PRIMEX SOVREIGN — System Blueprint (ARCHITECT v1.0)

> **Purpose**: Provide a complete, high-level architecture and modular blueprint for PRIMEX SOVREIGN (dual-mode AI platform: public + private "sovereign" mode), spanning product, tech, infra, security/compliance, GTM and ops. This document is designed for execution—each block maps to a backlog and an owner.

---

0) Macrodeck

Objectives

- Ship a dual-mode AI assistant that toggles between **Public Mode** (consumer/pro) and **Sovereign Mode** (Tyler-only, elevated capabilities, full data control).
- Launch iOS, Android, and Web with a **unified design system**, privacy-first defaults, and **on-device assist** for speed + trust.
- Build a durable moat via **offline/edge intelligence**, **private knowledge orchestration**, and **provenance/watermarking**.

Assets

- Brand/IP (PRIMEX, SOVREIGN), domains, design system, component library, data assets, model gateways, eval harness, growth loops, partner integrations, content provenance pipeline, prompts/agents marketplace.

Risks (top)

- Platform policy shifts (App Store/Play, EU DMA), AI safety/regulatory, copyright/data licensing, GPU supply shocks, retention economics.

Command Loop

- Weekly: roadmap checkpoint → ship plan → KPI readout → risk triage → unblock.

---

1) Product Pillars

1. **Dual-Mode UX**

2. **Public Mode**: Conversational AI, agents, workflows, RAG with user vault, voice+vision, shareable canvases.

3. **Sovereign Mode (Owner)**: Full admin console, custom tools, unrestricted graph orchestration, private model endpoints, raw telemetry, model routing rules, scripting (“Ops Console”).
4. Single binary/app; mode switch via **role + policy**; UI theme swap + capability gates.
5. **Knowledge Orchestration**
6. Personal/Team vaults (notes, files, links, emails), connectors (Drive, Notion, Slack, GitHub).
7. RAG pipeline with policy-aware retrieval (PII filtering, per-space ACLs).
8. Memory with **scoped recall** (session / project / account) and TTLs.
9. **Agent Framework**
10. Tool abstraction (web search, calendars, sheets, code, DB),
11. Deterministic planners + reflection loop,
12. Guardrails (policy, rate limits, safety classifiers),
13. Task hub: runbooks, automations, scheduled jobs.
14. **Edge Intelligence**
15. On-device ASR/TTS, small LLM for intent + redaction + offline quick answers.
16. Progressive enhancement: upgrade to cloud models when consented/available.
17. **Trust & Authenticity**
18. Content provenance (C2PA/Content Credentials) for media generation.
19. Clear AI disclosures in UI, data controls, model cards, per-feature consent.
20. **Growth System**
21. Shareable outputs (links, embeds), referral loops, creator payouts (prompts/agents marketplace), affiliate & influencer mechanics with compliant disclosures.

---

2) System Architecture (High-Level)

```
[ Clients ]  iOS (SwiftUI) | Android (Kotlin/Jetpack) | Web (Next.js/React)
     |
     v
[ Edge Layer ] On-device ASR/TTS • Lightweight LLM (intent, redaction) • Caches
```

```
     |
     v
[ API Gateway ]  GraphQL/REST • AuthN/AuthZ • Rate limiting • Feature flags
     |
     +--> [ Orchestration ]  Conversations • Agent runtime • Workflows • Jobs
     |        |
     |        +--> [ Tooling Bus ] Web, Calendar, Email, Files, Search, Code, DB
     |        +--> [ Policy Engine ] Permissions • Safety • Data residency
     |        +--> [ Evals/Observability ] Traces • Metrics • Feedback
     |
     +--> [ Knowledge Layer ]  RAG Services • Vector DB • Document store
     |
     +--> [ Model Gateway ]  Routing across Open/Closed models • On-device •
Self-hosted
     |
     +--> [ Media Services ]  Vision • OCR • Transcription • Synthesis •
Provenance (C2PA)
     |
     +--> [ Payments/Billing ] IAPs • Subscriptions • Entitlements • Invoices
     |
     v
[ Data Plane ]  Postgres (OLTP) • Object Storage (S3) • Vector DB (pgvector/
Qdrant) • Redis
     |
     +--> [ Analytics ]  Events (Kafka/Redpanda) • Warehouse (BigQuery/
Snowflake) • BI (Metabase)

[ Sec/Compliance ] KMS/Vault • Secrets • Audit logs • MASVS/ASVS controls • DLP
• PII vault
[ DevOps/MLOps ] K8s (GKE/EKS) • CI/CD • Model registry • Canary • A/B • Infra
as Code (Terraform)
```

Architecture style: Start with a **modular monolith** (NestJS or FastAPI + background worker) to speed iteration; separate heavy services (vector/RAG, media, model-router) behind internal APIs. Extract to microservices as load and team size grow.

---

3) Core Modules & Owners

A) Client Apps

- **Design System**: Tailwind/React primitives → SwiftUI + Jetpack Compose tokens.
- **Offline Core**: local caches, on-device embeddings; graceful degradation.
- **Voice/Camera**: streaming ASR, VAD, wake word, live transcription, frames to vision API.
- **Compliance Surfaces**: age gates, report/block, content filters, account deletion, data export.

B) API & Identity

- **Auth**: email/pass + passkeys; OAuth (Google/Apple); orgs + roles + scopes.
- **Entitlements**: plans, trials, usage quotas; server-authoritative; feature flags.
- **Rate Limiters**: per-user, per-org, per-tool; circuit breakers for costly routes.

C) Conversation/Agent Runtime

- **State**: threads, messages, tool calls, artifacts, feedback.
- **Planning**: multi-step tool plans; function-calling; retry + timeouts.
- **Guardrails**: prompt templates, safety policies, regex/AST filters, allow/deny tools per role.
- **Evals**: regression suites; golden sets for truthfulness, safety, latency, cost.

D) Knowledge & RAG

- **Ingestion**: connectors (Drive/Notion/Slack/GitHub/Web), chunkers, MIME parsers.
- **Index**: embeddings, hybrid search, metadata security labels.
- **Retrieval**: re-rankers, query planners, **privacy filters** before model exposure.
- **Rewrite**: answer synthesis with source attributions and confidence.

E) Model Gateway

- **Providers**: OpenAI, Anthropic, Google, Mistral, Meta (Llama), local (GGUF), vision/STT/TTS vendors.
- **Router**: policy-aware (PII, region), cost/latency SLOs, fallback trees.
- **Telemetry**: token/cost accounting, per-model win-rates, drift alerts.

F) Media Intelligence

- **Vision**: OCR, doc QA, layout parse; image understanding.
- **Synthesis**: image/video generation via approved providers; **C2PA provenance** stamping.
- **Redaction**: face/PII blur; watermark detection.

G) Trust, Safety & Compliance

- **Policy Engine**: age gating, UGC moderation, in-app report/block, appeals.
- **Data Controls**: per-connector consent, delete/export, residency pinning.
- **Audits**: immutable logs; model usage attestations; DPIAs; vendor DPA library.
- **Security**: MASVS/ASVS control mapping, MASTG checks, supply-chain scanning.

H) Monetization

- **Payments**: Apple IAP / Google Play Billing / Stripe (web).
- **Plans**: Free, Pro, Team, **Sovereign License** (owner-only bundle).
- **Marketplace**: prompts/agents/tools with rev-share, ratings, provenance.

I) Analytics & Growth

- **Growth Events**: acquisition → activation → retention → revenue → referral.
- **ASO & Store Surfaces**: screenshots, preview video, localizations, custom listings.

- **Attribution**: SKAN / Play Install Referrer; privacy-preserving cohorting.
- **Creator Program**: referral codes, UGC templates, compliance auto-disclosures.

---

4) Security, Privacy & Compliance Baseline

Security Standards

- **OWASP MASVS** (mobile) + **OWASP ASVS** (web/API), periodic pen-tests, SBOMs, SCA.
- Secrets in KMS/Vault, least-privilege IAM, static e2e encryption (AES-256 + TLS 1.3).

Privacy

- GDPR/CCPA/CPRA aligned: DSR portal (access/delete/export), purpose-limited processing, sensitive data toggles.
- Account deletion in-app (iOS/Android policy), **age-appropriate design** for minors, COPPA guardrails if applicable.

AI Transparency

- AI disclosures at first interaction; deepfake/provenance labeling; content credentials for generated media.

Store Compliance Surfaces

- UGC: report/block, proactive moderation, custom EULA, spam controls.
- AI: in-app reporting for harmful outputs; safety filters; restricted domains (no illegal facilitation).

Data Governance

- PII vault, region pins, data retention schedule, redaction before training, vendor DPIAs, DLP patterns in ingestion.

---

5) Infrastructure & SRE

- **Cloud**: GCP or AWS.
- **Compute**: K8s (GKE/EKS), autoscaling; spot for batch.
- **GPU**: cloud GPUs + specialist providers for training/fine-tune and heavy inference.
- **Datastores**: Postgres (OLTP), Redis (cache/queues), S3 (assets), Vector DB (pgvector/Qdrant), Warehouse (BigQuery/Snowflake).
- **Observability**: OpenTelemetry, Prometheus/Grafana, Sentry, ClickHouse for logs, feature flags (GrowthBook/Unleash).
- **CI/CD**: GitHub Actions, Canary, Blue/Green; Infra as Code (Terraform), policy as code (OPA).
- **Backups/DR**: PITR for Postgres, cross-region object replication, RTO/RPO: 30m/5m.

---

6) Data & Model Strategy

- **Model Mix**: API LLMs for quality + local small models for latency/privacy; dynamic routing by task.
- **RAG-first**: prioritize retrieval + tool-use over long-context brute force.
- **Evals**: automated win-rate dashboards per task (QA, coding, planning, safety).
- **Fine-tuning**: on curated, licensed, consented corpora; keep a clean audit trail.
- **Provenance**: stamp generated media; attach sources/links in answers.

---

7) Monetization & Pricing

- **Free**: daily cap, watermarking, community tools.
- **Pro**: higher limits, team spaces, advanced RAG, API credits.
- **Team/Business**: SSO, SCIM, audit logs, DLP, regional processing.
- **Sovereign**: owner-exclusive feature set, private endpoints, unlimited tools, root console, early model access.
- **Marketplace**: 70/30 rev share (creator/PRIMEX) initially; adjust post-PMF.

---

8) Go-To-Market System

- **Positioning**: “Your Private AI Ops System: fast on-device, powerful in cloud.”
- **Launch Stack**: waitlist → invite codes → phased store launch (soft regions) → global.
- **Channels**: App stores, web, creators, affiliates, B2B founder sales, developer API.
- **ASO**: localized listings, screenshot narratives, custom store listings per segment.
- **Influencers**: clear disclosures, approved claims, unique redemption flows.
- **Community**: prompts/agents gallery, weekly “Ops Runbooks,” revenue share highlights.

---

9) Roadmap (Quarter-by-Quarter)

Q0 (4–6 weeks): Foundation

- Brand/IP search, core UX wireframes, tech spike (model router + RAG), store compliance checklist, privacy policy/EULA, observability scaffold.

Q1: MVP

- Chat + RAG + voice, iOS/Android/Web beta, basic marketplace (prompt templates), subscriptions (Free/Pro), analytics/Growth v1, provenance v1.

Q2: Sovereign Mode + Automations

- Owner console, scripting + scheduled jobs, advanced connectors, team spaces, revenue share v1, creator portal.

Q3: Scale & Partner

- On-device LLM v2, enterprise controls (SSO, SCIM), co-marketing partners, merch drop, education content series.

---

10) Org & Roles

- **ARCHITECT**: system design, platform guardrails, DX.
- **CORTEX**: strategy, modeling of growth/monetization, risk sims.
- **CENTURION**: QA gates, AppSec enforcement, release approvals.
- **GHOSTLINE**: privacy playbooks, threat modeling, comms hygiene.
- **GOODJEW**: compliance templates, DPAs, ToS/EULA, policy mapping.
- **OVERSEER**: program mgmt, sprints, deadlines.
- **SCRIBE**: docs, changelogs, decisions, knowledge base.
- **MINT/VAULT**: pricing, finance stack, runway, vendor contracts.

Hiring (external): Full-stack (TypeScript), iOS, Android, ML engineer (RAG/evals), DevOps/SRE, Designer.

---

11) Budgets & Targets (first 2–3 quarters)

- **Core Cloud**: \$6–15k/mo pre-scale (mix of CPU + burst GPU), observability + CDNs + storage.
- **GPU Burst (experiments)**: \$5–20k/mo depending on model tests/fine-tunes.
- **Team**: 5–8 core builders.
- **KPIs**: D1/D7/D30 retention; ARPPU; cost per 1k messages; RAG source-citation rate; creator GMV; refund/churn.

---

12) Risk Register & Mitigations

- **Platform Rejection** → Pre-submission audits, UGC controls, staged rollouts, alt-distribution (EU).
- **Regulatory Shifts** → Configurable disclosures, provenance by default, policy versioning, legal reviews.
- **Copyright/Data** → Licensed/consented training data, no shadow scraping, opt-out honoring, provenance + source links.
- **Compute Costs** → Router cost caps, caching, hybrid search, batch jobs on spot.
- **Model Supply** → Multi-provider contracts, fallbacks, on-device micro-models.

---

13) App Store & Play Compliance Checklist (Shipping)

- Age ratings, data safety, account deletion, report/block, UGC EULA, harmful content filters, AI output reporting, privacy policy, parental gating (if minors).
- IAP receipts server-side validation; billing entitlements consistent across platforms.

14) Brand & Collateral

- **Name Lock**: PRIMEX SOVREIGN; classes 9/42 core; defensive registrations.
- **Visuals**: minimal neon accents, dark/light modes, motion micro-interactions.
- **Apparel/Tech**: limited drops aligned to milestones (e.g., “Ops Console” hoodie), NFC-tagged authenticity.

---

15) Appendices

- **Policy Mappings**: MASVS/ASVS → controls; privacy DSR flows; AI transparency copy blocks.
- **Data Schemas**: user/org/entitlement; conversation/thread/message; document chunk indexes; telemetry events.
- **Runbooks**: incident severity ladder; abuse response SLAs; model rollback; hotfix protocol.
- **Creator Program**: terms, payout schedules, acceptable use, review process.

Execution Notes

- Build the **Model Gateway + RAG** first; every feature composes on top.
- Keep **observability from day 1**; ship with evals; measure model win-rates weekly.
- **Sovereign Mode** is a superset; ship it in parallel with feature flags + owner policy.
- Land **trust features** (provenance, disclosures, DSRs) at MVP to de-risk GTM.