Product Requirements Document (PRD)
Project Name: AI-Agent + AI-Design Web Application
Version: 2.1 (Production Ready with Enhancements)
Date: August 26, 2025

1. Executive Summary

This PRD defines all requirements for a production-ready, enterprise-grade AI web application that integrates AI task management with AI-powered design visualization (inspired by Gendo.ai functionality). The application must support secure, scalable, and compliant workflows for both general task automation and design-to-image rendering.
Key Objectives:

Provide robust task orchestration and AI agent capabilities.
Deliver professional-grade AI visualization tools (design-to-CGI, editing, style transfer, text-to-image).
Enforce compliance (GDPR, SOC2, HIPAA-ready) and enterprise controls.
Scale to thousands of users with credit-based billing and multi-tier subscriptions.
Maintain observability, monitoring, and DevOps readiness.
AI Model Customization: Enable users to fine-tune AI models for personalized outputs.
Sustainability Metrics: Track and report AI operational environmental impact.
Hybrid AI-Human Workflows: Support real-time human review of AI suggestions.

Implementation Details for New Objectives:

AI Model Customization: Implement via a low-code interface using transfer learning APIs from libraries like Hugging Face Transformers. Backend will expose endpoints for uploading datasets, triggering fine-tuning jobs on GPU clusters, and storing custom models in user-specific S3 buckets. Ensure versioning and rollback via Git-like model registry.
Sustainability Metrics: Integrate CodeCarbon library in ML pipelines to calculate CO2 emissions per task/render. Store metrics in PostgreSQL and expose via API for dashboard visualization. Implementation includes periodic reporting jobs using cron-like schedulers in Kubernetes.
Hybrid AI-Human Workflows: Use WebSockets for real-time collaboration; backend queues human approval requests via Redis, with frontend polling for updates. Implement approval flows as state machines in the task orchestration engine.

2. Scope & Stakeholders

In-Scope:

Web application with modular workspaces: Task Manager & Design Studio.
APIs and SDKs for task automation and AI rendering.
Subscription & billing with credit system.
Full security, compliance, and privacy features.
Multi-language and accessibility compliance.
Third-Party Integrations: Connectors to external tools for workflow enhancement.
Progressive Web App (PWA): For mobile-first access.

Out-of-Scope: Mobile-native apps (though PWA provides similar functionality); advanced metaverse integrations (deferred to Phase 4).
Stakeholders: Product, Engineering, Security, Compliance, QA, Marketing, End Users (Designers, Analysts, Teams, Enterprises), AI Ethics Review Board (for bias audits), Customer Success Teams (for enterprise onboarding).
Implementation Details for New Scope Items:

Third-Party Integrations: Use OAuth2 for authentication with services like Slack (webhooks for notifications), Jira (API sync for tasks), Adobe/Figma (import/export via SDKs). Backend microservices will handle API calls, with rate limiting via Redis.
Progressive Web App (PWA): Implement using React's service workers for offline capabilities. Manifest.json for installability, with push notifications via Firebase Cloud Messaging.

3. User Personas & Journeys
Personas:

Analyst – creates, monitors, and reports tasks.
Designer – uploads designs, generates images, edits materials.
Compliance Officer – audits logs and ensures legal compliance.
Administrator – manages roles, credits, billing.
Enterprise Buyer – oversees organization-wide usage.
Developer – builds custom agents via SDK.
Freelancer – needs quick, low-cost bursts with flexible credits.

Core Journeys:

Sign-up & Onboarding: Secure account creation, MFA, tutorial.
Task Management: Define AI workflows, monitor execution.
Design-to-Image: Upload sketch → CGI render.
Generative Editing: Edit parts of an image using segmentation.
Text-to-Image: Generate new visuals from prompts.
Style Transfer: Convert images to sketch, watercolor, CGI, etc.
Audit Review: Access logs, filter by user/date.
Subscription Management: Track credits, buy packs, manage billing.
Collaboration Journey: Real-time co-editing of tasks or designs.
Feedback Loop Journey: Post-render surveys for iterative improvements.
Error Recovery Journey: Guided troubleshooting for failures.
Export/Sharing Journey: Secure sharing with expiration and watermarks.

Implementation Details for New Personas & Journeys:

New Personas: Extend RBAC to include Developer (API access tokens) and Freelancer (burst credit packs). Implementation via database schema updates in PostgreSQL for role-specific permissions.
Collaboration Journey: Use Socket.io for real-time updates; store session states in MongoDB. Frontend components in React with collaborative cursors.
Feedback Loop Journey: Integrate post-task forms with Net Promoter Score (NPS); aggregate data in Elasticsearch for AI retraining triggers.
Error Recovery Journey: Build a rule-based expert system in the backend (Python with Flask) to suggest fixes, logging errors to Sentry.
Export/Sharing Journey: Generate signed URLs via AWS S3 with TTL; add watermarking using Pillow library in image processing pipelines.

4. Functional Requirements
4.1 Authentication & Authorization

MFA, SSO (SAML, OIDC), RBAC.

Roles: Admin, Designer, Analyst, Compliance, Auditor.
Passwordless Login: Magic links, WebAuthn, passkeys.
Session Management: Auto-logout, device management, geo-fencing.
Federated Identity: Google Workspace, Microsoft Entra ID.

Implementation Details:

Passwordless Login: Use Auth0 or similar for magic links; WebAuthn via browser APIs. Backend verifies via JWT tokens.
Session Management: Store sessions in Redis with expiration; geo-fencing via IP lookup services like MaxMind.
Federated Identity: Configure OIDC providers in the auth microservice.

4.2 Task & AI-Agent Management

Task orchestration (scheduling, retry logic).
Parameterized pipelines with templates.
Logging, monitoring, and error alerts.
Agent Marketplace: Pre-built templates with ratings.
Version Control: Git-like branching for pipelines.
Dependency Management: Auto-resolve conflicts.
Analytics Dashboard: Visualizations for performance.

Implementation Details:

Agent Marketplace: MongoDB collection for templates; frontend marketplace UI with React and ratings via API endpoints.
Version Control: Integrate GitLab API or custom VCS in backend for pipeline YAML files.
Dependency Management: Use DAG (Directed Acyclic Graph) libraries like Airflow for orchestration and conflict detection.
Analytics Dashboard: Use Chart.js in frontend; backend aggregates data from Prometheus.

4.3 AI Design Studio Features

Design-to-Image Rendering: Convert 2D sketches/images to high-fidelity CGI.
Generative Editing: Smart segmentation for selective modifications.
Text-to-Image: Prompt-based image generation.
Style Transfer: Artistic transformations (sketch, watercolor, ink).
3D Model Support (Phase 2).
Output Management: Variants, high-res PNG, version history.
Batch Processing: Parallel rendering for multiple inputs.
Real-Time Preview: Low-res previews with adjustments.
Undo/Redo Stack: Multi-level history with branching.
Advanced Tools: Inpainting/outpainting, upscaling, material libraries.
AR/VR Integration: WebXR previews.
Prompt Engineering Aids: Auto-suggestions.

Implementation Details:

Batch Processing: Queue jobs in RabbitMQ; process in parallel on Kubernetes pods with GPU sharing.
Real-Time Preview: Use progressive rendering with WebSockets; low-res via downsampled diffusion models.
Undo/Redo Stack: Store states as JSON in Redis; implement as a tree structure for branching.

Advanced Tools: Inpainting via Stable Diffusion models; upscaling with ESRGAN; material libraries as S3-hosted assets with metadata in DB.
AR/VR Integration: Use Three.js for WebXR; render previews client-side where possible.
Prompt Engineering Aids: NLP-based suggestions using spaCy or LLMs fine-tuned on prompt datasets.

4.4 Subscription & Credits System

Plans: Free, Individual, Professional, Team, Enterprise.
Credits: Monthly allocation, one-off packs, rollover for purchased packs.
Licensing Rules:

Free/Individual: Attribution required.
Professional/Team/Enterprise: Commercial rights, NDA-friendly.

Billing Dashboard: Upgrade/downgrade, invoices, usage reporting.
Dynamic Credit Allocation: AI-predicted usage suggestions.
Referral Program: Credit bonuses for referrals.
Usage Analytics: Granular breakdowns with forecasting.
Promo Codes & Discounts: Campaign support.

Implementation Details:

Dynamic Credit Allocation: ML model (scikit-learn) on historical data; API endpoint for predictions.
Referral Program: Track via unique codes in DB; credit awards via Stripe webhooks.
Usage Analytics: Query logs from ELK; forecasting with Prophet library.
Promo Codes: Store in PostgreSQL; validate in billing microservice.

4.5 Notifications & Alerts

Email/SMS/in-app alerts for task completion, credit usage, billing.
Webhooks for enterprise integrations.
Customizable Channels: User preferences.
Proactive Alerts: AI-driven warnings.
Integration Hooks: No-code triggers.

Implementation Details:

Customizable Channels: User settings in DB; dispatch via Twilio for SMS, SendGrid for email.
Proactive Alerts: Rules engine with ML anomaly detection (Isolation Forest).
Integration Hooks: Use Zapier-compatible webhooks.

4.6 Localization & Accessibility

i18n with multiple languages.
WCAG 2.1 AA compliance.
RTL language support.
Auto-Language Detection: Browser-based switch.
Voice Commands: Speech-to-text integration.
Custom Themes: Accessibility modes.

Implementation Details:

Auto-Language Detection: Use navigator.language in JS; fallback to user prefs.
Voice Commands: Web Speech API for input; process via backend NLP.
Custom Themes: CSS variables in Tailwind; toggle via user settings.

5. Non-Functional Requirements

CategoryRequirementPerformanceImage render < 30s; 99th percentile task latency < 500ms; edge caching < 100ms latencyScalability10,000+ concurrent users; burst compute autoscaling; petabyte-scale storageAvailability99.9% uptime SLA; multi-region failoverReliabilityFull rollback & canary deploys; agent drift detection with auto-retrainingSecurityOWASP Top 10 mitigations; SAST/DAST in pipeline; zero-trust; RASPComplianceGDPR, SOC2 Type II, HIPAA-ready (Enterprise tier); CCPA/CPRA; automated consentMaintainability90% unit test coverage, modular APIs, semantic versioning; auto-generated docs
Implementation Details for Updates:

Performance: CDN via Cloudflare for assets; GPU acceleration with CUDA.
Scalability: Kubernetes with HPA; S3 for storage.
Availability: Active-active setup with AWS Global Accelerator.
Reliability: Drift detection via model monitoring tools like Evidently AI.
Security: Implement zero-trust with service mesh (Istio); RASP via agents in containers.
Compliance: Consent via cookie banners; automated via OneTrust integration.
Maintainability: JSDoc for docs; coverage with Jest.

6. Security & Compliance

Encryption: TLS 1.3 in transit, AES-256 at rest.
Privacy Workflows: Right to be Forgotten, Data Access/Export.
Key Management: KMS/Vault, quarterly key rotation.
Vendor Audits: SOC2, DPAs, annual reviews.
Breach Policy: 72-hour disclosure, audit trail.
Threat Modeling: Continuous pentesting, bug bounty.
Data Residency: User-selectable regions.
AI-Specific Security: Prompt injection defenses, model watermarking.
Incident Response: Quarterly drills, SIEM integration.

Implementation Details:

Threat Modeling: Use STRIDE methodology; bug bounty via HackerOne.
Data Residency: Multi-region DB replication in AWS.
AI-Specific Security: Sanitize inputs with regex/LLM guards; watermark via invisible embeddings.
Incident Response: Splunk for SIEM; runbooks in Confluence.

7. Architecture & Infrastructure

Frontend: React + TypeScript, Tailwind, i18n.
Backend: Node.js/Python microservices, REST/gRPC APIs.
Data Layer: PostgreSQL (core), MongoDB (logs), S3 (assets).
ML/AI Layer: LLMs & diffusion models for tasks and design.
IaC: Terraform, ArgoCD, Helm.

CI/CD: GitHub Actions → ArgoCD (Blue/Green, Canary).
Serverless Components: For bursty tasks.
Edge Computing: ML inference at edge.
Data Pipeline: Real-time streaming.
Backup Strategy: Daily with recovery.

Implementation Details:

Serverless: AWS Lambda for non-GPU tasks; invoke via API Gateway.
Edge Computing: Cloudflare Workers for lightweight inference.
Data Pipeline: Kafka for events.
Backup: AWS Backup service with PITR.

8. Monitoring & Observability

Metrics: Prometheus, Grafana dashboards.
Logs: Centralized ELK with JSON structure.
Tracing: OpenTelemetry.
Alerts: Alertmanager with on-call escalation.
AI-Powered Insights: Anomaly detection.
User Behavior Analytics: Engagement tracking.
SLO/SLI Definitions: Tied to alerts.

Implementation Details:

AI-Powered Insights: Use TensorFlow for anomaly models on metrics.
User Behavior Analytics: Mixpanel integration.
SLO/SLI: Define in code with Checkly.

9. QA & Testing Strategy

Unit, Integration, E2E (Cypress/Playwright).
Adversarial prompt testing.
Accessibility audits (axe-core).
Load/stress tests (k6).
Security tests (OWASP ZAP).
Chaos Engineering: Failure simulations.
User Beta Program: Feedback loops.
AI Testing: Hallucination/bias checks.

Implementation Details:

Chaos Engineering: Gremlin for tests.
User Beta: LaunchDarkly for feature flags.
AI Testing: Use datasets for bias (Fairlearn); hallucination via custom eval scripts.

10. Documentation & Support

User Docs: Task management, design studio, billing.
API Docs: OpenAPI spec, Swagger UI with code samples.

Runbooks: Incidents, rollback, migrations.
Knowledge Base: Tutorials, troubleshooting, community forum.
Interactive Tutorials: Embeddable demos.
AI Chat Support: In-app bot.
Community Features: Forums with upvoting.

Implementation Details:

Interactive Tutorials: Use Shepherd.js for tours.
AI Chat Support: Integrate with Grok-like LLM via API.
Community Features: Discourse forum integration.

11. Pricing Strategy

Free: Starter credits, attribution required.
Individual: More credits, attribution required.
Professional: Commercial license, no attribution, NDA support.
Team: Shared credit pool, org billing.
Enterprise: Custom SLA, compliance, private infrastructure.
One-off Packs: Available for spikes, rollover credits.
Billing Dashboard: Invoices, history, upgrade/downgrade.
Tiered Add-Ons: Extras like priority queues.
Usage-Based Billing: Pay-per-render hybrid.
Transparency Tools: Credit burn calculator.

Implementation Details:

Tiered Add-Ons: Stripe products for add-ons.
Usage-Based: Metered billing via Stripe Usage API.
Transparency Tools: JS calculator in dashboard.

12. Roadmap

PhaseFeaturesTimelinePhase 0A/B testing wireframes; soft launchQ3 2025Phase 1Core tasks, Design-to-Image, Credits, APIQ3 2025Phase 23D support, Team tier, Attribution controlQ1 2026Phase 3Enterprise infra, advanced editing (landscaping, props, upscaling)Q3 2026Phase 4API marketplace, metaverse integrationsQ1 2027
Milestone Metrics: 50% user retention, 95% satisfaction.

13. Approval & Next Steps

Sign-off by Product, Security, Engineering, Compliance.
Finalize wireframes, OpenAPI spec, billing workflows.
Sprint 0: Env setup, CI/CD, auth module.
Phase 1 MVP release → Iterative rollout.