Simon AI

Profesyonel Proje Yönetimi - Genel Entegrasyon - Modül Mimarisi

Sürüm: v0.2 | Tarih: 24.12.2025

1. Yönetici Özeti

Simon AI; kullanıcıların farklı yapay zeka sağlayıcılarını tek bir kurumsal arayüzde, tek sohbet geçmişi ve tek proje yapısı altında kullanabildiği; hata/limit durumunda otomatik model geçişi (failover) yapabilen hibrit bir platformdur.

Key Modeli (Ürün Ekonomisi)

- FREE (Key yok): Kullanıcının kendi cihazında çalıştırdığı Ollama üzerinden açık kaynak modeller (yerel/kişisel kullanım).
- FREE+ (Sponsorlu): Simon AI sunucusunda yönetilen kısıtlı "server key havuzu" (kapalı modeller). Sert kota + rate limit + abuse kontrolü zorunlu.
- BYOK: Kullanıcı kendi API anahtarını girer. Kalite/limit yönetimi kullanıcıya aittir; platform yalnızca deneyim ve orkestrasyon sağlar.

MVP ilk sürüm hedefi (80/20)

- İlk sürümün odağı: Ultra hızlı web chat + Key Mode seçimi + model kataloğu + failover + temel telemetri.
- Sol/üst/sağ panellerin iskeleti (kurumsal UI), widget tarafı MVP'de minimum.
- Geri bildirim sistemi: toplama + otomatik triage raporu (admin onayıyla aksiyon).

2. Hedefler, Kapsam ve Başarı Kriterleri

Hedefler

- Kullanıcının 10 saniye içinde sohbet başlatabilmesi (ilk yükleme + ilk token).
- Tek arayüzden en popüler modeller: FREE'de en az 15 model (Ollama), BYOK'ta varsayılan 4 üst seviye model.
- Failover ile kesintisiz deneyim: sağlayıcı hatası/kota dolumu durumunda otomatik rota değişimi.
- Kurumsal, sade ve dikkat çeken UI: ChatGPT benzeri sol menü + üst bar + sağ widget paneli.

Kapsam

- Web (öncelik): Next.js tabanlı kurumsal chat arayüzü + admin alanı + widget paneli.
- Backend: FastAPI + AI Router (LiteLLM tabanlı) + veri katmanı + rate limit + audit.
- Mobil: MVP sonrası (Faz 4).

Başarı kriterleri (ilk 30 gün, MVP sonrası)

Kategori	KPI	Hedef
Performans	P95 yanıt gecikmesi (BYOK)	< 2.5 sn
Stabilite	P95 hata oranı (5xx/timeout)	< %1.0
Kullanım	Haftalık aktif kullanıcı (WAU)	Artan trend
Kalite	Geri bildirim memnuniyeti (CSAT/NPS)	>= 4.3/5

Operasyon	Geri bildirimden triage raporuna süre	< 10 dk (otomatik)

3. Paydaşlar ve Roller

- Kullanıcı: sohbet, projeler, widget alanı, BYOK key yönetimi.
- Admin (Ceyhun Bey): model kataloğu yönetimi, FREE+ key havuzu, abuse kontrolü, geri bildirim onayları, multi-model tartışma (debate), prompt-to-product stüdyosu.
- Sistem: telemetri, hata izleme, otomatik triage, güvenlik ve oran sınırlama (rate limit).

4. MVP Gereksinimleri (Fonksiyonel)

Chat ve Proje

- Sohbet: mesaj gönder/al, kopyala, düzenle, streaming yanıt (token token).
- Proje/Klasör: proje altında sohbetler; hızlı arama; pinleme.
- Model seçimi: Key Mode + model (15/4 listeleri) + failover politikası.
- Oturum: MVP'de zorunlu login yok (opsiyonel). Faz 3'te login & çoklu cihaz senkron.

Panel Yerleşimi (UI)

- Sol Sidebar: logo + chat/proje/klasör/araçlar/ayarlar/profil; daraltınca ikon moduna düşer.
- Üst Bar: sabitlenebilir/gizlenebilir ikon menüsü + hızlı model/Key Mode dropdown + mini dashboard (token, kelime, süre, toplam proje/sohbet).
- Sağ Panel: aç-kapa ikonları; seçilen uygulama için canlı widget alanı + sağ alt mini "slayt" (ör. haberler).

Geri Bildirim Sistemi (MVP)

- Kullanıcı: tek tuşla geri bildirim gönderir (ekran görüntüsü opsiyonel, cihaz/versiyon, son 50 log satırı, istek id).
- Backend: kaydı saklar; Triage Agent (admin BYOK anahtarıyla) problemi sınıflandırır: Problem - Olası Neden - Reprodüksiyon - Çözüm Önerisi - Risk.
- Admin: tek ekranda "Uygula mı?" onayı verir. MVP'de aksiyon: otomatik GitHub Issue açma; V1+'da PR üretimi + CI deploy.

Admin Özel Modüller (MVP'de iskelet)

- Debate Playground: 2-4 modeli aynı prompt ile çalıştır, yanıtları karşılaştır, "hakem" modeliyle sentez üret.
- Prompt-to-Product Studio: (V1+) prompt ile tasarım/uygulama/doküman üretim iş akışları (tek tuş).

5. NFR: Performans, Güvenlik, KVKK/GDPR ve Kullanılabilirlik

Performans ve deneyim

- Streaming: ilk token hedefi < 1.5 sn (BYOK, bölgesel ağ koşulları dahilinde).
- UI: kritik path'lerde SSR/partial hydration; minimal bundle; skeleton loading.
- Backend: istek başına timeout + retry/backoff; bağlantı havuzu; cache (model kataloğu, ayarlar).

Güvenlik

- API anahtarları: sunucuda şifreli saklama (KMS/Secrets) + maskeleme + audit log.
- Rate limit: IP + kullanıcı + key mode bazlı; FREE+ için agresif abuse koruması.
- CORS/CSP, XSS/CSRF korumaları; dosya upload antivirüs (V1+).

KVKK/GDPR ve veri minimizasyonu

- Varsayılan: minimum PII; sohbet geçmişi opsiyonel saklama (MVP'de lokal).
- Geri bildirimde PII redaksiyonu (otomatik maskeleme).
- Silme/anonimleştirme iş akışı (Faz 3-4).

6. Modül Mimarisi (Web, Backend, Admin, Servisler)

MVP'de minimum parça ile maksimum değer hedeflenir. Tüm modüller tek repoda, net sınırlarla tutulur.

Web (Next.js)

- Chat UI: mesaj akışı, composer, model dropdown, streaming render.
- Layout: sol sidebar + üst bar + sağ widget paneli (responsive).
- Telemetry: performans ölçümü (web vitals), hata yakalama.

Backend (FastAPI)

- AI Router: LiteLLM üzerinden sağlayıcı/Key Mode seçimi ve failover.
- Key Manager: BYOK anahtarları + FREE+ pool yönetimi + policy.
- Usage & Limits: token sayımı, oturum süresi, kota.
- Feedback Service: kayıt + triage agent çıktısı + admin onayı.

Servisler

- DB (Faz 3): PostgreSQL/Supabase (kullanıcı, sohbet, projeler, feedback).
- Cache: Redis (opsiyonel) - düşük maliyet için başlangıçta in-memory + CDN cache.
- Observability: Sentry/OTel + basit log toplama.

7. Genel Entegrasyon ve Veri Akışları

Akış A - Chat isteği

- Web -> Backend: mesaj + proje bağlamı + Key Mode + model tercihi + istek id.
- Backend: policy kontrol (rate limit, key var mı?) -> LiteLLM çağrısı -> streaming yanıt.
- Failover: hata/limit -> sıradaki uygun modele otomatik geçiş -> kullanıcıya "model değişti" etiketi.

Akış B - Geri bildirim ve otomatik triage

- Web: geri bildirim formu + otomatik diagnostik paket.
- Backend: kayıt + triage agent (LLM) -> rapor üretimi -> admin ekranına düşer.
- Admin: onay -> GitHub Issue (MVP) / PR+Deploy (V1+).

Akış C - Debate Playground

- Admin prompt -> seçili 2-4 model paralel/seri çalışır.

- Yanıtlar karşılaştırılır; hakem modeli “sentez” üretir; çıktı proje notlarına eklenir.

8. Veri Modeli Özeti (MVP)

MVP’de login zorunlu olmadığı için veri modeli iki katmanlı önerilir: (1) tarayıcı-local (IndexedDB) (2) opsiyonel sunucu DB (Faz 3).

Tablo/Koleksiyon	Amaç	Kritik alanlar
users (Faz 3)	Kullanıcı hesabı ve rol	id, email, role, created_at
api_keys (Faz 3)	BYOK anahtar depolama	user_id, provider, key_enc, created_at, last_used_at
projects	Proje/kategori	id, owner_id(optional), name, pinned
chats	Sohbet başlığı ve meta	id, project_id, title, model_mode, created_at
messages	Mesajlar	chat_id, role, content, token_usage, created_at
model_catalog	Model listesi	provider, model_id, key_mode, enabled, priority
feedback_reports	Geri bildirim	id, user_ctx, logs, triage_report, status, created_at
usage_events	Telemetri	timestamp, request_id, latency_ms, tokens_in/out, error_code

9. UI/UX Tasarım Sistemi ve Yerleşim (Sol - Üst - Sağ Paneller)

Tasarım prensipleri

- Ultra hızlı ve sade: tek ana CTA, minimal animasyon, yüksek kontrast (dark mode).
- Kurumsal görünüm: tipografi hiyerarşisi, tutarlı boşluk, keskin olmayan köşeler, soft shadow.
- Dikkat çeken ama kontrollü renk: tek aksan renk + durum renkleri (success/warn/error).

Yerleşim (önerilen)

[Sol Sidebar] [Chat Alanı (Mesajlar)] [Sağ Panel]
[Logo + Nav] [Üst Bar: Mode/Model + Dashboard + Kısayollar] [Widgets]
[Proje/Chat] [Mesaj balonları] [Mini-slayt]
[Ayar/Profil] [Composer + araçlar] [(sağ alt)]

Sağ panel widget mantığı (MVP -> V1)

- MVP: seçilebilir “widget kartları” (haber akışı, sosyal, skor vb.) - güvenli embed (allowlist).
- V1: seçili kaynağın içeriklerini backend’den çekip “mini slayt” olarak döndürme (otomatik geçiş).

10. Yetkilendirme ve Güvenlik (Key Yönetimi dahil)

Rol modeli

- guest: login yok, lokal sohbet, FREE (Ollama) ve sınırlı FREE+.
- user: BYOK anahtar yönetimi, sohbet senkron (Faz 3).
- admin: model kataloğu, key pool, feedback onayı, debate, otomasyon stüdyosu.

Key yönetimi - minimum güvenlik standardı

- Anahtarları UI'da asla tam göstermeme (mask).
- Sunucuda şifreli saklama + rotate desteği.
- BYOK çağrılarında anahtar sızıntısını önlemek için: sadece server-to-provider; tarayıcıdan direkt çağrı yok.

11. DevOps, Sürümleme, Otomasyon, Yedekleme ve Rollback

Tek repo (single-repo) disiplini

- /apps/web (Next.js), /apps/api (FastAPI), /apps/admin (opsiyon), /packages/shared, /infra (docker, scripts, docs).
- Branch: main (stabil) + develop; release tag: v0.x (rollback için).

Tek tuş kurulum hedefi

- Docker Compose ile: web + api + (opsiyonel) db + redis tek komutla kalkar.
- CI: lint/test/build; release artifact ve changelog.

Yedekleme ve geri dönüş

- Her büyük değişiklik öncesi: tag + yedek branch.
- DB (Faz 3): günlük otomatik yedek + aylık export prosedürü.

12. Faz Planı (Faz 0-4) ve Teslimatlar

Hedef: 10 saat içinde çalışan MVP. Sonrasında iteratif iyileştirme.

Faz	Net süre (hedef)	Teslimatlar (çıktı)
Faz 0 - Repo & Otomasyon	1.5 saat	Tek repo normalize + docker compose + CI iskeleti + sürüm tag/rollback
Faz 1 - Chat MVP	3.0 saat	Web chat UI + Key Mode seçimi + AI Router + streaming + temel telemetri
Faz 2 - Panel/UI Pro	2.0 saat	Sol/üst/sağ panel iskeleti + hızlı arama + tema/ayarlar
Faz 3 - Login & DB	3.0 saat	Kullanıcı yönetimi + BYOK key kasası + sohbet senkron + kullanım

		kotaları
Faz 4 - Mobil	4.0+ saat	Flutter uygulama iskeleti + push + mağaza hazırlığı

13. Maliyet Analizi (40 USD hedef - Low Cost / Premium)

Önerilen düşük maliyet (hedef: ~40 USD/ay)

- Web hosting: Vercel/Cloudflare Pages (ücretsiz/low tier).
- API hosting: tek küçük VM veya PaaS (Render/Railway/Fly.io) - düşük trafik için.
- DB (Faz 3): Supabase/PostgreSQL düşük plan.
- Observability: ücretsiz katman + örnekleme.
- FREE+ key pool: sponsor bütçesi yoksa çok sınırlı açılır (kota şart).

Premium seçenek (ölçek)

- Managed Postgres + Redis + ayrı API worker'lar.
- CDN + WAF + global edge caching.
- Ücretli APM + uzun süreli log saklama.

14. KPI, Ölçümleme ve Raporlama

- Ürün: WAU/MAU, proje başına sohbet sayısı, tekrar kullanım (retention).
- Teknik: p50/p95 latency, timeout oranı, failover oranı, model bazlı hata.
- Ekonomi: FREE+ harcama, BYOK oranı, sponsor gösterimi.
- Destek: feedback sayısı, triage doğruluğu, çözüm süresi (MTTR).

15. Riskler, Varsayımlar ve Azaltım Planları

- Maliyet patlaması (FREE+): kota + rate limit + abuse koruması + sponsor/gelir olmadan genişletmeme.
- Yerel Ollama gerçeği: FREE modelin global sunucu maliyeti yüksek. Çözüm: FREE'yi "kullanıcının cihazı" olarak konumlandırma.
- API sağlayıcı şartları: yalnız resmi API; scraping yok; açık ve şeffaf kullanım.
- Anahtar güvenliği: şifreli kasa + rotate + erişim logu.
- UI karmaşıklığı: panel mimarisi modüler; MVP'de iskelet, sonra iteratif.

16. İleri Öneriler ve Roadmap (V1+)

- RAG (doküman sohbeti) + proje klasörlerinden bağlam (chunking + embeddings).
- Araçlar: dosya işlemleri, web tarama, takvim/e-posta bağlayıcıları (izin tabanlı).
- Marketplace: widget ve araç ekosistemi (allowlist + güvenlik).
- Prompt-to-Product Studio: tasarım -> kod -> PR -> deploy tek tuş (admin onaylı).
- Mobil: Flutter ile tek kod tabanı + push + offline cache.

17. Ekler

A) Varsayılan model kataloğu

Not: Liste sistemde konfigüre edilebilir; aşağıdaki öneriler başlangıç setidir.

FREE (Ollama - en az 15 model)

- gemma3 (Genel)
- qwen2.5 (Genel)
- qwen2.5-coder (Kod)
- qwen3 (Genel)
- phi4 (Genel)
- phi4-mini (Hızlı)
- llama3.3 (Genel)
- llama4:scout (Üst seviye)
- mistral (Genel)
- deepseek-r1 (Akıl yürütme)
- qwq (Akıl yürütme)
- llava (Görsel)
- moondream (Hafif görsel)
- codellama (Kod)
- granite3.3 (Kurumsal)

BYOK (varsayılan en iyi 4)

- OpenAI GPT-5.2 (genel amaç / kod / ajan)
- Anthropic Claude Opus 4.5 (ajan / kod / uzun görev)
- Google Gemini 3 Pro (multimodal / 1M context)
- xAI Grok 4.1 Thinking (akıl yürütme / genel)

B) Failover politikası (öneri)

- Öncelik: Kullanıcının seçtiği model -> aynı sağlayıcı küçük model -> alternatif sağlayıcı eşdeğer.
- Maks 2 failover denemesi (sonsuz döngü yok).
- Kullanıcıya etiket: “Model değiştirildi (Sebep: timeout/limit)”.

C) Geri bildirim süreci - admin onay kapısı

- 1) Topla: form + diagnostik paket
- 2) Analiz et: triage agent raporu
- 3) Onayla: admin “uygula”
- 4) Aksiyon: Issue/PR/Deploy
- 5) Kapat: kullanıcıya geri dönüş

Kaynaklar (seçilmiş)

OpenAI model listesi ve GPT-5.x rehberi; Google Gemini 3 geliştirici kılavuzu; xAI API sayfası; Ollama model kütüphanesi ve örnek modeller; Anthropic Claude Opus 4.5 duyurusu.