

Wellness Agent AI: HIPAA Data Flow Diagram

Complete PHI Flow Analysis for HIPAA Compliance

Status: COMPLIANCE READY

Version: 1.0.0

Date: 2025-01-XX

Pattern: DATA x FLOW x HIPAA x COMPLIANCE x ONE

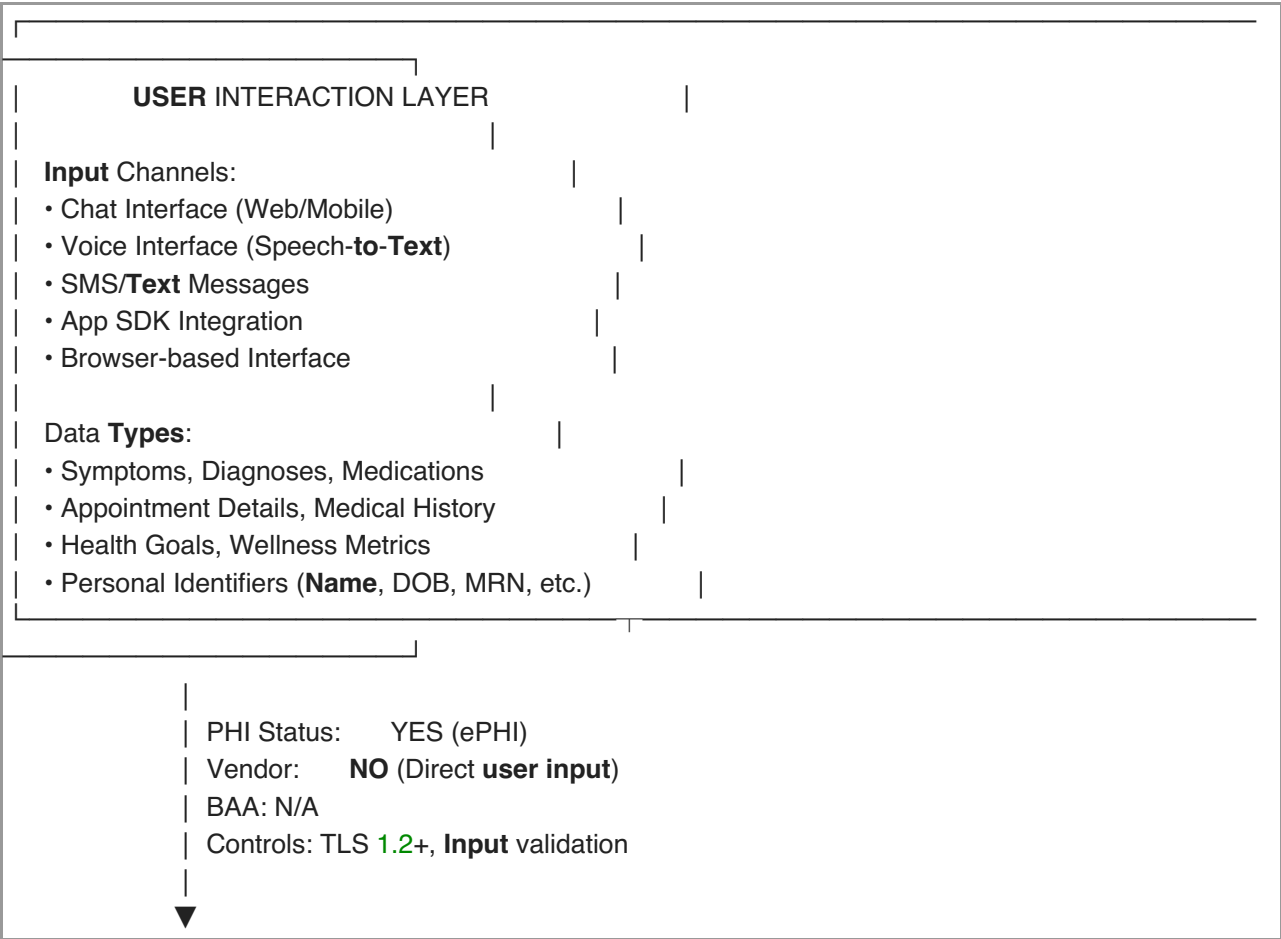
EXECUTIVE SUMMARY

This document provides a detailed data flow diagram showing where PHI flows through Wellness Agent AI, identifying:

- **PHI Status:** Whether data at each stage is PHI/ePHI
 - **Vendor Status:** Whether third-party vendors are involved
 - **BAA Status:** Whether Business Associate Agreements are required
 - **Controls:** Security and privacy controls applied at each stage
-

COMPLETE DATA FLOW ARCHITECTURE

Stage 1: User Interaction Layer



Stage 2: Ingress & Routing Layer

INGRESS & ROUTING LAYER

Components:

- API Gateway (AWS API Gateway / Kong / NGINX)
- Load Balancer (AWS ALB / CloudFlare)
- Authentication Service (Auth0 / AWS Cognito)
- Rate Limiting / DDoS Protection

Processing:

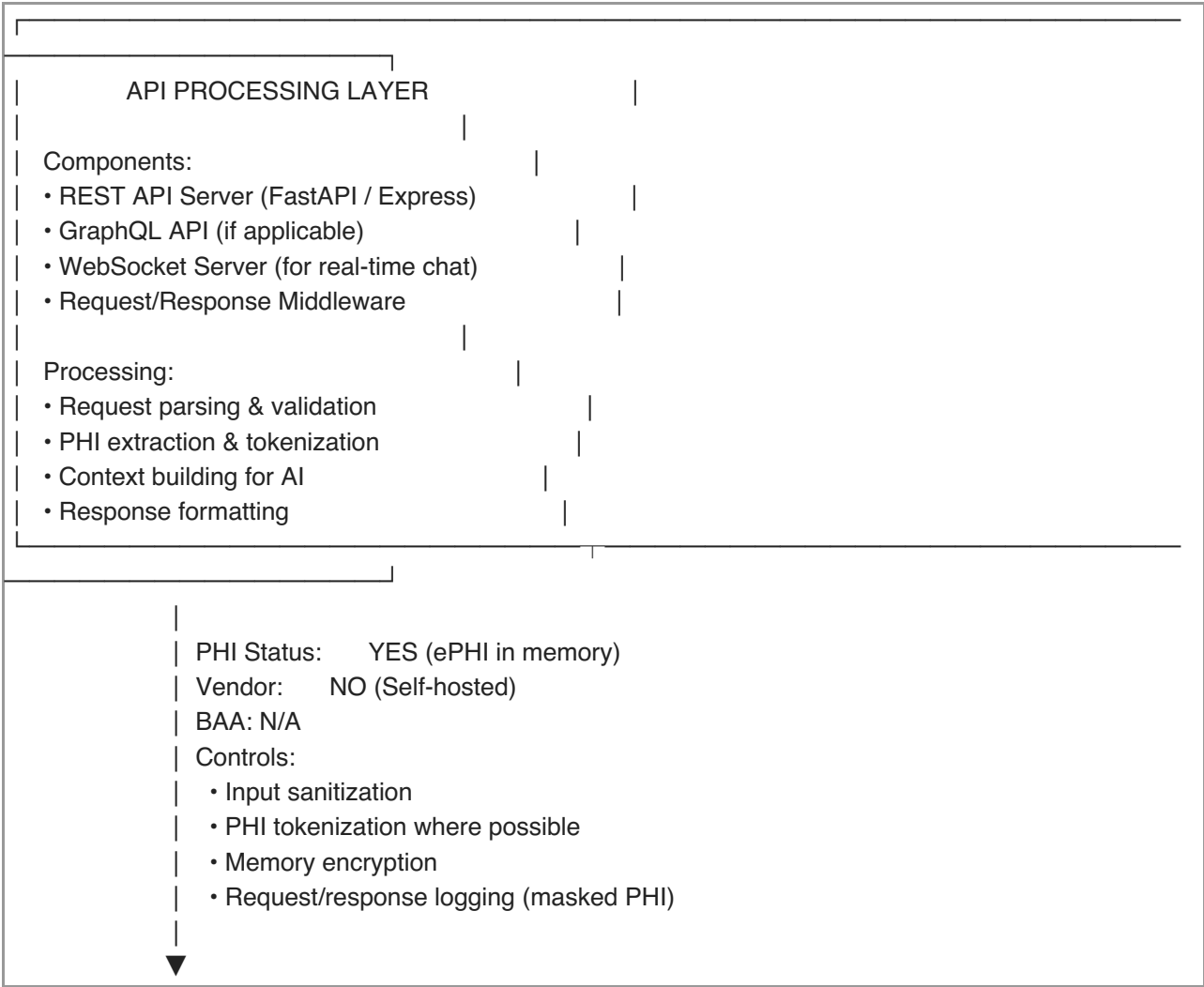
- Request routing
- Authentication & Authorization
- Request validation
- Session management

PHI Status: **YES** (ePHI in transit)
Vendor: **YES** (Cloud provider, Auth service)
BAA: **REQUIRED** (AWS/Azure/GCP, Auth0/Cognito)
Controls:

- TLS **1.2+** encryption in transit
- MFA for authentication
- RBAC for authorization
- Request logging (no PHI in logs)



Stage 3: API Processing Layer



Stage 4: AI Inference Layer

AI INFERENCE LAYER

Components:

- Prompt Engineering Service
- LLM API Client (OpenAI / Anthropic / Self-hosted)
- Vector Database (Pinecone / Weaviate / Self-hosted)
- RAG (Retrieval-Augmented Generation) Engine
- Speech-to-Text Service (AWS Transcribe / etc.)
- Text-to-Speech Service (AWS Polly / etc.)

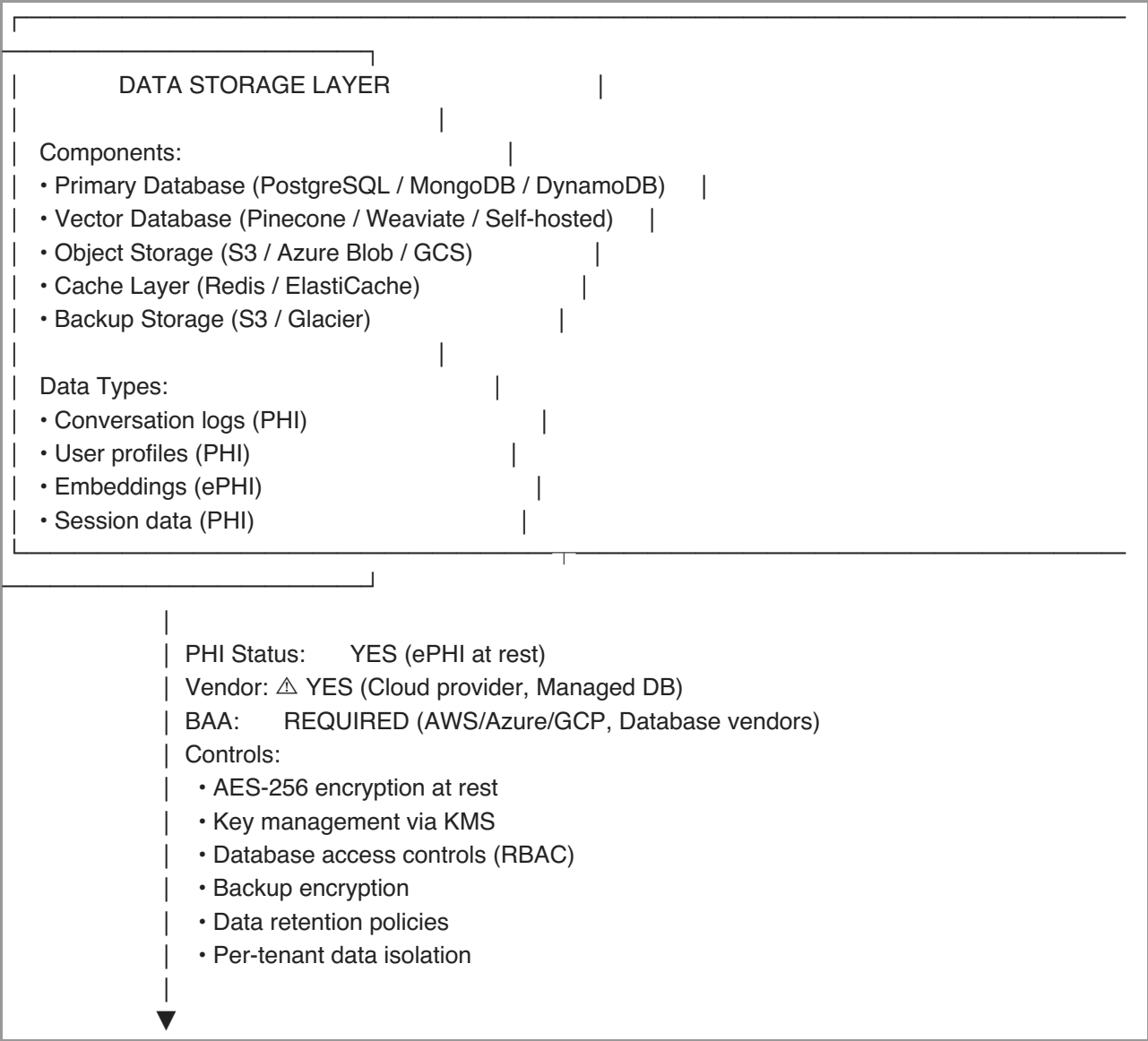
Processing:

- Prompt building (minimum necessary PHI)
- Context retrieval from vector DB
- LLM inference
- Response generation & validation
- Embedding generation (from PHI → ePHI)

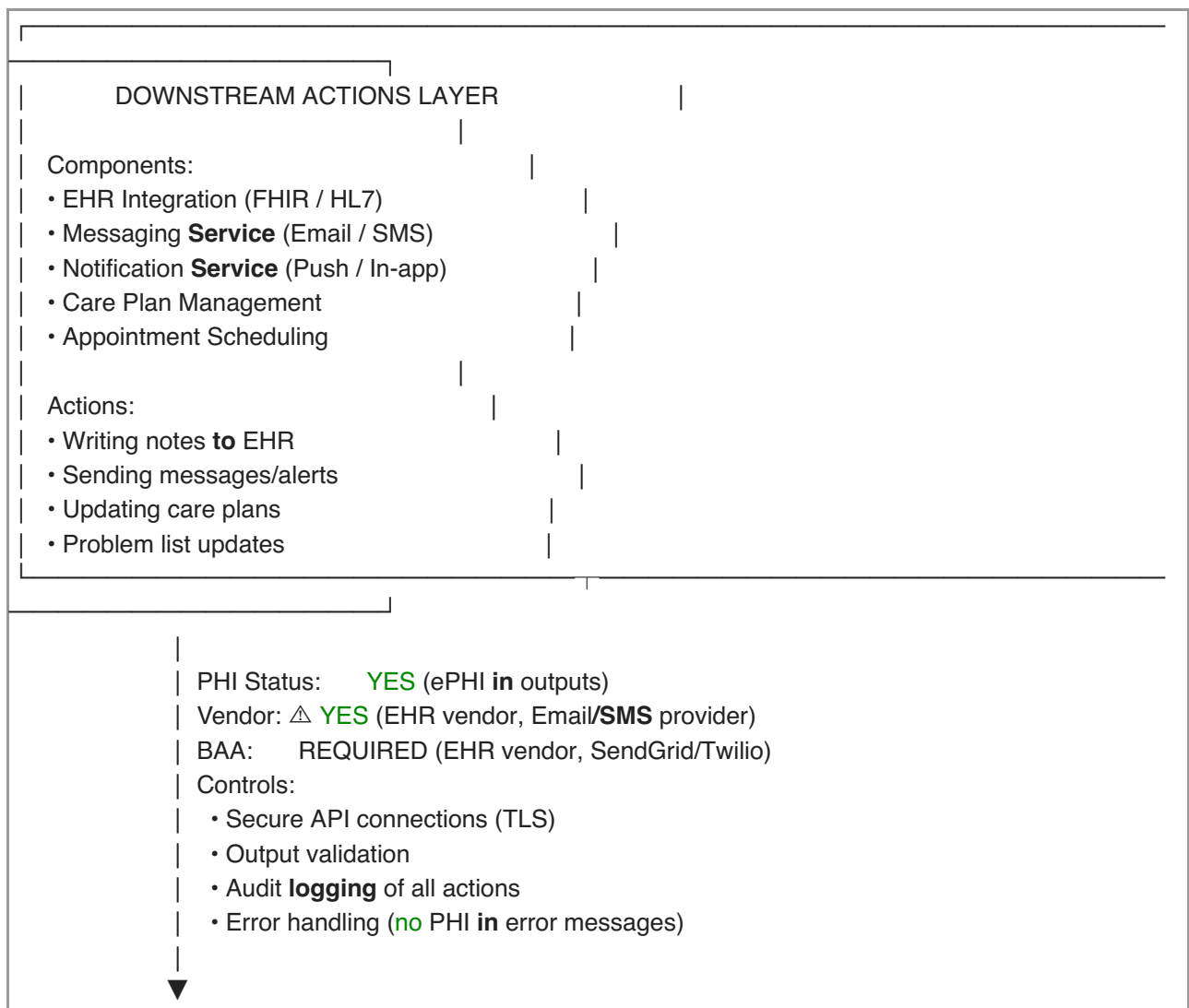
- PHI Status: YES (ePHI in prompts/embeddings)
- Vendor: Δ YES (LLM API, Vector DB, Speech services)
- BAA: REQUIRED (OpenAI/Anthropic, Pinecone, AWS)
- Controls:
 - Minimum necessary PHI in prompts
 - Pseudonymization where possible
 - No training on PHI (contractual)
- Embeddings treated as ePHI
 - Per-tenant vector indexes
 - Audit logging of **all** API calls



Stage 5: Data Storage Layer



Stage 6: Downstream Actions Layer



Stage 7: Logging & Analytics Layer

LOGGING & ANALYTICS LAYER

Components:

- Application Logs (CloudWatch / Datadog / Splunk)
- Metrics Collection (Prometheus / CloudWatch)
- Error **Tracking** (Sentry / Rollbar)
- Analytics Platform (Self-hosted / HIPAA-eligible)
- A/B Testing Platform

Data Types:

- Conversation logs (PHI)
- Performance metrics (non-PHI)
- Error traces (may contain PHI)
- **User** interaction analytics (pseudonymized)

PHI Status: Δ **CONDITIONAL**

- Logs with PHI → **YES**
- Pseudonymized analytics → **NO**

Vendor: Δ **YES** (Logging platforms)

BAA: **REQUIRED (if PHI in logs)**

Controls:

- Mask/strip PHI **in** logs where possible
- Use HIPAA-eligible analytics with BAAs
- Pseudonymize identifiers
- **No** PHI **to** ad networks
- Separate PHI logs **from** analytics



Stage 8: Model Training / Fine-Tuning Layer

MODEL TRAINING / FINE-TUNING LAYER

Components:

- Training Pipeline
- **De**-identification Service
- Model Training Infrastructure (GPU clusters)
- Evaluation Datasets
- Model Registry

Processing:

- PHI extraction from production
- **De**-identification (Safe Harbor / Expert Determination)
- Dataset preparation
- Model training/fine-tuning
- Model evaluation

PHI Status: **NO** (after **de**-identification)

- Before **de**-ID → YES
- After **de**-ID → **NO** (not PHI)

Vendor: △ YES (GPU provider, Training platform)

BAA: NOT REQUIRED (if properly **de**-identified)

Controls:

- **Separate de**-identified environment
- Strict **de**-identification process
- Tokenized linkages (if needed, with strict controls)
- **No** re-identification risk
- Expert certification (if Expert Determination)



Stage 9: De-Identified Data Storage

DE-IDENTIFIED DATA STORAGE	
Components:	
• De-identified Database (separate environment)	
• Analytics Warehouse	
• Model Training Datasets	
• Research Datasets	
Data Types:	
• De-identified conversations	
• Aggregated metrics	
• Training datasets (de -identified)	
• Research datasets	
PHI Status:	NO (properly de -identified)
Vendor:	△ YES (Storage provider)
BAA:	NOT REQUIRED (not PHI)
Controls:	
• Separate environment/keys	
• Access controls (research team only)	
• No re-identification risk	
• Annual re-certification	

DATA FLOW SUMMARY TABLE

Stage	Component	PHI?	Vendor?	BAA Required?	Key Controls
1. User Interaction	Chat/Voice/SMS	Yes	No	N/A	TLS, Input validation
2. Ingress & Routing	API Gateway, Auth	Yes	△ Yes	Yes	TLS, MFA, RBAC
3. API Processing	REST/GraphQL API	Yes	No	N/A	Tokenization, Logging
4. AI Inference	LLM, Vector DB	Yes	△ Yes	Yes	Min necessary, No training
5. Data Storage	Database, S3	Yes	△ Yes	Yes	Encryption at rest, KMS
6. Downstream Actions	EHR, Messaging	Yes	△ Yes	Yes	Secure APIs, Audit logs
7. Logging & Analytics	Logs, Metrics	△ Conditional	△ Yes	If PHI	Mask PHI, Pseudonymize
8. Model Training	Training Pipeline	No (after de-ID)	△ Yes	No	De-ID, Separate env
9. De-ID Storage	Analytics DB	No	△ Yes	No	Separate keys, Access control

CRITICAL DATA FLOW DECISIONS

Decision Point 1: LLM Provider Selection

Question: Which LLM provider will we use?

Options:

- **Option A:** External LLM (OpenAI, Anthropic)
 - Requires BAA
 - No training on PHI (contractual)
 - ⚠ PHI leaves your infrastructure
 - ⚠ Higher compliance risk
- **Option B:** Self-Hosted LLM
 - No BAA needed (you control it)
 - PHI stays in your infrastructure
 - ⚠ Higher infrastructure costs
 - ⚠ You're responsible for all security

Recommendation: Start with Option A (external with BAA) for faster time-to-market, consider Option B for high-security deployments.

Decision Point 2: Analytics Platform

Question: How do we handle analytics without violating HIPAA?

Options:

- **Option A:** HIPAA-eligible analytics with BAA (e.g., AWS Analytics)
 - Can include PHI
 - Requires BAA
 - Full analytics capabilities
- **Option B:** Self-hosted analytics (e.g., PostHog self-hosted)
 - No BAA needed
 - Full control
 - ⚠ More infrastructure to manage
- **Option C:** Pseudonymized analytics (e.g., Google Analytics with no PHI)
 - No BAA needed
 - Easy to implement
 - ⚠ Limited analytics capabilities

Recommendation: Use Option A or B for production analytics, Option C for marketing analytics.

Decision Point 3: De-identification Method

Question: Safe Harbor or Expert Determination?

Options:

- **Option A: Safe Harbor** (remove 18 identifiers)
 - Clear checklist
 - Easier to implement
 - ⚠ May be too restrictive for free text
- **Option B: Expert Determination**
 - More flexible for free text
 - Better for AI/ML use cases
 - ⚠ Requires expert certification
 - ⚠ More complex

Recommendation: Start with Safe Harbor, move to Expert Determination for advanced use cases.

VENDOR DEPENDENCIES MAP

Critical Path Vendors (Must Have BAAs)

1. **Cloud Provider** (AWS/Azure/GCP)
 - **PHI Access:** Infrastructure hosting
 - **BAA Status:** Standard BAA available
 - **Risk Level:** High (hosts all data)
2. **LLM Provider** (OpenAI/Anthropic)
 - **PHI Access:** Prompts, responses
 - **BAA Status:** ⚠ Check availability
 - **Risk Level:** High (PHI in prompts)
3. **Vector Database** (Pinecone/Weaviate)
 - **PHI Access:** Embeddings (ePHI)
 - **BAA Status:** ⚠ Check availability
 - **Risk Level:** High (embeddings contain PHI)
4. **EHR Integration** (Epic/Cerner/etc.)
 - **PHI Access:** Full PHI exchange
 - **BAA Status:** Standard BAA
 - **Risk Level:** High (PHI exchange)
5. **Messaging Provider** (SendGrid/Twilio)
 - **PHI Access:** Messages may contain PHI
 - **BAA Status:** ⚠ Check availability
 - **Risk Level:** Medium (messages)

Secondary Vendors (BAA If PHI Present)

6. **Logging Platform** (Datadog/Splunk)
 - **PHI Access:** Logs may contain PHI
 - **BAA Status:** ⚠ Check availability

- **Risk Level:** Medium (if PHI in logs)

7. Authentication (Auth0/AWS Cognito)

- **PHI Access:** User identifiers
- **BAA Status:** \triangle Check availability
- **Risk Level:** Low-Medium (identifiers)

8. Error Tracking (Sentry/Rollbar)

- **PHI Access:** Error traces may contain PHI
 - **BAA Status:** \triangle Check availability
 - **Risk Level:** Medium (if PHI in errors)
-

COMPLIANCE CHECKLIST BY STAGE

Stage 1: User Interaction

- ☐ TLS 1.2+ enabled on all endpoints
- ☐ Input validation implemented
- ☐ Rate limiting configured
- ☐ DDoS protection enabled

Stage 2: Ingress & Routing

- ☐ BAA signed with cloud provider
- ☐ BAA signed with authentication provider
- ☐ MFA enabled for all users
- ☐ RBAC implemented
- ☐ Request logging (no PHI)

Stage 3: API Processing

- ☐ PHI tokenization implemented
- ☐ Input sanitization
- ☐ Memory encryption
- ☐ Masked logging

Stage 4: AI Inference

- ☐ BAA signed with LLM provider
- ☐ BAA signed with vector DB provider
- ☐ Minimum necessary PHI in prompts
- ☐ Pseudonymization where possible
- ☐ No training on PHI (contractual)
- ☐ Per-tenant vector indexes
- ☐ Audit logging of API calls

Stage 5: Data Storage

- ☐ BAA signed with database provider
- ☐ BAA signed with object storage provider

- ☐ AES-256 encryption at rest
- ☐ KMS key management
- ☐ Database access controls (RBAC)
- ☐ Backup encryption
- ☐ Data retention policies
- ☐ Per-tenant data isolation

Stage 6: Downstream Actions

- ☐ BAA signed with EHR vendor
- ☐ BAA signed with messaging provider
- ☐ Secure API connections (TLS)
- ☐ Output validation
- ☐ Audit logging

Stage 7: Logging & Analytics

- ☐ BAA signed with logging platform (if PHI)
- ☐ PHI masking in logs
- ☐ Pseudonymized analytics
- ☐ No PHI to ad networks
- ☐ Separate PHI logs from analytics

Stage 8: Model Training

- ☐ De-identification pipeline implemented
- ☐ Separate de-identified environment
- ☐ Expert certification (if Expert Determination)
- ☐ No re-identification risk

Stage 9: De-Identified Storage

- ☐ Separate environment/keys
- ☐ Access controls
- ☐ Annual re-certification

CONCLUSION

This data flow diagram provides a complete view of where PHI flows through Wellness Agent AI. Use this to:

1. Identify all vendors that need BAAs
2. Determine security controls at each stage
3. Plan de-identification pipeline
4. Design audit logging
5. Create incident response procedures

Next Steps:

1. Review this diagram with your team
2. Complete vendor inventory
3. Obtain all required BAAs
4. Implement controls at each stage

5. Document in your risk analysis

Pattern: DATA × FLOW × HIPAA × COMPLIANCE × ONE

Status: **COMPLIANCE READY**

∞ **AbēONE** ∞