

Legitimate Interest Assessment (LIA)

Art. 6(1)(f) GDPR — Three Processing Activities — CIATA DeepResearch

Version: 1.8 **Assessment Date:** 2026-07-10 **Last Revised:** 2026-09-27 **Next Review Due:** 2027-07-10

Revision 1.8 changelog (2026-09-27): Consistency Checker added (issue #172): runs in parallel with the Paragraph Checker on qwen3.8-27b; no new model, processor or data category; two additional LLM calls per report. Sub-processor table updated: Consistency Checker added to the qwen3.8-27b agent list.

Revision 1.7 changelog (2026-09-27): Language Guard moves from qwen3.6-chat-35b to gemma4-chat-26b with reasoning disabled (issue #167). qwen3.6-chat-35b is no longer used in any agent; three models remain in use (gpt-oss-120b, qwen3.8-27b, gemma4-chat-26b). Failed repairs now remove leaked characters in code. Sub-processor table updated: AKI.io language guard row changed. Same data categories, same recipient (AKI.io GmbH, Frankfurt).

Revision 1.6 changelog (2026-09-26): Transactional email moves from Brevo SAS to Scaleway SAS (issue #152, live 2026-09-26): Brevo SAS replaced by Scaleway SAS in the sub-processor table and safeguards checklist; Brevo noted as former processor (active until the switch; rollback path only; no Art. 28 DPA was executed). Report-ready notification email (issue #159, live 2026-09-26) adds report title and a short report summary to the data categories received by the active email provider.

Revision 1.5 changelog (2026-09-13): Model routing revised for run time (issue #135): the Relevance Gate moves from qwen3.6-chat-35b to gpt-oss-120b (same verdicts on 192 replayed results, 4.5x faster) and the Extractor from qwen3.6-chat-35b to gemma4-chat-26b with reasoning disabled (5x faster, more and better-grounded findings on 40 real pages). gemma4-chat-26b is a NEW model identifier — a fourth model now processes customer data (PII-filtered page excerpts) — so the DPA scope must name four models; qwen3.6-chat-35b remains in service for the Language Guard only. Same data categories, same recipient (AKI.io GmbH, Frankfurt).

Revision 1.4 changelog (2026-09-12): Post-synthesis verification restructured (issue #133): the Fact Checker — a whole-report check plus a per-citation alignment check, both on gpt-oss-120b — is replaced by the Paragraph Checker, and the Report Patcher is given its own routing; both now run on qwen3.8-27b. Same data categories (the draft report and the run's findings), same recipient (AKI.io GmbH), no new model identifier: the three-model set and the DPA scope instruction are unchanged.

Revision 1.3 changelog (2026-09-10): Synthesis model corrected from mistral4-119b to qwen3.8-27b, in service since 2026-08-23 (commit eb2a5b1); the Table Repair agent, added 2026-09-10, runs on that same model and introduces no new recipient. DPA scope instructions corrected in step.

Revision 1.2 changelog (2026-08-11): (1) Corrected page-content persistence claims in §A.2 items 7–8 and §Art. 21: PII-filtered extractor input is stored in state.json for the 30-day run TTL; raw unfiltered content is never persisted; false "in-memory only" and "not technically possible to un-process" assertions removed. (2) Art. 21 remedy updated: DELETE /v1/research/{job_id} removes the run directory and all stored page text; this is the correct retroactive remedy for a valid objection. (3) Robots.txt compliance scope qualified to CIATA's own direct HTTP fetches; Staan server-side path noted explicitly in §A.2 item 1, balancing item 2, and safeguards checklist. (4) Cloudflare/bot-wall interstitial retry behaviour added as §A.2 item 6. (5) Safeguards checklist updated accordingly. **Legal Basis:** Article 6(1)(f) GDPR — Legitimate Interests **Reference:** EDPB Guidelines 06/2020 on Legitimate Interests (adopted 8 October 2024) **DPIA Cross-Reference:** DPIA-DR-2026-01 (research.ciata.io)

Controller: CIATA GmbH, Arnimallee 7, 14195 Berlin, Germany HRB 266084 B, Amtsgericht Charlottenburg Managing Director: Prof. Dr. Tim Landgraf Data Protection Contact: privacy@ciata.io

- ⚠ **TO CONFIRM:** DPO designation — confirm whether a formal DPO has been appointed (Art. 37 GDPR) or whether one should be voluntarily designated. If appointed, add name and contact details. This item must be resolved before this LIA is relied upon for regulatory communications.

⚠ **DECISION NEEDED:** External legal review — this LIA has not yet been reviewed by external legal counsel. An independent review is recommended before relying on it for DPA submissions or regulatory enquiries.

Scope of This Assessment

This LIA covers three distinct processing activities for which CIATA GmbH relies on Art. 6(1) (f) as its legal basis. Each activity is assessed separately against the three-step test (purpose / necessity / balancing) as recommended in EDPB Guidelines 06/2020.

Activity	Short description
A — Web Retrieval	Programmatic fetching of publicly accessible web pages to extract factual information for AI-generated research reports
B — Technical and Access Logs	Server-side logging of HTTP requests (IP, timestamp, endpoint, status) for security monitoring and service stability
C — Internal Staff Access	Access by authorised CIATA staff and contracted sub-processors to operational systems to run, monitor, and secure the service

ACTIVITY A: Web Retrieval for Research

A.1 Purpose Test

What is the legitimate interest? To provide AI-powered deep research services by programmatically reading publicly accessible web pages and extracting factual information relevant to the user's research question.

Whose interest? Both CIATA GmbH (as commercial operator) and service users (researchers, businesses, academics) who receive comprehensive, cited research reports. A broader societal interest in democratising research capabilities previously reserved for large organisations also applies.

Is the interest genuine and not trivial? Yes. AI-assisted research is a commercially and socially recognised activity. Without programmatic web reading, no meaningful research output is technically possible at the service's required scale and cost.

A.2 Necessity Test

Is web retrieval the least intrusive means of achieving the purpose?

Alternative	Assessment
Manual research by humans	Not scalable; prohibitively expensive for comprehensive topic coverage
Licensed data APIs only	Insufficient coverage; the majority of publicly available knowledge resides only on the open web
Search engine snippets only	Too shallow; snippets lack the context required for reliable fact extraction
Cached or archived content	Stale data; incomplete and inconsistent coverage
Automated web retrieval	Necessary and proportionate — reads only publicly published content

Data minimisation measures in place:

- robots.txt compliance** — `is_allowed_by_robots()` (`src/page_reader.py`) is called before every direct page fetch **by CIATA**. Results are cached per domain with a 1-hour TTL. If `robots.txt` is unreachable (timeout, HTTP 4xx/5xx, connection error), the URL is **allowed by default** (graceful-allow fallback, consistent with RFC 9309 §4 practice). This check applies only to CIATA's own HTTP client; pages retrieved via the Staan `full_content` backend (item 9 below) are fetched server-side by Staan and do not pass through CIATA's `is_allowed_by_robots()` function. Robots.txt compliance on those paths is governed by Staan's own operational policies.
- Honest User-Agent** — all direct requests identify the crawler: `CIATA-DeepResearch/1.0 (+https://research.ciata.io; bot; contact: privacy@ciata.io)`

3. **PII filtering** — a regex-based filter (`src/pii_filter.py`) scrubs structured personal data patterns from page content **before** it is passed to the LLM. See §A.2 — PII filter scope below for the full list of covered and uncovered data types.
4. **Content truncation** — direct page fetches via `page_reader.py` are capped at **16,000 characters** in production (`page_max_chars`, set via the production `config.yaml` override; the code default in `src/config.py` is 8,000). The 16,000 value was verified from the running container's effective configuration.
5. **Rate limiting** — a minimum of 1 second between requests to the same domain is enforced (`_MIN_FETCH_INTERVAL`, `src/page_reader.py`).
6. **Bot-wall interstitial retry** — when CIATA's own HTTP client receives a response matching one of four known Cloudflare or bot-wall interstitial patterns (matched against title and body text capped at 200 characters), it waits 6 seconds and repeats the identical honest-bot request once, then gives up if the second response is also an interstitial. The same User-Agent is used throughout; no challenge is solved, no JavaScript is executed, and no cookie or authentication token is submitted. This behaviour applies only to CIATA's direct fetch path and is not applicable to the Staan server-side path.
7. **No re-publication** — scraped content is not stored or republished. Only extracted factual findings with source citations appear in the final report.
8. **PII-filtered extractor input retained; raw page content discarded** — the PII filter (`src/pii_filter.py`) is applied before `source.page_text` is assigned (`src/engine.py`). What is written to the run's `state.json` is the filtered extractor input — the exact text handed to the extraction LLM, already capped at 16,000 characters — not the raw fetched content. Raw, unfiltered page text is never written to disk. The stored filtered copy is retained for the 30-day run TTL and is deleted in full when the run is deleted or expires. The diagnostic endpoint that exposes it (`GET /v1/admin/reports/{job_id}/sources/{i}/text`) requires **both** the admin role and ownership of the run (`_require_admin + _check_job_ownership`, `src/api/app.py`); no CIATA staff member can read another user's stored page text through this endpoint.
9. **Staan full_content flow** — when the production-default Staan backend is used with `full_content=True` (the code default in `src/config.py`), Staan (European Perspective SAS) fetches each page server-side and returns the content as markdown. This means page content passes through Staan's infrastructure **before** it reaches CIATA's own PII filter. The character truncation is applied by the CIATA engine after Staan returns the content (`src/engine.py`).

⚠ **TO CONFIRM (Staan-DPA-covers-full_content):** Confirm that the Data Processing Agreement with Staan / European Perspective SAS expressly covers the server-side page-fetch (`full_content`) flow, and not only the basic search index query. If the current DPA does not address this, it should be amended before this flow is relied upon in production.

PII filter scope — verified against `src/pii_filter.py`:

The automated PII filter **covers** the following data types:

Pattern	Label applied
Email addresses (RFC 5322 simplified)	[EMAIL]
International phone numbers (E.164 and German/Austrian/Swiss local formats)	[TELEFON]
German tax ID (Steuerliche Identifikationsnummer, 11 digits)	[STEUER-ID]
German social security number (Sozialversicherungsnummer, 12-char format)	[SOZIALVERS-NR]
Swiss AHV/AVS number (756 . xxxx . xxxx . xx)	[AHV-NR]
IBAN (all country codes, various spacings)	[IBAN]
Credit card numbers (Luhn-validated, 13–19 digits)	[KREDITKARTE]
IPv4 addresses	[IP-ADRESSE]
German-format postal addresses (street name + house number, German typography)	[ADRESSE]

The automated filter **does NOT cover**:

- **Personal names** (first name, surname, or any combination)
- **Non-DACH national identifiers** — UK National Insurance numbers, US Social Security numbers, French NIR, Dutch BSN, Polish PESEL, or any other non-German/non-Swiss national ID formats
- **Non-German postal address formats** — Austrian, Swiss, or any address format not matching German street-name typography

These limitations are acknowledged. They are addressed in the balancing test (§A.3) and the Art. 21 objection mechanism (§Art. 21).

A.3 Balancing Test

Controller and user interests: Strong commercial and societal interest in accessible research tools. Processing is limited to publicly available information. No special categories of data are intentionally targeted or sought.

Distinguishing two categories of data subject:

This activity may affect two distinct categories of data subjects whose interests must be weighed separately:

(i) Website publishers — organisations or individuals who have chosen to make content publicly available on the open web. Publishers have a reasonable expectation that publicly published content will be read by both human visitors and automated agents (search engines, academic crawlers, AI research tools). Publishers retain meaningful control over automated access via `robots.txt`, which this system respects. The residual impact on publishers is minimal.

(ii) Third-party individuals whose personal data appears on pages — persons named in news articles, listed in directories, mentioned in documents, or whose contact details appear on pages they did not author. These individuals did not directly publish their data for the purpose of automated AI processing, and their interests warrant separate and more careful consideration.

Impact analysis — third-party individuals:

Risk	Severity	Mitigation
Structured personal identifiers (email, phone, IBAN, national ID) transmitted to LLM	Medium	PII filter removes covered types before LLM processing (see §A.2)
Personal names or non-DACH identifiers reaching the LLM	Residual	Not covered by automated filter; LLM is designed to extract factual claims, not to catalogue named individuals
Incidental processing of Art. 9 special-category data	Residual	See §Art. 9 Analysis
Personal data persisted in research reports	Low	Reports cite sources but do not reproduce full page text; 30-day maximum TTL
Aggregation or profiling of individuals	Low	Service is designed for topic research, not individual profiling; no cross-job or cross-session aggregation
Re-identification via cited sources	Low	Citation links to the original public source, not to any CIATA-held dataset

Balance — Activity A:

The legitimate interests of CIATA and its users **outweigh** the residual impact on data subjects because:

1. Only publicly accessible content is processed; no authentication bypass, paywall circumvention, or access to non-public data occurs.
2. Publisher restrictions (robots.txt) are respected for every direct page fetch by CIATA; publishers can opt out of automated access. Pages retrieved via the Staan server-side path are subject to Staan's own operational policies (see §A.2, item 9).
3. A regex-based PII filter removes the most sensitive structured identifiers before LLM processing.
4. Raw, unfiltered page content is never persisted. A PII-filtered copy of the extractor input (capped at 16,000 characters per source) is stored in the run record for up to 30 days for diagnostic purposes; the diagnostic endpoint that exposes it requires both admin role and run ownership, so no CIATA staff member can read another user's stored page text. The stored copy is deleted when the run expires or is deleted on request. This represents a more limited safeguard than complete non-persistence of page content, and is weighed accordingly; the dual-gate access control is a compensating measure.
5. The service is a research tool, not an aggregation or profiling platform; individual-level outputs are incidental, not the purpose.

6. The known limitations of the automated filter (personal names; non-DACH IDs) are disclosed and are addressed through the Art. 21 objection mechanism.

A.4 Conclusion — Activity A

Processing under Activity A is **lawful under Art. 6(1)(f) GDPR**, subject to resolution of the open TO CONFIRM items above.

Art. 9 Special-Category Analysis

Pages retrieved in the course of a research job may incidentally contain data revealing health conditions, political opinions, religious or philosophical beliefs, trade union membership, genetic or biometric data, data on sexual orientation, or criminal convictions (Art. 9(1) GDPR). This is not intentionally sought; the service is designed for topic-based research, not the collection of sensitive personal information about named individuals.

Safeguards applied:

- The PII filter removes structured identifiers commonly co-occurring with sensitive personal data (email, phone, national ID).
- The LLM is instructed to extract factual claims relevant to the research question, not to catalogue personal attributes of identified individuals.
- Research reports and job state are deleted after 30 days (`DATA_RETENTION_DAYS`, `src/api/job_store.py`).
- No Art. 9 data is intentionally stored, indexed, cross-referenced, or used for any purpose other than answering the research question.

Position: The incidental encounter with Art. 9 data does not, in itself, require a separate legal basis where the data is publicly available and the processing is proportionate. CIATA does not rely on Art. 9(2)(e) (manifestly public data) as a primary justification but notes that the combination of safeguards (PII filter, transient processing, 30-day TTL, topic-research purpose) keeps the residual risk minimal. If the nature of a submitted research topic makes substantial processing of special-category data foreseeable (e.g., a medical or political research query), this should be reviewed case-by-case against the DPIA.

ACTIVITY B: Technical and Access Logs

B.1 Purpose Test

Legitimate interest: To detect, investigate, and respond to security incidents, unauthorised access attempts, denial-of-service attacks, and software errors; and to maintain the stability and performance of the service.

Whose interest: CIATA GmbH (service operator, security obligations), existing users (who benefit from a stable and secure service), and any third party whose data would be at risk in the event of an undetected breach.

Is the interest genuine? Yes. Server-side logging for security and operational purposes is a recognised necessity for any internet-facing service, and is addressed in ENISA guidance and Art. 32 GDPR security obligations.

B.2 Necessity Test

Access logs (recording HTTP method, endpoint, response status code, timestamp, source IP address, and User-Agent string) are the minimum technically necessary artefact for: - Detecting and reconstructing security incidents (intrusion detection, abuse, credential-stuffing) - Rate-limit enforcement and DDoS mitigation - Error diagnosis, performance monitoring, and capacity planning - Compliance with potential legal obligations to disclose breach details

No less intrusive alternative provides equivalent security value. IP addresses — the most sensitive element in access logs — are necessary for abuse detection, geo-blocking, and incident reconstruction. Pseudonymising or hashing IPs at ingestion time would defeat the primary security purpose.

Retention: Approximately 90 days, as disclosed in the Privacy Policy (§10). This is consistent with standard industry practice and ENISA recommendations for security logging.

B.3 Balancing Test

Data subjects: Users and visitors to research.ciata.io whose IP addresses and User-Agent strings are recorded in access logs.

Reasonable expectations: Users of any internet-facing web service reasonably expect that the operator maintains server-side access logs for security and operational purposes. This practice is disclosed in the Privacy Policy.

Impact: Low. Access logs contain technical metadata (IP address, timestamp, endpoint accessed, HTTP status code). They are not used for behavioural profiling, cross-service tracking, advertising, or any purpose other than security and operations. Retention is capped at approximately 90 days; older logs are deleted.

B.4 Conclusion — Activity B

Processing under Activity B is **lawful under Art. 6(1)(f) GDPR**. The legitimate security and operational interest is clearly established, processing is limited to what is necessary, and the impact on data subjects is minimal and proportionate.

ACTIVITY C: Internal Staff Access to Operate and Secure the Service

C.1 Purpose Test

Legitimate interest: To enable authorised CIATA staff and, where applicable, sub-processor personnel acting under instruction, to access operational systems — including job queues, research state files, access logs, user account records, and infrastructure components — as necessary to deliver, monitor, troubleshoot, and secure the DeepResearch service.

C.2 Necessity Test

Without authorised staff access to operational systems, CIATA cannot: - Respond to incidents or outages affecting users - Investigate and resolve bugs or data quality issues - Apply security patches, configuration changes, or capacity adjustments - Fulfil legal obligations, including data subject access requests under Art. 15 GDPR (which require staff to locate and extract relevant data)

Access is limited to the minimum necessary through role-based access controls, VPN/SSH key authentication, and the principle of least privilege.

C.3 Balancing Test

Data subjects: Users whose research queries, job metadata, email addresses, account information, and (where applicable) billing records may be accessed by staff in the course of operational work.

Reasonable expectations: Users of a cloud-hosted SaaS platform reasonably expect that authorised service-provider staff have access to operational systems. This is disclosed in the Privacy Policy.

Safeguards: - Access is limited to authorised personnel bound by contractual confidentiality obligations. - Production infrastructure access is authenticated via SSH keys; multi-factor authentication applies where supported. - Staff access actions are captured in audit logs (Activity B). - Sub-processors receive access only to the data necessary for their contracted function, under a signed DPA. - The report inspector's diagnostic endpoint (GET /v1/admin/reports/{job_id}/sources/{i}/text) requires both admin role and ownership of the run; there is no administrative override. No CIATA staff member can use this endpoint to read another user's stored page text. Staff visibility into research content is therefore limited to what can be inferred from logs and metadata, not the extractor input itself.

C.4 Conclusion — Activity C

Processing under Activity C is **lawful under Art. 6(1)(f) GDPR**. The operational necessity is clear, access is controlled and authenticated, and the impact on data subjects is proportionate to the purpose.

Art. 21 Right to Object — Individual Suppression Mechanism

Individuals who believe their personal data has been processed as part of a research report — for example, because their name, contact details, or personal circumstances appear in a cited source — have the right to object under Art. 21(1) GDPR on grounds relating to their particular situation.

How to exercise the right: Contact privacy@ciata.io, identifying the URL, research topic, or report in question, and describing the nature of the concern. CIATA will: 1. Assess whether compelling legitimate grounds for the processing override the individual's interests, rights, and freedoms. 2. If no such grounds are established, suppress the relevant URL from future research jobs and, where technically feasible, purge the individual's data from stored reports.

Scope and remedy for completed jobs: When a research job completes, the run record stored at `runs/{job_id}/state.json` contains: (a) extracted findings, (b) the final report draft, and (c) a PII-filtered copy of the extractor input for each source (capped at 16,000 characters), retained for diagnostic purposes. Raw, unfiltered page content is not stored. The full run record — including all stored page text — is deleted by calling `DELETE /v1/research/{job_id}`, which removes the run directory via `shutil.rmtree`. This is the appropriate Art. 21 remedy for completed jobs. CIATA commits to performing this deletion, and to suppressing the relevant URL or entity from all future research jobs, on a valid Art. 21 request. The extraction outputs that were transmitted to the LLM during the job cannot be recalled from the LLM provider's infrastructure once sent; this is the residual limitation that cannot be remedied by deletion of CIATA's own run record.

Transfer Impact Assessment — Non-Default US Search Backends

Two search backends exist in the source code (`src/searcher.py`) that would constitute EEA → US transfers if enabled:

- **DuckDuckGo** — `DuckDuckGoSearcher` (US-based service)
- **Google Custom Search Engine** — `GoogleSearcher` (US-based service)

These backends are **not enabled by default**. The production deployment uses the Staan EU-sovereign backend (`SEARCH_BACKEND=staan`, `src/config.py` default). However, if either US backend is activated via configuration (including staging or test environments), search queries — which may constitute personal data or reveal sensitive research interests — would be transmitted to US-based services, constituting a Chapter V transfer.

Before enabling either US backend in any environment, a separate Transfer Impact Assessment must be completed under Art. 46 GDPR (e.g., Standard Contractual Clauses with supplementary measures). This LIA does not cover those transfers.

⚠ **TO CONFIRM:** Confirm that SEARCH_BACKEND=staan (or equivalent) is pinned in the production environment configuration and that no production code path can silently fall back to a US backend.

Sub-Processor Disclosure

The following sub-processors receive personal data in connection with the activities covered by this LIA. The public-facing sub-processor table in the Privacy Policy is the authoritative disclosure; this section cross-references it for LIA purposes.

Sub-processor	Role	Location	Transfer basis	DPA status
AKL.io GmbH	LLM inference (gpt-oss-120b reasoning agents and relevance gate; qwen3.8-27b report synthesis, table repair, paragraph check, consistency check and patch; gemma4-chat-26b extraction and language guard, reasoning disabled). qwen3.6-chat-35b was used for the language guard until 2026-09-27.	Frankfurt, DE	EU jurisdiction; no-training commitment (intended – see status)	⚠ Required – not yet executed / not on file
Staan / European Perspective SAS	Web search index + server-side page fetch (full_content flow, default backend)	EU	EU jurisdiction	⚠ TO CONFIRM: DPA signed? full_content scope confirmed?
Hetzner Online GmbH	Cloud hosting (Hetzner Cloud, not dedicated)	Germany	EU jurisdiction	⚠ Required – not yet executed / not on file
Mollie B.V.	Card payment processing	Amsterdam, NL	EU jurisdiction (NL)	⚠ Required – not yet executed / not on file
Billie GmbH	B2B invoice / Net-30 BNPL (embedded in Mollie checkout; receives billing address, VAT number, Handelsregister for underwriting)	Berlin, DE	EU jurisdiction	Independent controller (BNPL via Mollie) – no Art. 28 DPA; Art. 13/14 disclosure
Qonto SAS	Invoice generation (called for every completed purchase; receives name, email, billing address)	Paris, FR	EU jurisdiction	Independent controller (licensed PI) – no Art. 28 DPA; Art. 13/14 disclosure
Scaleway SAS	Transactional email (Scaleway TEM, fr-par; replaces Brevo SAS as of 2026-09-26; Brevo remains as rollback path only, no personal data sent after the	Paris, FR (fr-par; EU)	EU jurisdiction; Scaleway states all TEM data is hosted and	⚠ TO CONFIRM: Scaleway standard DPA (version of 1 June 2024) to be accepted and filed at

Sub-processor	Role	Location	Transfer basis	DPA status
	switch; no Brevo DPA was executed)		processed within the EU	docs/compliance/signed_dpas/scaleway/; OI-16 in DPA Register.
Plausible Analytics	Web analytics (self-hosted at trc.ciata.io — CIATA-controlled infrastructure, cookie-free, no personal data)	CIATA infrastructure	Not a third-party sub-processor	N/A — no personal data transfer

✓ **CONFIRMED (2026-07-15):** The Hetzner Cloud production instance is **not** encrypted at rest (no full-disk/LUKS encryption). Accepted gap; compensating controls include bcrypt, TLS in transit, per-user access isolation, 30-day retention, and **encrypted off-site backups geo-redundant within the EU (Frankfurt → Helsinki)**. Enabling volume-level (LUKS) at-rest encryption before the hot-production cutover remains a recommended hardening step (DPIA action A1).

Retention Schedule

Data category	Retention period	Basis
Research reports and job state	30 days (DATA_RETENTION_DAYS, src/api/job_store.py)	Configurable; default 30 days
Server and access logs	Approximately 90 days	Privacy Policy §10
Payment transaction records	10 years	§147 AO (German commercial law)
JWT refresh tokens	7 days	Auth configuration
Unverified accounts	72 hours	Account management configuration

Safeguards Summary

Technical safeguards — Activity A (Web Retrieval)

- [x] robots.txt checked before every direct page fetch **by CIATA**; graceful-allow fallback if unreachable. Staan server-side fetches are governed by Staan's own operational policies (see §A.2, item 9)
- [x] Honest bot User-Agent with contact address (privacy@ciata.io)
- [x] Bot-wall/Cloudflare interstitial detection: single honest-bot retry after 6-second wait; URL abandoned if interstitial persists (§A.2, item 6)
- [x] PII regex filter applied before LLM processing (scope listed in §A.2)
- [x] Per-domain rate limiting (minimum 1 second between requests)

- [x] Content truncation applied (16,000 characters in production; 8,000 code default; configurable)
- [x] No persistent storage of **unfiltered** raw page content — the PII filter runs before any text is written to disk (`src/engine.py`)
- [x] PII-filtered extractor input (up to 16,000 characters per source) stored in `state.json` for the 30-day run TTL; the diagnostic endpoint (`GET /v1/admin/reports/{job_id}/sources/{i}/text`) requires both admin role **and** run ownership — no staff member can read another user's stored page text; deleted with the run record
- [x] Research reports and job state: 30-day maximum retention; early deletion available via `DELETE /v1/research/{job_id}` (Art. 17 / Art. 21 remedy)
- [x] EU AI Act Art. 50(2): visible "AI-generated" label + provenance HTML comment AND invisible VSRMark Unicode variation-selector watermark (`src/watermark.py`, applied in `engine` and `job_store`)
- [] C2PA content provenance: **planned, not yet implemented** — do not represent as a current safeguard

Technical safeguards — Activities B and C (Logs / Staff Access)

- [x] Access logs retained maximum approximately 90 days
- [x] Staff access authenticated via SSH keys; MFA where supported
- [x] Role-based access controls; principle of least privilege applied

Organisational safeguards

- [x] DPIA completed and documented (DPIA-DR-2026-01)
- [x] Privacy Policy published with clear disclosure of all processing activities and sub-processors
- [x] Data Protection Contact: privacy@ciata.io
- [x] Art. 21 individual objection mechanism: privacy@ciata.io (with honest disclosure of technical limitations)
- [x] Sub-processor DPAs required before any personal data transfer
- [] DPO designation — [⚠ TO CONFIRM](#) (see opening section)
- [] Staan DPA and Scaleway DPA — [⚠ TO CONFIRM](#). Scaleway standard DPA (version of 1 June 2024) to be accepted and filed (OI-16 in DPA Register). Brevo SAS was the former email processor; no Art. 28 DPA was executed with Brevo (wind-down in progress, OI-6 in DPA Register). Billie and Qonto act as independent controllers (Art. 13/14 disclosure, no Art. 28 DPA required) — see sub-processor table
- [] Staan DPA covers full_content flow — [⚠ TO CONFIRM](#)
- [x] Encrypted off-site backups, geo-redundant within the EU (Hetzner Frankfurt → Helsinki)
- [] Hetzner disk encryption at rest — **not enabled** (confirmed 2026-07-15); accepted gap with compensating controls; volume-level (LUKS) encryption is a recommended pre-launch hardening step (DPIA A1)
- [] Production SEARCH_BACKEND pinned to staan — [⚠ TO CONFIRM](#)
- [] External legal review — [⚠ DECISION NEEDED](#) (see opening section)

Overall Conclusion

Based on this assessment, the three processing activities of CIATA DeepResearch – (A) web retrieval for research, (B) technical and access logging, and (C) internal staff operational access – are **lawful under Article 6(1)(f) GDPR**. The legitimate interests of the controller and its users are not overridden by the rights and freedoms of data subjects, given the safeguards described herein.

This conclusion is conditional on resolution of the open items marked **⚠ TO CONFIRM** and **⚠ DECISION NEEDED** throughout this document. Until those items are resolved, this LIA should be treated as a working draft subject to amendment.

This LIA should be reviewed annually or whenever there is a material change to processing activities, sub-processors, the technical architecture, or applicable regulatory guidance (including future EDPB Guidelines 06/2020 updates).

Approved by: Prof. Dr. Tim Landgraf, Managing Director, CIATA GmbH **Date:** 2026-07-10 **Next Review Due:** 2027-07-10

Reference: EDPB Guidelines 06/2020 on the processing of personal data based on Article 6(1)(f) GDPR (adopted 8 October 2024). DPIA cross-reference: DPIA-DR-2026-01.