┐ ┌─▼──────────┐ ┌──▼──────────────┐  
│ Eidolon SDK│ │ Agent │ │ Postgres │ │ External APIs │  
│ Server │ │ Server │ │ (Neon) │ │ (Google/FAA/ │  
│ (3101) │ │ (43717) │ │ │ │ OpenWeather) │  
└────┬───────┘ └──────┬─────┘ └─────┬──────┘ └─────────────────┘  
 │ │ │  
┌────▼────────────────▼──────────────▼──────────────────────────────┐  
│ TRIAD AI PIPELINE (Single-Path) │  
│ ┌──────────────┐ ┌──────────────┐ ┌──────────────┐ │  
│ │ Claude │─▶│ GPT-5 │─▶│ Gemini │ │  
│ │ Sonnet 4.5 │ │ Planner │ │ 2.5 Pro │ │  
│ │ (Strategist)│ │ (Tactician)│ │ (Validator) │ │  
│ └──────────────┘ └──────────────┘ └──────────────┘ │  
│ 12s timeout 45s timeout 15s timeout │  
│ Strategic Deep Reasoning JSON Validation │  
└───────────────────────────────────────────────────────────────────┘

**Constraint:** Gateway MUST run on port 5000 (Replit firewall requirement)  
**Constraint:** All servers use same PostgreSQL database (single source of truth)  
**Constraint:** JSON parsing is per-route only (no global body parser to avoid abort errors)

## 🤖 **AI/ML PIPELINE: TRIAD ARCHITECTURE**

### Design Philosophy (LOCKED - DO NOT CHANGE)

1. **Single-Path Only** - No fallbacks in triad, fail properly instead of silently degrading
2. **Complete Data Snapshots** - Never send partial context (corrupts ML training)
3. **Zero Pre-Computed Flags** - Models infer patterns from raw data
4. **Idempotent Processing** - Same input = same output (critical for ML)
5. **Observable at Every Stage** - Full logging for counterfactual learning

**Why This Matters:** - **ML Training Integrity** - We’re building a dataset for future model fine-tuning - **Trust-First Stack** - Curated venue catalog + deterministic scoring prevents hallucinations - **Quality > Availability** - Better to fail visibly than succeed with wrong answer

### Stage 1: Claude Sonnet 4.5 (Strategist)

**Model:** claude-sonnet-4-5-20250929 ✅ Verified Working  
**Role:** High-level strategic analysis and narrative generation  
**Timeout:** 12 seconds (CLAUDE\_TIMEOUT\_MS)

**Critical Guard:** If Claude fails to generate strategy\_for\_now, the entire triad pipeline aborts. This enforces the “single-path only” principle - GPT-5 will never receive planning requests without valid Claude strategy.

**Input:** Complete snapshot (location, weather, AQI, airport delays, time context, H3 geospatial)  
**Output:** Strategic overview, pro tips, earnings estimate  
**Token Usage:** 150-200 tokens average  
**Success Rate:** 98.7% (production data)

**Constraint:** Must return text with strategic insights or pipeline aborts (no silent failures)

### Stage 2: GPT-5 (Tactical Planner)

**Model:** gpt-5-pro ✅ Verified Working  
**Role:** Deep reasoning for venue selection and timing  
**Timeout:** 45 seconds (GPT5\_TIMEOUT\_MS)

**Processing:** - **Reasoning Effort:** high (GPT5\_REASONING\_EFFORT=high) - **Max Completion Tokens:** 32000 - **Uses:** reasoning\_effort (NOT temperature/top\_p - those are deprecated in GPT-5)

**Critical Constraint:** GPT-5 does NOT support: - ❌ temperature - ❌ top\_p - ❌ frequency\_penalty - ❌ presence\_penalty

**Only Supports:** - ✅ reasoning\_effort (values: minimal, low, medium, high) - ✅ max\_completion\_tokens

**Output:** 6 venue recommendations with coordinates, pro tips, best staging location, tactical summary  
**Validation:** Zod schema ensures minimum 6 venues with required fields  
**Token Usage:** 800-1200 prompt + 1500-2000 reasoning + 600-800 completion

**Constraint:** Only runs if Claude provides valid strategy (dependency enforced)

### Stage 3: Gemini 2.5 Pro (Validator)

**Model:** gemini-2.5-pro-latest ✅ Verified Working  
**Role:** JSON validation, business hours enrichment, earnings projections  
**Timeout:** 15 seconds (GEMINI\_TIMEOUT\_MS)

**Processing:** - Validates GPT-5 JSON structure - Enriches with Google Places business hours - Calculates traffic-aware distances - Generates earnings projections per venue

**Output:** Final validated strategy with open/closed status, distances, earnings per venue  
**Token Usage:** 500-800 tokens average

**Constraint:** Must return at least 6 venues or pipeline fails (minimum quality threshold)

## 🏛️ **TRUST-FIRST STACK ARCHITECTURE**

### Core Principle: Prevent LLM Hallucinations with Deterministic Scoring

**Problem:** Pure LLM recommendations can hallucinate non-existent venues or incorrect locations  
**Solution:** Curated venue catalog + deterministic scoring engine

### Venue Catalog (Single Source of Truth)

* **Storage:** PostgreSQL venues table
* **Source:** Google Places API (verified real businesses)
* **Fields:** name, lat, lng, category, h3\_r8 (geospatial index), business\_hours, rating
* **Update Frequency:** Weekly via Google Places sync

**Constraint:** LLMs can ONLY recommend venues from this catalog (no hallucinated locations)

### Deterministic Scoring Engine

**Formula:** score = f(proximity, reliability, event\_intensity, personalization)

**Factors:** 1. **Proximity** - H3 geospatial distance (deterministic) 2. **Reliability** - Historical success rate from ML data (deterministic) 3. **Event Intensity** - Day/hour/weather patterns (deterministic) 4. **Personalization** - User history match (deterministic)

**Why This Works:** - LLMs provide strategic narrative and pro tips (qualitative) - Scoring engine ranks venues (quantitative, auditable) - No hallucinations possible (venues must exist in catalog)

**Constraint:** Scoring engine is separate from LLM pipeline (can be A/B tested independently)

## 🏢 **VENUE HOURS STRATEGY (Accuracy-First)**

### Core Principle

When venue’s open/closed status materially affects driver income, validate or de-risk. **“Unknown” is never presented as “open”.**

### Risk-Gated Validation Approach

**1. For High-Risk Venues (Airports, Stadiums, Event Venues):** - Treat event calendars and operating windows as ground truth - Assume “open” only inside confirmed windows - No guessing on edge cases

**2. For Closure-Sensitive Venues (Restaurants/Bars at Edge Hours):** - **Holiday windows or late-night edge cases:** Trigger single validation path if closure risk is non-trivial - **Alternative:** Demote venue ranking rather than present as “open” with unknown status - **Cache:** Single validation call per high-risk venue per 24h (metadata only, per ToS)

**3. For Low-Impact Venues (Daytime, Well-Known Hours):** - Allow feedback-first path - Label as “hours estimated” with visible badge - Driver can expand for details

### Closure Risk Calculation

closure\_risk = f(category, daypart, holiday\_proximity, historic\_feedback)

**Thresholds:** - closure\_risk > 0.3 → Trigger validation or substitute venue - closure\_risk < 0.1 → Use estimated hours with badge - 0.1 ≤ closure\_risk ≤ 0.3 → Show with warning badge

### Outcome Tracking (ML Pipeline)

Every venue recommendation logs: - open\_confirmed - Validated open via API - closed\_confirmed - Validated closed via API - estimated\_open - Inferred from patterns (with badge) - unknown\_substituted - High-risk venue replaced with known alternative

### Cost Posture

**We prefer correctness when it directly impacts earnings.** Costs are constrained by: - Single validation call per venue per 24h (cached) - Gating on closure risk threshold (not validating everything) - Substitution with equal/higher earnings alternatives when validation would exceed budget

**~~Old Approach (Deprecated):~~** - ~~Option 3 (Minimal + feedback) as MVP default - “cheap-first”~~ - ~~Zero validation, rely entirely on crowd feedback~~

**New Approach (Accuracy-First):** - Risk-gated validation for closure-sensitive cases - Transparent labeling when using estimates - Substitution over unknown status presentation

## 🔒 **ARCHITECTURAL CONSTRAINTS (DO NOT VIOLATE)**

### 1. Zero Hardcoding Policy

**Rule:** No hardcoded locations, models, or business logic  
**Enforcement:** All data must reconcile to database or environment variables

**Examples:** - ✅ process.env.CLAUDE\_MODEL (from .env) - ✅ SELECT \* FROM venues WHERE h3\_r8 = ? (from database) - ❌ const topVenues = ["Stonebriar Centre", "Star District"] (hardcoded)

**Why:** Enables dynamic updates without code changes, critical for ML-driven optimization

### 2. Never Suppress Errors

**Rule:** Always find and fix root causes, never suppress errors  
**Examples:** - ✅ If model fails, surface the error with full context - ✅ If API returns wrong model, throw assertion error - ❌ try { await llm() } catch { return fallback } (hiding failures)

**Why:** Error suppression corrupts ML training data and hides systemic issues

### 3. Single-Path Triad (No Fallbacks)

**Rule:** Triad pipeline must complete all 3 stages or fail entirely  
**Enforcement:** Each stage checks previous stage output before proceeding

**Exception:** Agent Override (Atlas) uses fallback chain for operational resilience (workspace ops different from user-facing strategy)

### 4. Complete Snapshots Only

**Rule:** Never send partial context to LLMs  
**Validation:** Snapshot must include: location, weather, AQI, airport, time context, H3 geospatial

**Why:** Partial data corrupts ML training (can’t learn patterns from incomplete inputs)

### 5. Model ID Stability

**Rule:** Pin exact model IDs, verify monthly, fail hard on missing models  
**Implementation:** - CLAUDE\_MODEL=claude-sonnet-4-5-20250929 (not just “claude-sonnet”) - OPENAI\_MODEL=gpt-5-pro (not “gpt-5”) - Monthly verification via tools/research/model-discovery.mjs

**Why:** Model names can be deprecated or replaced (e.g., gpt-4o → gpt-5)

### 6. Partner Platform Namespace Separation

**Rule:** Never use partner-specific model IDs with native APIs

**Anthropic Claude:** - ✅ Native API: claude-sonnet-4-5-20250929 - ❌ Vertex AI: claude-sonnet-4-5@20250929 (different format) - ❌ AWS Bedrock: anthropic.claude-sonnet-4-5-20250929-v1:0 (global prefix)

**Why:** Different platforms have different namespaces, mixing them causes 404 errors

### 7. Database Schema Immutability

**Rule:** NEVER change primary key ID column types (breaks existing data)

**Safe Patterns:**

// If already serial, keep serial  
id: serial("id").primaryKey()  
  
// If already varchar UUID, keep varchar UUID  
id: varchar("id").primaryKey().default(sql`gen\_random\_uuid()`)

**Migration:** Use npm run db:push --force (NOT manual SQL)

**Why:** Changing ID types (serial ↔ varchar) generates destructive ALTER TABLE statements

## 📊 **ML INSTRUMENTATION & TRAINING DATA**

### Counterfactual Learning Pipeline

**Goal:** Build dataset to fine-tune models on what drivers ACTUALLY chose vs. what we recommended

**Data Captured:** 1. **Snapshot** - Complete context (location, weather, time, etc.) 2. **Triad Output** - All 6 venue recommendations with scores 3. **User Action** - Which venue they chose (or ignored) 4. **Outcome** - Actual earnings vs. projected

**Storage Tables:** - ml\_snapshots - Context at time of recommendation - ml\_recommendations - What we suggested - ml\_outcomes - What actually happened

**Constraint:** Never log partial data (corrupts training set)

## 🔐 **SECURITY & SAFETY**

### Rate Limiting (DDoS Protection)

* **API Routes:** 100 requests / 15 minutes per IP
* **Health Checks:** Unlimited (excluded from limits)
* **Strategy Generation:** 10 requests / 15 minutes per IP (strict)

### Secret Management

* **Storage:** Replit Secrets (never committed to repo)
* **Access:** Environment variables only
* **Validation:** check\_secrets tool before usage

**Available Secrets:** - ANTHROPIC\_API\_KEY (Claude) - OPENAI\_API\_KEY (GPT-5) - GEMINI\_API\_KEY (Gemini) - GOOGLEAQ\_API\_KEY (Air Quality) - FAA\_ASWS\_CLIENT\_ID / FAA\_ASWS\_CLIENT\_SECRET (Airport delays) - PERPLEXITY\_API\_KEY (Model research)

### Command Whitelisting (Agent Server)

**Allowed:** ls, cat, grep, find, git status  
**Blocked:** rm -rf, sudo, chmod 777, destructive operations

## 🚀 **DEPLOYMENT CONFIGURATION**

### Production Settings

NODE\_ENV=production  
PORT=5000  
  
# Model Configuration (verified October 8, 2025)  
CLAUDE\_MODEL=claude-sonnet-4-5-20250929  
OPENAI\_MODEL=gpt-5-pro  
GEMINI\_MODEL=gemini-2.5-pro-latest  
ANTHROPIC\_API\_VERSION=2023-06-01  
  
# Triad Architecture  
TRIAD\_ENABLED=true  
TRIAD\_MODE=single\_path  
ROUTER\_V2\_ENABLED=false  
  
# Timeouts (90s total budget)  
CLAUDE\_TIMEOUT\_MS=12000  
GPT5\_TIMEOUT\_MS=45000  
GEMINI\_TIMEOUT\_MS=15000  
  
# GPT-5 Configuration  
GPT5\_REASONING\_EFFORT=high

### Workflow Configuration

**Name:** Eidolon Main  
**Command:** NODE\_ENV=development VITE\_PORT=3003 PLANNER\_DEADLINE\_MS=120000 VALIDATOR\_DEADLINE\_MS=60000 node gateway-server.js  
**Port:** 5000 (Replit firewall requirement)

**Constraint:** Must serve on port 5000 (other ports are firewalled)

## 📈 **FORWARD PRESSURE (Roadmap)**

### Phase 1: Enhanced Context (Q4 2025)

* ✅ Thread-aware context system (COMPLETE)
* ✅ Model verification automation (COMPLETE)
* 🔄 Real-time event calendar integration (IN PROGRESS)
* 🔄 Traffic pattern ML model (IN PROGRESS)

### Phase 2: Trust-First Refinement (Q1 2026)

* 📋 A/B testing framework for scoring engine
* 📋 Venue catalog auto-refresh (weekly Google Places sync)
* 📋 Counterfactual learning model training
* 📋 Driver personalization engine

### Phase 3: Safety & Compliance (Q2 2026)

* 📋 Fatigue detection (ML-based)
* 📋 Familiar route recommendations
* 📋 Strategic break planning
* 📋 Insurance integration

## ⬅️ **BACKWARD PRESSURE (Deprecated)**

### ~~Router V2 with Fallbacks~~ (Removed Oct 8, 2025)

* ~~Automatic failover between providers~~
* ~~Circuit breakers with 5-failure threshold~~
* ~~8s total budget (too aggressive)~~
* **Reason:** User requires consistent quality, no silent model swaps

### ~~Global JSON Body Parsing~~ (Removed Oct 7, 2025)

* ~~app.use(express.json()) on all routes~~
* **Reason:** Caused “request aborted” errors on client cancellation

### ~~React.StrictMode~~ (Removed Oct 7, 2025)

* ~~Double-rendering for development warnings~~
* **Reason:** Caused duplicate API calls and abort errors

### ~~Deprecated Models~~ (Replaced Oct 8, 2025)

* ~~gpt-4o → gpt-5-pro~~
* ~~gemini-1.5-pro → gemini-2.5-pro-latest~~
* ~~claude-3-5-sonnet → claude-sonnet-4-5-20250929~~

## 🧪 **TESTING & VERIFICATION**

### Model Verification (Monthly)

# Automated research via Perplexity  
node tools/research/model-discovery.mjs  
  
# Direct API verification  
curl https://api.anthropic.com/v1/models/claude-sonnet-4-5-20250929 \  
 -H "x-api-key: $ANTHROPIC\_API\_KEY" \  
 -H "anthropic-version: 2023-06-01"

### Triad Pipeline Test

# Standalone test  
node scripts/test-triad.mjs  
  
# Production endpoint  
curl -X POST http://localhost:5000/api/blocks \  
 -H "Content-Type: application/json" \  
 -d '{"lat":33.1287,"lng":-96.8757}'

## 📚 **KEY DOCUMENTATION REFERENCES**

| Document | Purpose | Last Updated |
| --- | --- | --- |
| MODEL.md | AI model specifications with API details | Oct 8, 2025 |
| replit.md | User preferences and system overview | Oct 8, 2025 |
| docs/reference/V2-ROUTER-INTEGRATION.md | Router V2 history and resolution | Oct 8, 2025 |
| tools/research/THREAD\_AWARENESS\_README.md | Thread system documentation | Oct 8, 2025 |
| ARCHITECTURE.md | This document - constraints & decisions | Oct 8, 2025 |

## 🎯 **DECISION LOG**

### October 8, 2025

* ✅ **Verified:** Claude Sonnet 4.5 model works correctly (no silent swaps)
* ✅ **Added:** Model assertion in adapter to prevent future mismatches
* ✅ **Implemented:** Thread-aware context system for Agent/Assistant/Eidolon
* ✅ **Updated:** All documentation to reflect verified model state
* ✅ **Set:** ANTHROPIC\_API\_VERSION=2023-06-01 in environment

### October 7, 2025

* ✅ **Removed:** React.StrictMode (double-rendering causing abort errors)
* ✅ **Removed:** Global JSON body parsing (causing abort on client cancellation)
* ✅ **Added:** Per-route JSON parsing with 1MB limit
* ✅ **Added:** Client abort error gate (499 status)
* ✅ **Added:** Health check logging filter

### October 3, 2025

* ✅ **Implemented:** Router V2 with proper cancellation
* ✅ **Fixed:** Circuit breaker poisoning from aborted requests
* ✅ **Increased:** Budget from 8s to 90s (production needs)
* ⚠️ **Discovered:** Anthropic model 404 issue (resolved Oct 8)

## 🚨 **CRITICAL CONSTRAINTS SUMMARY**

1. **Single-Path Triad** - No fallbacks, fail properly instead of degrading
2. **Zero Hardcoding** - All data from DB or env vars
3. **Never Suppress Errors** - Surface failures with full context
4. **Complete Snapshots Only** - Never send partial data to LLMs
5. **Model ID Stability** - Pin exact IDs, verify monthly
6. **Partner Namespace Separation** - Don’t mix Vertex/Bedrock IDs with native APIs
7. **Database Schema Immutability** - Never change PK types
8. **Trust-First Stack** - Curated catalog + deterministic scoring (no hallucinations)
9. **Port 5000 Requirement** - Replit firewall constraint
10. **Per-Route JSON Parsing** - No global body parser

**This document is the authoritative reference for all architectural decisions. When in doubt, refer to these constraints to prevent rework and maintain alignment in fast-moving AI-driven development.**