SkillPilot Security & Compliance Dossier · Volume V

Security, Data Privacy, and AI Transparency

Pseudonymous Storage and Deletion Boundaries, OpenAI MCP Client Binding

(mTLS, OAuth, 24h Learning Session), and EU AI Act Compliance

Scope of this dossier

This volume consolidates SkillPilot's technical security, data-privacy, and AI-transparency specifications: the data-flow and storage architecture across browser, pseudonymous backend, active Claude beta lane, and OpenAI integration candidate including the 365-day inactivity deletion boundary; the four-proof OpenAI MCP client-binding architecture (HTTPS, connector mTLS, OAuth with PKCE `S256`, and 24-hour learning sessions); and the EU AI Act compliance assessment (Article 50 transparency and Annex III boundaries).

As of 11 November 2026 · License: Apache-2.0 (technical security and compliance documentation; see LICENSING.md).

Technical architecture and inventory documentation — does not replace legal advice or the binding public policies at skillpilot.com/privacy and skillpilot.com/legal.

Contents

1 Data Privacy and Storage Concept **3**
1.1 Core Philosophy: Privacy by Design . . . 3
1.2 Data Partitioning . . . 4
1.3 Data Flow Scenarios . . . 7
1.4 Retention and Deletion . . . 8
1.5 Backup, Recovery, and Residual Operating Rules . . . 9

2 OpenAI MCP: Authorized Client Binding **11**
2.1 Protection Objective . . . 11
2.2 Four Independent Proofs . . . 11
2.3 Binding OAuth Client Profile . . . 12
2.4 OAuth Flow and Resource Validation . . . 12
2.5 Learning Session as Separate Application Binding . . . 12
2.6 mTLS at the MCP Edge . . . 13
2.7 Fail-Closed Startup Invariants and Secret Lifecycle . . . 13
2.8 Acceptance, Negative Tests, Residual Assumptions, and References . . . 13

3 EU AI Act: Compliance Baseline for SkillPilot **15**
3.1 Regulatory Calendar . . . 15
3.2 Verified Product Inventory . . . 15
3.3 Article 50: Classification and Implementation . . . 16
3.4 Annex III: Preliminary High-Risk Boundary . . . 17
3.5 Evidence Status, Release Gates, and Ongoing Maintenance . . . 17

1 Data Privacy and Storage Concept

Technical data-flow and storage baseline

Status: updated on 2026-09-13 (with October 2026 neutral OpenAI endpoint cutover alignment) for the active Claude-first beta, the subsequent ChatGPT integration candidate and Windows Desktop Git beta, the coach-tool free-text boundary, voluntary feedback, and the SkillPilot-ID deletion and retention boundary. Existing teacher-view and personal-subject-schedule details remain applicable.

This is a technical data-flow and storage description. It does not replace the learner-facing Skill Pilot Privacy Policy, the provider's privacy terms, or a legal review before a public release. The public policy is authored in `app/src/utils/privacyViewCopy.ts`. Repository defaults below are not evidence of the configuration or retention of every deployed service, log copy, or backup.

The current work order is Claude beta → stable candidate → focused ChatGPT integration acceptance → official submission. The provider-neutral coach boundary and Claude runbook govern the active provider lane. The separate OpenAI MCP identity and 24-hour session contract and 2 describe the subsequent OpenAI integration candidate and explicitly authorized Desktop Git beta, not an accepted hosted directory release.

1.1 Core Philosophy: Privacy by Design

SkillPilot separates **pseudonymous learner identity**, **persistent learning state**, **temporary provider access**, and **provider-side conversation content**:

- The SkillPilot backend stores learning progress under a random, pseudonymous SkillPilot ID. It does not require a learner name, email address, or password. This is pseudonymity, not anonymity: the ID links a learning history, and a person or institution with an ID-to-person mapping can identify its owner.
- The browser is the learner's SkillPilot login surface. It may hold the permanent SkillPilot ID locally and can optionally save it encrypted with a user-chosen password.
- The permanent SkillPilot ID stays in first-party SkillPilot surfaces and outside the AI chat, OAuth principal, MCP tools, and provider-hosted widgets. Every explicit WebGUI start creates a separate temporary learning-session reference that the backend resolves to that ID.
- The AI provider processes the chat, uploads, projected learning context, and tool results under the provider account and its terms.
- SkillPilot does not receive the complete provider chat transcript. It does receive every explicit API or MCP tool request sent to it, including the arguments needed to read or change learning state.

Architectural invariant for coach tools: learner answers, explanations, chat summaries, and coach-authored feedback must stay in the provider chat. Neither `workFeedback`, `outcomeFeedback`, Recall `feedback`/`lastFeedback`, nor renamed equivalents belong in accepted coach inputs, persisted learning state, replay responses, or diagnostics. The coach makes the pedagogical assessment and composes its own success message; SkillPilot accepts only the authorized structured decision and returns confirmed state (see provider-neutral coach boundary). The 12 September 2026 cleanup record records a historical production purge and the rollout obligations identified at that time; it is not evidence of the state of a later deployment.

Provider status must stay explicit:

1. **Claude V1 OAuth/MCP:** the active learning beta, distributed through the personal Claude marketplace. Installation and updates do not bypass OAuth, session isolation, or domain authorization.
2. **Multilingual OpenAI V1 OAuth/MCP App:** the ChatGPT integration candidate (`skillpilot-coach` on the neutral target endpoint `https://mcp-openai.skillpilot.com/mcp`, alongside the explicitly authorized Windows Desktop Git beta). Full host acceptance across Desktop/Voice, Web, and mobile remains required before official directory publication; repository code, local tests, or a successful Claude test do not establish ChatGPT host acceptance.
3. **Visible Session:** historical Custom-GPT rollback contract only, not an advertised beta or the current MCP contract. Its temporary `sps_...` bearer token is deliberately visible in the prepared message and state footer.

The provider lanes share the SkillPilot core while retaining strictly separate OAuth and session boundaries. Locale support and the selected frontend integration are release-specific; their presence in source is not a public platform promise.

1.2 Data Partitioning

A. Local Client (Browser / Local Storage) **Trust boundary:** controlled by the learner and the browser environment. The browser may hold:

- the active SkillPilot ID;
- optional named local profiles in which the SkillPilot ID is encrypted with a user-chosen password (the password itself is not stored);
- a browser-local OpenAI eligibility confirmation bound to the current pseudonymous SkillPilot profile for the lifetime of the browser tab;
- the selected curriculum and other UI preferences;
- local SRS scheduling state before or between synchronization;
- the freshly prepared provider launch URL or start message, which contains a natural start message with a temporary learning-session reference, never a permanent SkillPilot ID, OAuth token, or client secret.

Every V1 coach start calls a learner-specific SkillPilot UI endpoint with the locally active permanent SkillPilot ID, creating a fresh learning-session reference at that exact moment. The Claude start opens its prepared chat; the ChatGPT Desktop beta displays the provider-specific message for the learner to paste into a new desktop chat. Because a launch URL or message carrying the temporary reference is itself sensitive, it can appear in browser history or a copied link and must be treated accordingly.

For teacher-led usage, the teacher's browser or institution-controlled storage may additionally hold local class names, learner names or aliases, permanent SkillPilot IDs, and cached Level-2 personalization. That local class mapping is not stored centrally by SkillPilot. The existing-learner teacher view reads normal learner profile and mastery endpoints directly with the permanent ID. When the teacher first schedules a learning scope, one additional no-store read derives every atomic target ID in the learner's complete Level-2 personalization for the selected subject and its not-yet-mastered subset (unrestricted by the mutable Level-3 focus), storing those goal IDs, aggregate counts, and capture time as an immutable local planning baseline without storing the SkillPilot ID or numeric per-goal mastery values inside that baseline. The ordinary teacher view does not change learner-side planned goals and creates no server-side teacher, class, authorization record, or membership relationship.

The teacher view disables learner-state mutations in the frontend, while its separate browser-local teacher course plan remains editable working data. One explicit, confirmed action can copy the plan label, dated blocks, and client-materialized atomic goal IDs into a separate personal subject schedule under the known permanent SkillPilot ID. The server validates the IDs against the current personalized subject scope: newly added IDs are accepted only while open, whereas IDs already present in the personal schedule may remain in a confirmed replacement to preserve continuity. The write transfers no local class ID, alias, coverage journal, attestation, mastery value, learner-derived planning baseline, or teacher-plan history and creates no teacher-to-learner relation. Later local teacher-plan edits are not synchronized; teacher-entered labels and block titles are copied unchanged and may contain personal data.

The confirmed first-party action **Make planning effective** validates the complete set of current subject plans before writing all independent subject copies or none, enables plan-guided learning under the known permanent SkillPilot ID, and immediately selects the first due goal whose prerequisites are satisfied (pausable by the learner at any time). The subordinate **Update this subject only** action replaces only that subject copy without enabling plan-guided learning or selecting a goal. An explicit subject switch parks an unfinished active goal without changing mastery and selects an eligible due goal in the chosen subject; after confirmed completion, a valid plan containing the completed goal has priority, and otherwise valid plans follow deterministic due-urgency and subject order. Stale or invalid plans are ignored, no replacement is invented when no due eligible goal exists, generic Autopilot stays suppressed while plan-guided learning is enabled, and calendar progress alone causes no learner-state write. Plan exports omit the learner-derived baseline, though teacher-entered free text is exported unchanged.

The permanent SkillPilot ID remains the bearer secret and full-access key for the learner state (deleting a local class leaves the ID valid and unchanged). A downloaded local-class export is therefore always encrypted in the browser with a teacher-chosen password before leaving the application, using versioned PBKDF2-SHA-256 and authenticated AES-256-GCM; SkillPilot neither stores nor recovers the password. Legacy plaintext JSON class exports remain importable for migration (flagged as unprotected in the UI), while new exports are never written in plaintext. File encryption protects the downloaded file at rest; it does not encrypt browser local storage, secure an unlocked device, invalidate the ID, or make the ID read-only after decryption.

B. Persistent Learning State in the SkillPilot Backend

Trust boundary: pseudonymous, centralized SkillPilot storage. The backend stores:

- `Learner`: the permanent pseudonymous learner key;
- provider-specific active learning sessions, including their absolute expiry and direct internal link to the learner;
- `Mastery`: learning-goal mastery keyed by learner and goal;
- planned goals and learner configuration, including curriculum, personalization, scope, active goal, filters, and synchronized client state;
- optional personal subject schedules, their revisions and materialized atomic goal IDs, plus the current preference for plan-guided learning (initially off until confirmed activation);
- static curricula, competence definitions, SRS decks, resources, and review metadata.

No model provider needs the permanent SkillPilot ID to use the active Claude integration or the OpenAI candidate; it remains the internal join key from a temporary session to the learning state. The operator confirmed that the SkillPilot Core is hosted in Germany. This is not a statement that the selected AI provider, support services, or every external processor operates only in Germany; provider-side processing follows that provider's disclosed terms.

C. Historical Visible Session Data — Rollback Only This retained legacy fallback is neither the active Claude beta nor a current ChatGPT MCP contract. The backend stores a `ChatSession` containing the HMAC hash of the temporary token, its learner association, language, start/expiry/use timestamps, and optional revocation metadata; plaintext `sps_...` tokens are returned only at session creation and must never be logged. ChatGPT sees the temporary token in the prepared message and required state footer, projected coaching context, visible selection numbers/references, public canonical learning-goal IDs, and Recall card IDs/prompts. The token is a short-lived credential (expiring absolutely after at most 24 hours without extension on use), not an anonymization of the conversation.

D. Multilingual OpenAI V1 OAuth/MCP App Candidate Data This section describes the implemented OpenAI V1 security and data boundary (`skillpilot-coach` prepared for the neutral protected endpoint `https://mcp-openai.skillpilot.com/mcp`, superseding the historical `https://mcp-coach-v1.skillpilot.com/mcp` origin; actual host acceptance, cutover rollout, and directory publication remain separate steps). The backend isolates this lane from Visible Session tables and Claude connection records.

The dedicated protected `/mcp` endpoint uses server-authenticated HTTPS, OpenAI connector mTLS, and a valid OAuth access token, mapped by a dedicated nginx virtual host to the loopback-only Spring handler `/internal/openai/v1/mcp` (see 2). For the hosted ChatGPT directory profile, the authorization server accepts the single configured confidential OAuth client (`client_secret_basic` + PKCE `S256` + exact client ID, redirect URI, resource, audience, and scopes, with the client secret stored solely in ChatGPT App configuration and the SkillPilot secret store); for the explicitly authorized native Windows Desktop route, a separate, strictly pinned public CIMD profile with PKCE `S256` is used and never acts as a fallback from confidential authentication. Open DCR, unpinned CIMD, `none`, `private_key_jwt`, and implicit profile fallback are excluded. The mTLS edge validates OpenAI's published CA chain, `clientAuth` EKU, and exact connector SAN, recording only bounded classifications (`VERIFIED`, temporary `OBSERVE_NO_CERT`, or loopback-only `LOCAL_OPERATOR`), never raw certificates (`observe` serves initial integration proof; hosted publication requires `enforce`).

The backend stores the OAuth authorization/consent/code/opaque-access/rotating-refresh records and a separate learning-session record per explicit **Lernen starten** containing only an HMAC/hash of the high-entropy reference, learner association, `started_at`, `expires_at`, and revocation metadata. OAuth authenticates the App, whereas the session addresses the learner: every domain MCP call requires both a valid resource/scope-bound OAuth token and a valid session reference. Supported launch intents are bounded (current unit, Verified Recall with goal and batch size, and reviewed Abi 2026 exam entry with goal and course level); free browser-supplied prompt instructions are not persisted.

Creating or loading a permanent SkillPilot ID, confirming provider notices, and configuring Level-2 personalization occur exclusively in the first-party SkillPilot WebGUI. ChatGPT holds OAuth access/refresh tokens strictly as transport credentials in the HTTP `Authorization` header (unseen by the user and outside model prompts/arguments), while the temporary `learningSessionId` is inserted into the launch message for domain tool calls without exposing the permanent SkillPilot ID. Through `CoachStateProjection`, MCP projects bounded workflow context (curriculum/scope summaries, active goal, progress, permitted actions, public goal/card IDs, task content, selected resources); exam solutions/scoring require the authorized evaluation path, and Recall answers unlock only after the learner-answer step.

E. Active Claude V1 OAuth/MCP Data Claude is the active learning-beta provider: the personal marketplace supplies the plugin package and coach instructions, while learning operations

call the isolated Claude V1 MCP boundary over HTTPS and OAuth. The Claude authorization principal is a generated, learner-independent technical connection reference rather than the permanent SkillPilot ID. Each explicit first-party start creates a fresh high-entropy 24-hour learning-session reference (stored as hash, learner association, creation/expiry timestamps, locale, and state version; never extended by chat activity or OAuth refresh). Both valid OAuth authorization and the temporary learning session are required for learner-specific calls.

Claude's client-authentication profiles are separate from OpenAI's (supporting the constrained Claude public CIMD profile and an explicitly configured confidential profile with PKCE and provider-specific validation; OpenAI connector mTLS does not apply to Claude, and a public client identifier is not a secret). Projected coaching context omits the permanent learner key while remaining pseudonymous learning data. Incoming operations accept only structured goal/card references, language, concurrency/idempotency identifiers, authorized capability references, and bounded decisions (optional numeric exam score; boolean pass/fail per Recall card); unsupported result fields are rejected before core mutation or replay persistence, while answers, photos, voice inputs, and coach explanations stay in the Claude chat (`ClaudeV1AppAuthenticationFilter`, `ClaudeV1LearningSessionService`, `ClaudeV1McpContractAdapter`, `ClaudeV1CoachContextProjector`).

F. Deliberately Submitted Learning-Goal Feedback The cockpit and learning-goal book open the same voluntary public feedback form (`credentials: 'omit'`, supplying public goal/book context without a learner ID and never auto-exporting provider chats). Under its dedicated notice and explicit confirmations, the form accepts a category, observation, and optional explanation/evidence, suggestion, source, and perspective without name, email, or attachment fields (users are instructed not to paste chats, credentials, or personal details, as length/field checks cannot detect arbitrary personal prose).

`goal_feedback_submission` stores submitted text with random feedback/submission IDs, receipt time, public goal/book/page bindings and fingerprints, notice version/language, confirmations, and technical digests, with **no learner foreign key** (rate limiting hashes the client IP in memory for short-lived counters without proving absence of network metadata in infrastructure logs). Pending submissions become due for deletion after 30 days (daily cleanup plus cleanup on intake/export, hence the form states at most 31 days online). Operators export batches to access-restricted local files under `tmp/goal-feedback/inbox/<export-id>` and acknowledge the batch digest after verified local storage, deleting online content and leaving a content-free batch receipt. Exported text may undergo disclosed technically assisted review; local raw files and Codex session/tool records form separate retention boundaries requiring manual post-review removal per the feedback intake runbook (`GoalFeedbackSubmissionService`, `GoalFeedbackRetentionService`, `GoalFeedbackExportService`, `GoalBookFeedbackPilotView.tsx`). Withdrawal or deletion requests use `support@skillpilot.com` (preferably citing the feedback ID).

1.3 Data Flow Scenarios

- **Individual Learner Login:** The learner opens `https://skillpilot.com`; the browser creates a SkillPilot ID, loads an encrypted local profile, or accepts an existing ID, stores curriculum/configuration under that pseudonymous ID in the backend, and enters the cockpit or starts an available coach integration.
- **Claude Beta Connection and Learning Start:** The learner installs the Claude plugin and completes OAuth (separate from the learner profile), configures context in the SkillPilot WebGUI, and triggers **Lernen starten** / **Start learning** to mint a 24-hour Claude learning

session; Claude calls MCP tools with OAuth + temporary session reference, evaluates work inside the Claude conversation, and sends only structured decisions back to SkillPilot.
- **Historical ChatGPT Visible Session (Rollback Only):** `POST /api/ui/learners/{skillpilotId}/visible-chat-start` creates a 24-hour `sps_...` token (HMAC-hashed on the server) and opens the legacy Custom GPT against `/api/ai/{lang}/sessions/.../visible`.
- **Candidate Acceptance — OpenAI V1 App Connection:** After browser-local country age/guardian confirmation (collecting no birth date and inferring no age from grade level), ChatGPT runs OAuth Authorization Code + PKCE `S256` against the pinned client profile (`client_secret_basic` for hosted confidential App or separate pinned desktop CIMD profile), issuing resource/scope-bound tokens that authorize the App without selecting a learner or session.
- **Open MCP Coach Without a Prepared Session:** Finding no `learningSessionId`, the coach calls no tool, asks for neither permanent nor temporary IDs, and instructs the learner to open `https://skillpilot.com/` and click **Lernen starten** / **Start learning**.
- **Candidate Acceptance — OpenAI Session With Backend-Pinned Locale:** Clicking **Lernen starten** mints a fresh 24-hour session (storing HMAC hash, learner link, `communicationLocale`, `started_at`, `expires_at`, revocation metadata; repeated starts create distinct sessions without extending older ones) and injects the temporary reference into the launch prompt; backend MCP handlers require both valid OAuth and valid session without fallback.
- **App Authorization Revocation & Browser Cockpit Review:** Server-side OAuth revocation invalidates the selected authorization and its access/refresh tokens while leaving separate session rows unusable without OAuth and preserving durable learning progress. Direct cockpit inspection (`GET /api/ui/learners/{skillpilotId}/state`) stays strictly between browser and SkillPilot backend.

1.4 Retention and Deletion

Repository-default credential validity periods for Claude V1 and the OpenAI V1 candidate are:

- **OAuth access token:** 1 hour;
- **Rotating refresh token:** 30 days;
- **Claude V1 learning session:** a fixed 24 hours from its first-party start;
- **OpenAI V1 learning session:** exactly 24 hours from the corresponding confirmed first-party **Lernen starten** action.

The learning-session deadline is never extended by MCP requests, access-token refresh, browser reload, a new provider chat, or context compaction. Refresh-token rotation can establish a later credential expiry and must not be described as automatic deletion of all connection data after 30 days; session cleanup jobs default to hourly removal of expired rows. OAuth validity is not a general database-retention deadline for all expired authorization, consent, or diagnostic records, and deleting a learner does not automatically delete every learner-independent technical provider authorization. Historical Visible Session tokens expire after at most 24 hours, and learning progress is never deleted merely because an AI connection expires or is revoked.

The active learning state stored under a SkillPilot ID in the SkillPilot database, including its associated SkillPilot learning sessions and SkillPilot connections, has one common deletion boundary:

- the learner can delete it manually with the designated WebGUI function; and

- after 365 consecutive days without a successful activity, it becomes due for automatic deletion and is removed during the next automatic deletion run.

Activity is recorded only for these successfully completed boundaries:

- successful creation of a SkillPilot ID;
- foreground loading or resuming of the learning state through the SkillPilot WebGUI;
- a server-completed import or export of signed learner data;
- a learner-state change successfully stored on the server; or
- a successfully completed SkillPilot session or AI-provider connection action;
- a valid Coach/MCP invocation that completes with a successful domain result.

Background GET requests, SSE traffic, OAuth token refreshes, merely selecting, opening, or reading a local file, and server operations that do not complete or are domain-rejected do not count and do not restart the 365-day period.

After deletion, the permanent SkillPilot ID can no longer access that learning state and its associated SkillPilot learning sessions and learner-bound connections cannot be used; a separately retained technical OAuth authorization cannot restore or select the deleted learner. Because Skill Pilot has no required real-world identity link, profile-related requests require proportionate proof of access through an agreed secure path (support must never solicit permanent IDs or temporary credentials in ordinary email; feedback requests use their feedback ID). This boundary does not delete local files or provider-side chats. Existing backup copies are not part of the active learning state. Neither manual deletion nor the 365-day expiry immediately deletes each backup copy individually.

Plaintext learning-session references, OAuth credentials/secrets, request bodies, learner answers, and launch secrets must not enter application, reverse-proxy, trace, or support logs. The general `RequestLoggingFilter` skips protected MCP, OAuth, feedback, and coach-result routes entirely, while bounded OAuth/MCP telemetry records operation, outcome, latency, correlation IDs, and HMAC-derived temporary-session fingerprints rather than arbitrary tool arguments. Other API routes can log field-redacted, size-limited bodies at `DEBUG`; optional AI tracing (`tmp/ai-trace.jsonl`, disabled by default) uses known-field/token-pattern redaction rather than general personal-data detection and has no automatic TTL or learner-deletion coupling. Likewise, legacy `ChatStartRequest.promptContext` (truncated to 2,000 characters and unpersisted in the session row) is not redacted as a whole by debug logging and must not be advertised or traced without reviewing text exposure.

1.5 Backup, Recovery, and Residual Operating Rules

Learners and institutions remain responsible for preserving access keys and external identity mappings: signed learner exports/imports move state across browsers, password-protected ID files store the SkillPilot ID encrypted, local-class exports encrypt teacher mappings (leaving password hygiene and obsolete-file deletion with the institution), logging out clears only the active browser login, and server-side learner deletion leaves local exports/ID files untouched until removed on the device.

`scripts/backup_db.sh` sets restrictive file permissions, defaults to `BACKUP_RETENTION_DAYS=30`, writes SQL dumps excluding feedback-content tables, and prunes aged files only during successful backup runs (without dump-level encryption by the script itself or an exact to-the-hour 30-day guarantee). `application.yml` defaults to `INFO` logging in `logs/skillpilot.log` with 10 MB rotation and `max-history: 7`; actual deployed log, proxy, trace, snapshot, WAL, and

off-site limits remain operational facts to confirm separately and must never be conflated with active learner or online feedback deletion.

Key operating rules and residual risks:

- **Credentials in transit and support:** Live learning-session references (`Claude`, `OpenAI`, or legacy `sps_...`) and OAuth access/refresh tokens are bearer credentials — never share unredacted chats, screenshots, browser history, HAR files, or logs containing them, whereas canonical goal and card IDs are public technical identifiers.
- **Scope of file encryption and provider chats:** Class-export encryption protects the downloaded file at rest and detects tampering, without encrypting live `skillpilot_classes` browser storage or securing an unlocked device. Learners must be guided not to type or upload personal data into the AI chat, which the provider processes under its own terms and which SkillPilot deletion cannot erase.
- **Truthful release gating:** Visible Session remains rollback-only, and OpenAI candidate code or Claude beta availability never establishes host acceptance on another client. Before any provider release or cutover, verify privacy notices, provider disclosures/eligibility, deployed retention/logging/OAuth profiles, real revocation behavior, and remaining processor/backup confirmations without weakening authentication or retaining unnecessary text.

2 OpenAI MCP: Authorized Client Binding

Binding security architecture for the OpenAI MCP coach

Original baseline: 31 July 2026 · Reconciled with the permanent neutral plugin decision of 19 October 2026 (`skillpilot-coach` on `https://mcp-openai.skillpilot.com/mcp`, superseding the earlier `https://mcp-coach-v1.skillpilot.com/mcp` endpoint ahead of directory submission while retaining OAuth issuer `https://skillpilot.com/api/openai/v1`). Identity, cutover, and lifecycle rules are governed by the Versioning and Lifecycle Plan and the Stable Origin Runbook.

2.1 Protection Objective

The permanent SkillPilot ID remains the learner-held secret that guards durable learning state. This chapter addresses an orthogonal security objective:

> An unauthorized third-party client must not be able to invoke learner-specific OpenAI V1 MCP APIs, even if it knows both the public server contract and a valid `learningSessionId`.

This boundary leaves the SkillPilot ID as well as learning-goal, mastery, curriculum, and coach semantics unchanged.

2.2 Four Independent Proofs

Proof	Purpose	Binding mechanism
SkillPilot Server	Confidentiality and server authentication	Standard HTTPS/TLS terminated at the dedicated reverse proxy.
OpenAI Connector	Only hosted ChatGPT connections signed by Open AI's CA chain reach the production edge	mTLS validated against OpenAI Root CA, Connectors intermediate, `clientAuth` EKU, and exact SAN `mtls.prod.connectors.openai.com`.
Authorized MCP App	Only the explicitly permitted client receives usable OAuth tokens	Pre-registered confidential OAuth client (`client_secret_basic` + exact callback allowlist + PKCE `S256`) for the hosted directory lane; separate strictly pinned desktop CIMD profile (`S256`) without fallback for native desktop.
Active Learning Session	The concrete chat addresses strictly the learner chosen at WebGUI launch	Fresh 24-hour `learningSessionId` minted on every **Lernen starten** / **Start learning** click and required on every domain MCP tool call.

Every domain MCP call must satisfy cumulatively:

```
valid client- and resource-bound OAuth access token
AND
valid explicit learningSessionId
AND (in production hosted enforce mode)
valid OpenAI connector client certificate
```

OAuth authenticates and authorizes the App; the learning session selects the learner. Neither proof substitutes for the other.

2.3 Binding OAuth Client Profile

For the hosted ChatGPT directory lane (**SkillPilot Coach**, technical name `skillpilot-coach`), exactly one pre-configured confidential OAuth client is provisioned:

- fixed author-chosen `client_id` and high-entropy random `client_secret`;
- byte-exact callback URL registered in the ChatGPT App management console;
- token-endpoint authentication exclusively via `client_secret_basic`, paired with Authorization Code + PKCE `S256` only;
- exact public Resource/Audience `https://mcp-openai.skillpilot.com/mcp` (no dual audiences alongside the retired `mcp-coach-v1` origin);
- separated read and write scopes, short access-token TTL, rotating refresh tokens, and explicit revocation.

The client secret exists exclusively in two protected locations: (1) the confidential OAuth settings of the ChatGPT App and (2) the protected SkillPilot server environment — never in repositories, browser bundles, launch prompts, chats, MCP tool arguments, responses, or logs.

In addition, a **strictly separate, byte-pinned public CIMD profile** (`7yTcOAOqrumX` client/callback on the same neutral resource with PKCE `S256`; see Native CIMD Runbook) is authorized for native ChatGPT Desktop integration. That profile never acts as a fallback from failed confidential authentication. Open Dynamic Client Registration (DCR), unpinned CIMD documents, arbitrary `none` clients, and `private_key_jwt` are excluded from production.

2.4 OAuth Flow and Resource Validation

The Authorization Server accepts an authorization-code flow only when (1) `client_id` matches an explicitly pinned profile, (2) `redirect_uri` matches its byte-exact allowlist, (3) PKCE `S256` is used, (4) `resource` equals the active MCP resource, and (5) only permitted scopes are requested. At the token endpoint, the confidential client must additionally prove possession of its secret via HTTP Basic.

On every MCP request, the Resource Server verifies token signature or active introspection, issuer (`https://skillpilot.com/api/openai/v1`), expiry and revocation, exact audience/resource, exact client ID, required read or write scope, and the explicit valid `learningSessionId`. A valid OAuth token can neither select a learner nor create a session; the OAuth subject is never an identity source or fallback.

2.5 Learning Session as Separate Application Binding

Per the OAuth and Learning-Session Contract, clicking **Lernen starten** / **Start learning** mints a fresh random ID on every start (valid for a fixed 24 hours without extension, stored server-side solely as an HMAC hash linked to the learner). New domain operations require at least `PT1H` remaining lifetime (exact `PT1H` is valid). An already committed write may replay its stored result without re-mutating state only when at least `PT1H` remains and tool name, canonical arguments, `clientRequestId`, pinned workflow/curriculum availability, and canonical learner revision all match.

`SESSION_REQUIRED`, `SESSION_RENEWAL_REQUIRED`, and `SESSION_VERSION_UNAVAILABLE` are session boundaries rather than OAuth errors: the coach outputs `instruction` verbatim (or selects the exact localized entry from `instructions`, appending `startUrl` only if absent) without domain output, retry on the expired session, or OAuth re-authentication. The learner starts a

fresh session in a new chat from the first-party WebGUI while the permanent SkillPilot ID stays server-side.

2.6 mTLS at the MCP Edge

Hosted ChatGPT connectors present an OpenAI-managed client certificate during TLS handshake. The dedicated MCP virtual host trusts solely the published OpenAI Root CA and Connectors intermediate (without pinning rotating leaf certificates), while a loopback-bound verifier checks certificate chain, `clientAuth` EKU, and exact SAN `mtls.prod.connectors.openai.com`:

- **`observe`:** Valid certificates classify as `VERIFIED`; connections without a certificate pass provisionally to the unchanged OAuth layer and count solely as `OBSERVE_NO_CERT` (for controlled integration and native desktop verification, not a hosted publication state). Any invalid presented certificate is rejected unconditionally.
- **`enforce`:** External `/mcp` calls require `VERIFIED`. Only direct loopback socket peers may exercise the unchanged OAuth path locally as `LOCAL_OPERATOR`; internet headers, query parameters, and `X-Forwarded-For` never create a bypass.

Protected-resource metadata and domain-verification challenges remain reachable without a client certificate. mTLS authenticates OpenAI connector infrastructure and substitutes for neither OAuth nor `learningSessionId`.

2.7 Fail-Closed Startup Invariants and Secret Lifecycle

When the OpenAI V1 provider is enabled, server startup fails closed if any invariant is missing or inconsistent: dedicated high-entropy `SKILLPILOT_SIGNING_SECRET` with at least 32 non-whitespace characters (no placeholders or defaults, never logged), consistent pinned client configuration (non-empty client ID, strong client secret of at least 32 — preferably 48+ — random bytes for `client_secret_basic`, exact HTTPS callback allowlist, PKCE `S256`, exact HTTPS resource and OAuth endpoints), exact mode agreement across root nginx, backend, and internal classification, and disabled open DCR / unpinned profiles. Port `8787` must remain closed to the public internet; the reverse proxy terminates TLS and forwards only approved routes to `/internal/openai/v1/mcp`.

The production OAuth client secret is generated outside the repository, stored identically only in ChatGPT App settings and the SkillPilot server environment, never echoed in status or error endpoints, and rotated immediately upon suspected compromise (with token revocation and App re-authorization). Rotating `SKILLPILOT_SIGNING_SECRET` intentionally invalidates active learning sessions as a planned session cutover. Rotating OAuth client credentials never preserves or reactivates sessions as an identity substitute.

2.8 Acceptance, Negative Tests, Residual Assumptions, and References

Before release, objective negative tests must fail closed for: token requests missing authentication, carrying a wrong secret, or naming a foreign client ID; unpinned auth methods (`client_secret_post`, open DCR, unpinned `none`/CIMD/`private_key_jwt`); mismatched callback URL or missing/invalid PKCE; missing/wrong resource or scope; revoked/expired tokens; valid OAuth without `learningSessionId` and valid session without OAuth; unknown, tampered, revoked, or expired sessions; direct bypass to port `8787`; and in `enforce` mode any external `/mcp` request lacking a valid OpenAI client certificate, presenting a wrong chain/SAN/EKU, or spoofing forwarding/classification headers. Positive end-to-end verification runs through the configured

App (public positive `enforce` mTLS proof can only be executed by the real hosted ChatGPT connector holding OpenAI's private key).

Residual assumptions: This architecture does not protect against compromise of the Skill Pilot host, reverse proxy, or ChatGPT App configuration, learner-side loss of a SkillPilot ID or live session link, leak of the production client secret, or deliberate sharing of a valid access token paired with a live session — mitigated via secret rotation, short token TTLs, fixed 24h session lifetimes, revocation, minimal scopes, and redacted logging.

Normative OpenAI references: Client identification · Client registration & authentication · Resource/Audience parameter · Mutual TLS (mTLS).

3 EU AI Act: Compliance Baseline for SkillPilot

Documented compliance baseline and live inventory

Baseline date of this assessment: **29 July 2026** (documented repository compliance record; does not constitute legal advice). The machine-readable AI transparency inventory is deterministically maintained as assets evolve; for the subsequent September 2026 provider work order (active Claude learning beta ahead of focused ChatGPT acceptance), see the Claude-first release strategy and 1.

3.1 Regulatory Calendar

Track	Effective date	Relevance for SkillPilot
Article 50 AI Act	**2 August 2026**	Transparency duties for direct AI interaction and specific synthetic content. The transition period until **2 December 2026** applies strictly to Art. 50(2) for generative systems placed on the market before 2 August 2026.
Annex III / Art. 6(2)	**2 December 2027**	Chapter III Sections 1–3 apply from this date to Annex III high-risk systems. Legally separate from Article 50.
Annex I / Art. 6(1)	**2 August 2028**	Later product-safety regulation track, not primary for SkillPilot's current scope.

Legal foundations are the EU AI Act (EU) 2024/1689, Amending Regulation (EU) 2026/1744 (in force since 27 July 2026), and the European Commission's Guidelines on Article 50 transparency obligations adopted on 20 July 2026 (interpretive guidance rather than statutory text).

3.2 Verified Product Inventory

The continuously verified machine-readable transparency inventory serves as technical evidence; as of 29 July 2026, the repository audit established the following facts and evidentiary boundaries:

Asset class	Repository fact on 29 July 2026	Evidentiary boundary
Goal visualizations	`app/public/assets/goal-visualizations/` holds 1,494 images (1,484 JPG, 10 PNG), matching the canonical source set.	File counts and formats prove neither initial public availability prior to a statutory cutoff nor deepfake status.
Image provenance	A byte scan detects a `c2pa` marker in 1,491 images and `Created by Google Generative AI` in 1,484 images; prompt sidecars frequently cite Gemini / Nano Banana Pro.	No cryptographic validation of C2PA signatures, certificate chains, or manifest integrity — provenance indicators rather than validated Art. 50 proof.
Narrative illustrations & comics	Five semantic collections comprising 24 runtime files (12 bilingual quickstart scene files plus 12 language/format variants across four comic sets including `docs/comic4/`).	File count does not equal distinct creative works; generation provider and first publication dates are not exhaustively evidenced.

Asset class	Repository fact on 29 July 2026	Evidentiary boundary
Whitepaper raster figures	`docs/whitepaper/` and `app/public/whitepaper/` bind 24 PNG files (2 shared with `docs/comic4/`), yielding 46 unique runtime images outside goal visualizations.	Mixed or partially evidenced provenance; inventorying explicitly does not imply all 24 whitepaper figures are AI-generated.
Learning coach (29 July 2026 baseline)	Two ChatGPT connection paths at baseline; the external AI provider executes the dialogue model while Skill Pilot supplies context and tools.	The Claude OAuth/MCP lane paused on 29 July 2026 was subsequently promoted to the active learning beta by Product Owner decision in September 2026 (see 1).
Texts and flashcards	Governed by domain and editorial QA lanes without uniform per-item provider generation logs.	No blanket claim that every text is purely AI-generated; the public-interest text exception requires demonstrable human review and editorial responsibility per case.

Audio Inventory

- **`app/public/audio/intro-de.m4a`** — Duration `1095.540680 s`, encoder tag `Google`, SHA-256: `11af439ea19741567bd797d7b05314c2a1f5c416413beed5f781e2fd7d1cf196`
- **`app/public/audio/intro-en.m4a`** — Duration `892.482177 s`, encoder tag `Google`, SHA-256: `e4d53b9f15932b7b1ce863cf06a48b36323a8a1efa09f0238fe96407a471a923`
- **`docs/notebooklm/SkillPilot_macht_Bildung_zum_Navigationssystem.m4a`** — Duration `801.413515 s`, encoder tag `Google`, SHA-256: `267aafede53d9849a980cccd0304abc47c7069edee5ecac5ab8263c06c594876`
- **`docs/notebooklm/Skill_Graphs_Tame_the_Smart_Liar.m4a`** — Duration `1134.410884 s`, encoder tag `Google`, SHA-256: `1542430f14f1beee4738dfa5b655b7fc1d819503a95e31dc1f02dbd580da43f8`

Published audio introductions are not byte-identical to the two archived NotebookLM recordings. An encoder tag or UI icon proves neither the specific synthesis service nor voice provenance; accordingly, the precautionary provider-neutral player notice (*"This audio introduction contains AI-generated voices"*) does not substitute for full segment and voice provenance records.

3.3 Article 50: Classification and Implementation

- **Role boundaries:** Art. 50(2) attaches to the provider of the generative AI system placed on the market (not determined by model hosting alone). SkillPilot treats itself precautionarily as a **deployer** for its integration and publication, re-evaluating potential **provider** obligations whenever placing a system on the market under its own name, substantially modifying it, or changing its intended purpose. Role assessments are documented per system and content class.
- **Direct interaction:** Coach dialogues run inside the selected provider's chat interface; additionally, the SkillPilot launch surface explicitly names the AI assistant and warns of potential mistakes rather than relying solely on the "obvious to a reasonably well-informed person" exception (applying heightened clarity standards for minors).
- **Four-part cumulative deepfake test (Art. 3(60)):** Content constitutes a deepfake only when cumulatively (1) AI-generated or AI-manipulated, (2) image, audio, or video, (3) appreciably resembling existing persons, objects, places, entities, or events (even without a named real-world individual), and (4) presented in a context appearing falsely authentic

or truthful to a person (applying stricter intelligibility for minors). Schematic didactic illustrations regularly lack element (4) as a class assumption — without granting a blanket exemption or claiming a global image badge. Both audio introductions display an explicit AI-voice notice directly on the player, since a standalone legal subpage alone would not satisfy Art. 50(4) where applicable.

- **Legacy stock and five-step gate for post-cutoff assets:** Lacking exhaustive pre-2 August 2026 first-publication records, SkillPilot does not claim legacy grandfathering, relying instead on substantive Art. 50 review and contextual player/coach notices. Any content AI-generated, AI-manipulated, or materially repurposed after 2 August 2026 must pass five pre-release steps: (1) determine actor and role, (2) evaluate content type and deepfake criteria, (3) for public-interest text, document human review and editorial responsibility or disclose at the content, (4) where required, label clearly and accessibly no later than first exposure, and (5) record decision, version, date, and reviewer role in the inventory. Existing C2PA structures are preserved on deployment without substituting for human-readable disclosure.

3.4 Annex III: Preliminary High-Risk Boundary

System boundaries and AI-system status under Art. 3(1) are assessed separately for the learning coach, deterministic state tools, and future grading modules based on intended purpose, documentation, marketing, and concrete deployment context (informed by non-binding examples from the official AI Act Service Desk on Education):

Deployment context	Preliminary classification
Self-directed learning outside an educational or vocational institution without formal grading, admission, or certification effect	Currently **not** treated as Annex III No. 3(b); AI feedback remains a soft study aid with formal decisions excluded.
Deployment by an educational or vocational institution where AI outputs evaluate learning outcomes or steer the learning process based on such evaluation	**Treated precautionarily as high-risk under Annex III No. 3(b)** unless a documented counter-assessment establishes otherwise; requires a dedicated release gate prior to activation.

Six triggers activating the high-risk review lane: (1) contracts, pilots, or marketing targeting schools, universities, or vocational institutions, (2) using AI outputs or mastery values for grades, certificates, admission, promotion, or graduation, (3) institutionally mandated learning-path steering driven by AI evaluation, (4) new automated scoring, ranking, or prediction features, (5) profiling within the meaning of Art. 4(4) GDPR inside an Annex III system (personalization alone does not automatically constitute profiling), and (6) placing on the market under SkillPilot's name, substantial modification, or change of intended purpose.

An exemption under **Art. 6(3)** never applies where profiling occurs and requires cumulatively no significant risk to health, safety, or fundamental rights, no material influence on decision outcomes, plus at least one of four **alternative** conditions: (1) narrow procedural task, (2) improving the result of a previously completed human activity, (3) detecting patterns or deviations without replacing proper human assessment, or (4) preparatory task for an Annex III assessment — backed by pre-market documentation and Art. 49(2) EU database registration by the provider.

3.5 Evidence Status, Release Gates, and Ongoing Maintenance

Measure	Documented status (29 July 2026 baseline)
Concise AI transparency section in `/legal` (DE/EN)	Implemented in repository without adding a new consent or liability-waiver gate or forcing re-consent.
General liability-notice gate	Kept technically and legally separate from Article 50 and unchanged (mandatory statutory rights remain unaffected).
Contextual AI notices on audio player & coach start (DE/EN)	Implemented in repository with passing local render, copy, and accessibility smoke checks.
Machine-readable inventory & CI/deploy gates	`check:ai-transparency-inventory`, `check:ai-transparency-artifact`, and deploy readiness checks verify UI, inventory, and built artifacts.
Legacy grandfathering & cryptographic C2PA validation	Neither is claimed; C2PA markers are documented strictly as byte structures.
Institutional high-risk release gate	Future trigger prior to any institutional pilot or release; not a deficit of the voluntary self-study product.

Maintenance triggers: Both this dossier and the JSON inventory are re-evaluated upon any provider or model change, new asset class, realistic voice or human likeness, change to coaching or evaluation mechanics, institutional distribution, or material change of intended purpose, alongside role-appropriate AI literacy support under Article 4 AI Act.