Product Design Document: AI PII Sanitizer Chrome Extension

Executive Summary

Product Name: AI PII Sanitizer (working title: "PrivatePrompt", "AliasAI", or "SafeChat")

Core Value Proposition: Use AI assistants (ChatGPT, Claude, Gemini, etc.) with real personal information context without exposing your actual PII to AI providers.

How It Works: Browser extension intercepts outgoing prompts to AI services, replaces real PII with contextually-appropriate aliases, then reverses the substitution in responses so users see their real data.

Primary Use Case: Privacy-conscious individuals and professionals in regulated industries (healthcare, legal, HR) who want AI assistance without compromising sensitive information.

Problem Statement

Current Pain Points

1. **Privacy Paralysis**: Users avoid using AI for truly useful tasks because they contain personal information
    - "Help me draft an email to my doctor about my symptoms" → Can't use real doctor's name
    - "Analyze this performance review for Sarah Chen" → Afraid to use actual names
    - "Plan a surprise party for my wife Emma" → Don't want this in AI training data
2. **Manual Redaction is Tedious**:
    - Replacing names with "[PERSON_1]" breaks AI context and reasoning
    - Users forget to redact everything, leading to partial exposure
    - Copy-paste-edit workflow interrupts creative flow
3. **Enterprise Cannot Adopt AI**:
    - HIPAA, GDPR, attorney-client privilege prevent AI use with real data
    - Companies ban ChatGPT rather than risk leaks
    - Massive productivity opportunity locked behind compliance walls
4. **Existing Solutions Fall Short**:
    - **Corporate proxies**: Expensive, complex to set up, IT-managed only
    - **PII detection tools**: Only flag issues, don't solve them
    - **Local AI models**: Less capable, require technical setup, expensive hardware

---

Product Vision

Target Users (Priority Order)

Phase 1: Privacy-Conscious Consumers

- Tech-savvy individuals who understand AI risks
- People dealing with sensitive personal situations (medical issues, legal problems, family matters)
- Willingness to configure aliases manually for key relationships

Phase 2: Regulated Professionals

- Healthcare workers (nurses, therapists, administrators)
- Legal professionals (paralegals, contract reviewers)
- HR professionals, recruiters
- Financial advisors

Phase 3: Enterprise Teams

- Small to medium businesses needing AI tools but lacking enterprise AI solutions
- Compliance-conscious organizations in healthcare, finance, legal

Success Metrics

MVP Success Criteria:

- 1,000 active users within 3 months
- 70%+ retention after first week
- 50%+ of users configure 3+ aliases
- <5% error rate in alias substitution
- Net Promoter Score >40

Growth Indicators:

- Average session length >10 minutes
- Users process 20+ prompts per week
- Organic sharing (viral coefficient >0.3)
- Enterprise inquiries (evidence of B2B demand)

Core Features (MVP)

1. Alias Management

User Interface: Simple sidebar panel or popup with alias dictionary

Capabilities:

- Add name mappings: "Joe Smith" → "John Doe"
- Auto-suggest contextually-appropriate aliases (matching gender, ethnicity, length)
- Visual feedback when aliases are active
- Quick enable/disable toggle per conversation
- Export/import alias dictionary

Smart Features (Post-MVP):

- Detect relationships automatically ("my wife Sarah" → preserve relationship in alias)
- Suggest aliases based on usage patterns
- Auto-generate temporary aliases for one-off names

2. Request Interception

Technical Approach: Service Worker intercepts fetch/XHR to AI domains

Supported Services (MVP):

- ChatGPT (chat.openai.com)
- Claude (claude.ai)
- Gemini (gemini.google.com)

Process Flow:

1. User types prompt containing "Joe Smith"
2. Extension detects outgoing POST request
3. Parse request body, identify PII from alias dictionary
4. Replace "Joe Smith" → "John Doe" in request
5. Forward modified request to AI service
6. AI processes prompt with alias

3. Response Reversal

Process Flow:

1. AI responds with "John Doe" in completion
2. Extension intercepts response before rendering
3. Replace all instances of "John Doe" → "Joe Smith"
4. Render response with real names to user

Edge Cases:

- Partial matches ("John" vs "John Doe")
- Possessives ("John's wife" → "Joe's wife")
- Pronouns (maintain context: "he" still refers to Joe)
- New entities mentioned by AI (flag for user review)

4. Privacy Controls

Transparency Features:

- Visual indicator when sanitization is active (colored extension icon)
- "Show original" toggle to see what AI actually received
- Audit log of all substitutions made
- Clear warning when sending unprotected prompts

User Control:

- Whitelist mode: Only sanitize specific sites
- Per-conversation override: "Use real data for this chat"
- Emergency disable: One-click turn off for all sites

User Experience

First-Time Setup

1. Install extension from Chrome Web Store
2. Welcome screen explains concept with simple example
3. Guided setup: "Add your first alias"
    - Input: "Joe Smith" (your name)
    - Auto-suggest: "John Doe"
    - Preview: See how it works in sample prompt
4. Choose supported AI services to protect
5. Optional: Set up additional aliases (family, colleagues, places)

Daily Usage Flow

Scenario: User wants AI to help draft sensitive email

1. Opens Claude.ai, extension icon shows green (protection active)
2. Types: "Help me write an email to Dr. Sarah Chen about my recurring headaches"
3. Sees yellow highlight on "Sarah Chen" as it's typed (indicating it will be sanitized)
4. Clicks send
5. Extension briefly shows toast: "1 name sanitized"
6. Claude responds with alias, extension reverses it, user sees "Dr. Chen" in response
7. User copies response, all names are correctly original

Error Recovery

Unknown Name Encountered:

- AI response contains "Dr. Michael Roberts" (not in alias dictionary)
- Extension flags it with warning: "Unknown name detected in response"
- User options:
    - "Add to dictionary" (map to real name)
    - "Ignore for this conversation"
    - "It's fictional" (no action needed)

Technical Architecture (High-Level)

Extension Components

1. **Content Script**: Runs on AI chat pages, monitors input fields, highlights PII
2. **Service Worker**: Intercepts network requests/responses, performs substitution
3. **Popup UI**: Manages alias dictionary, shows stats, toggle protection
4. **Storage**: Chrome.storage.local for alias dictionary (encrypted)

Data Flow

```
User Input → Content Script (detect PII)
  → Service Worker (intercept POST)
  → Alias Engine (substitute)
  → AI Service

AI Response → Service Worker (intercept)
  → Alias Engine (reverse)
  → Content Script (render)
  → User sees original names
```

Key Technical Decisions

Why Chrome Extension over Proxy/VPN:

- Zero configuration for end users
- Works with any AI service (no API keys needed)
- No additional infrastructure costs
- Transparent to AI providers

Why Client-Side over Server-Side:

- No data leaves user's machine (besides what goes to AI)
- No hosting costs for MVP
- Better privacy (we never see user's PII)
- Faster (no round-trip to our servers)

Non-Goals (For MVP)

❌ **Image/Document Analysis**: Only text-based chat initially ❌ **Multi-Language Support**: English only for MVP ❌ **Team/Enterprise Features**: Single-user only ❌ **Phone Numbers, Addresses, SSNs**: Names only initially ❌ **Perfect Accuracy**: 90% accuracy acceptable if user can fix errors ❌ **Voice Input**: Text only

Risks & Mitigation

Technical Risks

Risk	Impact	Mitigation
AI mentions alias in unexpected context	Medium	Implement context-aware reversal, allow manual override
Chrome updates break interception	High	Monitor Chrome dev channels, maintain compatibility layer
Users forget which aliases map to whom	Medium	Visual "cheat sheet" in popup, search function
Performance overhead on large responses	Low	Optimize regex matching, use Web Workers for processing

Business Risks

Risk	Impact	Mitigation
AI providers block the extension	High	Use standard browser APIs, avoid detection; emphasize this helps their GDPR compliance
Privacy paradox: Users don't care enough	High	Target regulated industries first where compliance is required
Free product expectations	Medium	Open source core, charge for enterprise features later

Legal/Compliance Risks

Risk	Impact	Mitigation
Users assume 100% protection	High	Clear disclaimers: "Reduces risk, doesn't eliminate it"
Violates AI ToS	Medium	Review each service's ToS, nothing explicitly blocks this

---

Go-to-Market Strategy

Launch Plan

Phase 1: Developer Preview (Month 1)

- Release on GitHub as open source
- Post to HackerNews, r/programming, r/privacy
- Focus on feedback over growth
- Goal: 100 power users, identify bugs

Phase 2: Public Beta (Month 2-3)

- Submit to Chrome Web Store
- Write blog post: "How I Built a Privacy Layer for ChatGPT"
- Post on ProductHunt
- Reach out to privacy-focused tech press (EFF, PrivacyTools.io)
- Goal: 1,000 users

Phase 3: Regulated Industry Outreach (Month 4-6)

- Case study: "Healthcare Worker Uses AI Safely"
- Reach out to medical/legal associations
- Create enterprise version with team features
- Goal: 10,000 users, 5 enterprise trials

Pricing Strategy (Future)

Free Tier:

- Unlimited name aliases
- 3 supported AI services
- Basic protection

Pro Tier ($5-10/month):

- Unlimited alias types (addresses, phone numbers, emails)
- All AI services
- Team sharing (shared alias dictionary)
- Audit logs and compliance reports

Enterprise ($50-100/user/year):

- SSO integration
- Admin controls
- Compliance certifications
- Priority support

Success Criteria & Timeline

3-Month Milestones

Month 1: MVP Launch

- ✅ Name-only aliasing works reliably
- ✅ Supports ChatGPT, Claude, Gemini
- ✅ Open source on GitHub
- ✅ 100 active users

Month 2: Polish & Growth

- ✅ 90%+ substitution accuracy
- ✅ Onboarding flow tested with 10 users
- ✅ Chrome Web Store approval
- ✅ 1,000 users, 50% week-1 retention

Month 3: Enterprise Validation

- ✅ Case study from healthcare or legal user
- ✅ 5 enterprise inquiries
- ✅ Team features prototype
- ✅ 5,000 users

Decision Points

After Month 1:

- If <50 users OR high error rate → Pivot to document anonymization instead
- If good traction → Continue

After Month 3:

- If no enterprise interest → Keep as free open source project
- If 5+ enterprise leads → Build paid tier

Why This Will Work

Unique Advantages

1. **First-mover in specific niche**: General PII tools exist, but none target AI chat specifically with bidirectional aliasing
2. **Solves real pain**: Validated by enterprise AI bans and privacy discussions on every AI subreddit
3. **Easy adoption**: No API keys, no setup, just install and add aliases
4. **Imperfect beats nothing**: Even 90% protection is way better than raw-dogging PII into ChatGPT
5. **Network effects**: Users share alias dictionaries (anonymized) to help others
6. **B2B potential**: Once users love it, their employers will pay for team version

Why Now

- AI adoption exploding, but privacy concerns holding back use cases
- GDPR/CCPA making companies paranoid about PII
- No good solutions exist for consumer/SMB market (enterprise has Microsoft Copilot, etc.)
- Chrome extension ecosystem mature and trusted by users

Appendix: FAQ

Q: Won't AI providers just block this? A: We use standard browser APIs. Blocking would affect all extensions. More likely they'll appreciate our help with their GDPR compliance.

Q: What if the AI generates new information about the alias? A: We flag it for user review. If AI says "John Doe lives in Seattle," user maps that back to Joe if relevant.

Q: Can I use this for my whole company? A: Not in MVP. Enterprise features (shared dictionaries, admin controls) are post-MVP.

Q: Does this work with API access? A: No, extension intercepts browser requests only. API users need different solution.

Q: What about images or uploaded documents? A: Text only for MVP. Future: OCR + sanitization for images.

Q: Is my alias dictionary synced across devices? A: Not initially. Post-MVP: Optional cloud sync (encrypted).

Q: What if I forget to turn it on? A: Extension active by default. You explicitly disable it, not enable it.

Q: Can the extension see my passwords/credit cards? A: No. It only monitors chat interfaces, not all web traffic, and only processes text you explicitly send to AI services.

Document Version: 1.0
Last Updated: October 2025
Owner: [Your Name]
Status: Draft for Feedback