# **Wellness Companion AI - Complete API Specification**

## **Service 1: Desktop App → Core Backend APIs**

### **Authentication APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/auth/login** | **Initiate OAuth login** | **{}** | **{auth\_url, session\_id}** |
| **POST** | **/api/auth/callback** | **OAuth callback handler** | **{code, state}** | **{access\_token, refresh\_token, user\_info}** |
| **POST** | **/api/auth/refresh** | **Refresh JWT token** | **{refresh\_token}** | **{access\_token, expires\_in}** |
| **POST** | **/api/auth/logout** | **Logout user** | **{access\_token}** | **{success}** |
| **GET** | **/api/auth/profile** | **Get user profile** | **Headers: Authorization** | **{user\_id, email, name, preferences}** |
| **PUT** | **/api/auth/profile** | **Update user profile** | **{name, preferences}** | **{updated\_profile}** |

### **Document Management APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/documents/upload** | **Upload document** | **multipart/form-data** | **{document\_id, status, processing\_id}** |
| **GET** | **/api/documents/list** | **List user documents** | **?page=1&limit=20&filter=pdf** | **{documents[], total, page\_info}** |
| **GET** | **/api/documents/{doc\_id}** | **Get document details** | **Path: doc\_id** | **{document\_info, metadata, status}** |
| **DELETE** | **/api/documents/{doc\_id}** | **Delete document** | **Path: doc\_id** | **{success, cleanup\_status}** |
| **GET** | **/api/documents/{doc\_id}/status** | **Check processing status** | **Path: doc\_id** | **{status, progress, eta}** |
| **POST** | **/api/documents/bulk-upload** | **Bulk document upload** | **multipart/form-data** | **{batch\_id, document\_ids[]}** |
| **GET** | **/api/documents/search** | **Search documents by metadata** | **?query=text&type=pdf** | **{documents[], relevance\_scores}** |

### **Chat & Conversation APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **GET** | **/api/chat/conversations** | **List conversations** | **?limit=50&offset=0** | **{conversations[], total}** |
| **POST** | **/api/chat/conversations** | **Create new conversation** | **{title?, initial\_message?}** | **{conversation\_id, created\_at}** |
| **GET** | **/api/chat/conversations/{conv\_id}** | **Get conversation history** | **Path: conv\_id** | **{messages[], metadata}** |
| **POST** | **/api/chat/conversations/{conv\_id}/messages** | **Send message** | **{message, context?, model\_preference?}** | **{message\_id, response, sources[], confidence}** |
| **GET** | **/api/chat/conversations/{conv\_id}/messages** | **Get messages** | **?limit=50&before=timestamp** | **{messages[], has\_more}** |
| **POST** | **/api/chat/conversations/{conv\_id}/regenerate** | **Regenerate last response** | **{message\_id, model\_preference?}** | **{new\_response, sources[], confidence}** |
| **DELETE** | **/api/chat/conversations/{conv\_id}** | **Delete conversation** | **Path: conv\_id** | **{success}** |
| **PUT** | **/api/chat/conversations/{conv\_id}** | **Update conversation** | **{title, archived?}** | **{updated\_conversation}** |

### **Search & Knowledge APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/search/semantic** | **Semantic vector search** | **{query, limit?, threshold?, document\_ids?}** | **{results[], confidence\_scores[], sources[]}** |
| **POST** | **/api/search/web** | **Web search fallback** | **{query, context?, max\_results?}** | **{web\_results[], processed\_content, sources[]}** |
| **POST** | **/api/search/hybrid** | **Combined local + web search** | **{query, confidence\_threshold?, web\_fallback?}** | **{local\_results[], web\_results[], combined\_response}** |
| **GET** | **/api/search/suggestions** | **Query suggestions** | **?partial\_query=text** | **{suggestions[], categories[]}** |
| **POST** | **/api/search/rerank** | **Rerank search results** | **{query, results[], rerank\_model?}** | **{reranked\_results[], scores[]}** |
| **GET** | **/api/knowledge/status** | **Knowledge base status** | **{}** | **{total\_documents, total\_chunks, last\_updated, health}** |
| **POST** | **/api/knowledge/rebuild** | **Rebuild knowledge base** | **{force?, document\_ids?}** | **{rebuild\_id, estimated\_time}** |
| **GET** | **/api/knowledge/rebuild/{rebuild\_id}/status** | **Check rebuild status** | **Path: rebuild\_id** | **{status, progress, eta}** |

### **System & Admin APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **GET** | **/api/system/health** | **Overall system health** | **{}** | **{status, services[], uptime, version}** |
| **GET** | **/api/system/metrics** | **System metrics** | **?duration=1h** | **{cpu, memory, requests, response\_times}** |
| **GET** | **/api/system/users** | **User management (admin)** | **?page=1&limit=20** | **{users[], total, statistics}** |
| **GET** | **/api/system/usage** | **Usage analytics** | **?user\_id&period=week** | **{queries\_count, documents\_count, storage\_used}** |
| **POST** | **/api/system/maintenance** | **Trigger maintenance** | **{maintenance\_type, schedule?}** | **{maintenance\_id, schedule\_info}** |
| **GET** | **/api/system/logs** | **System logs (admin)** | **?level=error&limit=100** | **{logs[], total}** |

## **Service 2: Core Backend → AI/ML Orchestration APIs**

### **RAG Orchestration APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/rag/query** | **Process RAG query** | **{query, context?, user\_id, conversation\_id?}** | **{response, sources[], confidence, processing\_time}** |
| **POST** | **/api/rag/query/stream** | **Streaming RAG query** | **{query, context?, stream=true}** | **Server-Sent Events** |
| **GET** | **/api/rag/models** | **List available models** | **{}** | **{llm\_models[], embedding\_models[], status[]}** |
| **POST** | **/api/rag/models/switch** | **Switch active model** | **{model\_name, model\_type}** | **{success, switch\_time, model\_info}** |
| **GET** | **/api/rag/config** | **Get RAG configuration** | **{}** | **{chunk\_size, overlap, similarity\_threshold, models}** |
| **PUT** | **/api/rag/config** | **Update RAG configuration** | **{chunk\_size?, overlap?, similarity\_threshold?}** | **{updated\_config}** |

### **Vector Search APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/vector/search** | **Vector similarity search** | **{query\_vector, limit?, threshold?, filters?}** | **{matches[], scores[], metadata[]}** |
| **POST** | **/api/vector/search/text** | **Text-based vector search** | **{query\_text, limit?, threshold?, document\_ids?}** | **{results[], confidence\_scores[]}** |
| **GET** | **/api/vector/stats** | **Vector database statistics** | **{}** | **{total\_vectors, dimensions, collections[], health}** |
| **POST** | **/api/vector/reindex** | **Trigger reindexing** | **{collection?, force?}** | **{reindex\_id, estimated\_time}** |
| **GET** | **/api/vector/reindex/{reindex\_id}/status** | **Check reindex status** | **Path: reindex\_id** | **{status, progress, eta}** |

### **LLM Management APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **GET** | **/api/llm/models** | **List available LLM models** | **{}** | **{models[], loaded[], available[]}** |
| **POST** | **/api/llm/models/load** | **Load LLM model** | **{model\_name, quantization?}** | **{success, load\_time, memory\_usage}** |
| **POST** | **/api/llm/models/unload** | **Unload LLM model** | **{model\_name}** | **{success, freed\_memory}** |
| **POST** | **/api/llm/generate** | **Generate text** | **{prompt, model?, max\_tokens?, temperature?}** | **{generated\_text, tokens\_used, processing\_time}** |
| **GET** | **/api/llm/status** | **LLM service status** | **{}** | **{loaded\_models[], memory\_usage, gpu\_usage?}** |
| **POST** | **/api/llm/benchmark** | **Benchmark models** | **{models[], test\_prompts[]}** | **{benchmark\_results[], performance\_metrics}** |

### **Web Search APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/websearch/search** | **Web search via Tavily** | **{query, max\_results?, include\_domains?, exclude\_domains?}** | **{results[], processed\_content, sources[]}** |
| **POST** | **/api/websearch/enhanced** | **Enhanced web search** | **{query, context?, rerank?}** | **{enhanced\_results[], relevance\_scores[]}** |
| **GET** | **/api/websearch/config** | **Get web search config** | **{}** | **{api\_limits, enabled\_sources, filters}** |
| **PUT** | **/api/websearch/config** | **Update web search config** | **{max\_results?, domains?, filters?}** | **{updated\_config}** |

## **Service 3: Core Backend → Data Layer APIs**

### **Document Processing APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/data/documents/process** | **Process uploaded document** | **{document\_id, file\_path, user\_id}** | **{processing\_id, estimated\_chunks}** |
| **GET** | **/api/data/documents/process/{process\_id}/status** | **Check processing status** | **Path: process\_id** | **{status, progress, chunks\_processed, eta}** |
| **POST** | **/api/data/documents/reprocess** | **Reprocess document** | **{document\_id, new\_settings?}** | **{processing\_id, changes\_detected}** |
| **GET** | **/api/data/documents/{doc\_id}/chunks** | **Get document chunks** | **Path: doc\_id, ?page=1&limit=50** | **{chunks[], total, metadata}** |
| **DELETE** | **/api/data/documents/{doc\_id}** | **Delete document data** | **Path: doc\_id** | **{success, cleanup\_status}** |

### **Embedding Pipeline APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/data/embeddings/generate** | **Generate embeddings** | **{texts[], model?, batch\_size?}** | **{embeddings[], processing\_time}** |
| **POST** | **/api/data/embeddings/store** | **Store embeddings** | **{embeddings[], metadata[], document\_id}** | **{stored\_count, vector\_ids[]}** |
| **GET** | **/api/data/embeddings/models** | **List embedding models** | **{}** | **{models[], active\_model, capabilities}** |
| **POST** | **/api/data/embeddings/models/switch** | **Switch embedding model** | **{model\_name}** | **{success, migration\_required}** |
| **POST** | **/api/data/embeddings/batch** | **Batch embed documents** | **{document\_ids[], priority?}** | **{batch\_id, estimated\_time}** |
| **GET** | **/api/data/embeddings/batch/{batch\_id}/status** | **Check batch status** | **Path: batch\_id** | **{status, progress, completed[], failed[]}** |

### **Database Management APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **GET** | **/api/data/vector/collections** | **List vector collections** | **{}** | **{collections[], stats[], health}** |
| **POST** | **/api/data/vector/collections** | **Create collection** | **{name, dimension, distance\_metric?}** | **{collection\_id, created\_at}** |
| **DELETE** | **/api/data/vector/collections/{collection\_id}** | **Delete collection** | **Path: collection\_id** | **{success, deleted\_count}** |
| **GET** | **/api/data/postgres/stats** | **PostgreSQL statistics** | **{}** | **{tables[], row\_counts, size, health}** |
| **POST** | **/api/data/postgres/backup** | **Trigger database backup** | **{backup\_type?, schedule?}** | **{backup\_id, estimated\_size}** |
| **GET** | **/api/data/redis/stats** | **Redis cache statistics** | **{}** | **{memory\_usage, keys\_count, hit\_rate}** |
| **POST** | **/api/data/redis/flush** | **Flush Redis cache** | **{pattern?, confirm=true}** | **{flushed\_keys, success}** |

### **File Storage APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/data/storage/upload** | **Upload file to S3** | **multipart/form-data** | **{file\_id, s3\_url, upload\_time}** |
| **GET** | **/api/data/storage/{file\_id}** | **Get file info** | **Path: file\_id** | **{file\_info, download\_url, metadata}** |
| **DELETE** | **/api/data/storage/{file\_id}** | **Delete file** | **Path: file\_id** | **{success, freed\_space}** |
| **GET** | **/api/data/storage/usage** | **Storage usage stats** | **?user\_id** | **{total\_size, file\_count, quota, breakdown}** |
| **POST** | **/api/data/storage/cleanup** | **Cleanup orphaned files** | **{dry\_run?}** | **{cleanup\_id, files\_to\_delete[]}** |

### **Synchronization APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **POST** | **/api/data/sync/trigger** | **Trigger cloud sync** | **{user\_id, sync\_type?, force?}** | **{sync\_id, estimated\_time}** |
| **GET** | **/api/data/sync/{sync\_id}/status** | **Check sync status** | **Path: sync\_id** | **{status, progress, conflicts[], eta}** |
| **POST** | **/api/data/sync/resolve-conflicts** | **Resolve sync conflicts** | **{sync\_id, resolutions[]}** | **{resolved\_count, remaining\_conflicts}** |
| **GET** | **/api/data/sync/history** | **Sync history** | **?user\_id&limit=20** | **{sync\_history[], statistics}** |

## **Service 4: Internal Health & Monitoring APIs**

### **Service Health APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **GET** | **/health** | **Basic health check** | **{}** | **{status: "healthy", timestamp}** |
| **GET** | **/health/detailed** | **Detailed health check** | **{}** | **{services[], dependencies[], performance}** |
| **GET** | **/health/dependencies** | **Check external dependencies** | **{}** | **{postgres, redis, s3, ollama, qdrant}** |
| **GET** | **/metrics** | **Prometheus metrics** | **{}** | **Prometheus format metrics** |
| **GET** | **/metrics/custom** | **Custom application metrics** | **{}** | **{business\_metrics, performance\_metrics}** |

### **Logging & Debug APIs**

| **Method** | **Endpoint** | **Description** | **Request** | **Response** |
| --- | --- | --- | --- | --- |
| **GET** | **/api/debug/logs** | **Service logs** | **?level=error&limit=100&service=core** | **{logs[], total}** |
| **POST** | **/api/debug/trace** | **Enable request tracing** | **{trace\_id, duration?}** | **{success, trace\_info}** |
| **GET** | **/api/debug/performance** | **Performance diagnostics** | **?duration=1h** | **{response\_times[], bottlenecks[]}** |
| **POST** | **/api/debug/simulate-error** | **Simulate error (dev only)** | **{error\_type, severity?}** | **{simulated\_error}** |

## **Service 5: WebSocket APIs**

### **Real-time Communication**

| **Event** | **Direction** | **Payload** | **Description** |
| --- | --- | --- | --- |
| **connect** | **Client → Server** | **{auth\_token, user\_id}** | **Establish WebSocket connection** |
| **subscribe** | **Client → Server** | **{channels[]}** | **Subscribe to channels** |
| **message** | **Client → Server** | **{channel, message, type}** | **Send message** |
| **response** | **Server → Client** | **{response, sources[], confidence}** | **AI response** |
| **notification** | **Server → Client** | **{type, message, action?}** | **System notification** |
| **status\_update** | **Server → Client** | **{service, status, details}** | **Service status change** |
| **processing\_update** | **Server → Client** | **{process\_id, progress, eta}** | **Processing progress** |
| **disconnect** | **Both** | **{reason?}** | **Connection termination** |

## **External API Integrations**

### **Third-party APIs Used**

| **Service** | **Purpose** | **Key Endpoints** |
| --- | --- | --- |
| **Google OAuth** | **Authentication** | **/oauth2/auth, /oauth2/token** |
| **AWS Cognito** | **User management** | **/oauth2/authorize, /oauth2/token** |
| **Tavily API** | **Web search** | **/search, /extract** |
| **Ollama API** | **Local LLM** | **/api/generate, /api/chat** |
| **Sentry API** | **Error tracking** | **/api/projects/{id}/events/** |

### **Rate Limiting & Quotas**

| **API Category** | **Rate Limit** | **Quota** |
| --- | --- | --- |
| **Authentication** | **10/min per IP** | **1000/day per user** |
| **Document Upload** | **5/min per user** | **100MB/day per user** |
| **Chat Messages** | **30/min per user** | **1000/day per user** |
| **Search Queries** | **60/min per user** | **5000/day per user** |
| **System APIs** | **100/min per user** | **No daily limit** |

### **Error Response Format**

**{**

**"error": {**

**"code": "ERROR\_CODE",**

**"message": "Human readable message",**

**"details": {},**

**"timestamp": "2024-08-02T10:00:00Z",**

**"request\_id": "req\_123456"**

**}**

**}**

### **Authentication Headers**

**Authorization: Bearer <jwt\_token>**

**X-API-Key: <api\_key> (for service-to-service)**

**X-Request-ID: <unique\_request\_id>**

**X-User-ID: <user\_id> (for user context)**