

CIATA DeepResearch — System Architecture & MVP Definition for Legal Review

Current Version: v1.14 · describes main at the commit that introduces this revision, 2,932 tests passing · Last updated: **2026-10-08**

Purpose: This document provides the law firm with a complete description of the system's functionality, data flows, architecture, and existing compliance measures. It is designed to enable a thorough assessment of GDPR and EU AI Act compliance.

Changes in v1.14 (2026-10-08):

- **Staan's DPA is on file, and it contradicts our zero-retention statement.** Staан keeps search queries for 25 months on OVH servers in France (DPA Appendix 1) and states that queries are not personal data and fall outside the DPA (§2). The DPA does not mention the `full_content` page fetch. The provider is European Perspective SAS, Paris, not "European Search Perspective GmbH". Staан's FAQ says an unnamed US fallback can supplement results, so some queries may leave the EU; the DPA does not list it. CIATA has asked Staан to switch it off for its account (issue #194). Question for counsel: can derived queries be personal data for us, and is Staан's DPA then sufficient?

Changes in v1.13 (2026-10-04):

- **Technical logs are kept for the stated period (issue #161).** The data table said "~90 days (Docker log rotation)", but no rotation is configured and a log lived only until the next deploy replaced the container. Each deploy now saves the container log to a compressed, root-only archive on the same server, deleted 90 days after the container started.
- **Research content leaves the server log (issue #164).** Research progress messages (questions, sub-questions, search queries, source titles) are no longer written to the server log: they go to a log file inside the run's own directory, which is deleted with the run after 30 days. The server log keeps operational events and warnings, with e-mail addresses hashed or redacted and web addresses cut to their host name.

Changes in v1.12 (2026-09-27):

- **Consistency Checker added (issue #172).** A new post-synthesis agent, the Consistency Checker, checks the finished report draft against itself, running in parallel with the Paragraph Checker on qwen3.8-27b with reasoning effort medium. Two checks run per report: framing sections (executive summary, any decision section, research gaps) checked statement by statement against the topic sections; and lines carrying numbers, dates or article/section references compared for the same quantity. Flags go to the

Report Patcher together with the Paragraph Checker's flags. No new model, no new processor, no new data category: the report draft already goes to AKI.io for the Paragraph Checker and the Report Patcher. Two additional LLM calls per report. §7 agent table and §7.1 Synthesis Model row updated. Can be disabled with `CONSISTENCY_CHECK=false`.

Changes in v1.11 (2026-09-27):

- **Language Guard model:** qwen3.6-chat-35b → gemma4-chat-26b (**issue #167**). The Language Guard (post-synthesis repair of stray non-target-language text) moves from qwen3.6-chat-35b to gemma4-chat-26b with reasoning disabled. qwen3.6-chat-35b is no longer used in any agent. Three models remain in use at AKI.io: gpt-oss-120b, qwen3.8-27b, gemma4-chat-26b. When a repair call fails, the leaked characters are now removed in code rather than kept. §7 (agent table), §7.1 (Language-Guard Model row, DPA row, [△](#) TO CONFIRM note, model change note) and §12.2 Q9 updated.

Changes in v1.10 (2026-09-26):

- **Transactional email provider: Brevo → Scaleway (issues #152, #159).** The email provider changes from Brevo SAS to Scaleway SAS (Scaleway Transactional Email, TEM). The provider is selected by `EMAIL_PROVIDER`; Brevo remains in the code as a rollback path and continues to hold sending logs of mails delivered before the switch (wind-down open item). §2.1, §3.1, §3.2, §5, §12.1 Q5, and §13 updated.
- **Report-ready notification email disclosed (issue #159).** From 2026-09-26 the system sends the account owner an email when a report is ready. The email carries the report title and a short summary (up to about 500 characters of the executive summary). Data categories in §5 updated accordingly.

Changes in v1.9 (2026-09-15):

- **Output caps and retries.** Each agent's cap on tokens generated per call is raised to a comfortable budget sized from measured usage, between 2,048 and 32,768 (§7, "Output cap" column). Before, a call that reached its cap was re-sent with double the cap; that retry is removed. A call that reaches its cap is returned flagged as incomplete, or fails and the stage's fallback applies, and the event is counted. Nothing changes in what data is sent or to whom; the number of requests per report falls, and the upper bound on tokens generated by one call rises.

Changes in v1.8 (2026-09-15):

- **The Report Patcher now rewrites per section for every report**, one call per section that holds a flagged line, instead of one whole-report call for reports under 40,000 characters. Each call sends one section of the draft rather than the whole draft; the agent, the model, the data categories and the recipient are unchanged. §4.1 and §7 updated.

Changes in v1.7 (2026-09-15, later the same day):

- **The Figure Sheet call is switched off** (`research.figure_sheet: false` in the deployed configuration; the code default is also off). It ran in production only between the two deployments of 2026-09-15. No call is made and no data is sent for it unless an operator sets `FIGURE_SHEET=true`. §4.1, §7 and §7.1 mark it as disabled.
- **No other change** to agents, models, recipients, data categories or retention.

Changes in v1.6 (2026-09-15):

- **The production report writer changed on 2026-09-15** (commit 466a41b). Reports are now written section by section: one `qwen3.8-27b` call per topic section, then one call for the title, summary and research gaps. This agent, the Section Writer, ran in production from that date before this document listed it. §4.1 and §7 now describe it.
- **Two further calls support it, both on** `gpt-oss-120b`. The Glossary call merges the entity names the findings use into one name each. The Figure Sheet call, added by the change that introduces this revision, labels quantities that two different findings state with different values. Both send the research question and excerpts of the findings, which the Outliner already receives.
- **The Paragraph Checker retries a failed batch once**, so the same cited paragraphs and findings may be sent a second time.
- **No new model, recipient or data category, and no change to how long data is kept.**

Changes in v1.3 (2026-09-10):

- **The synthesis model named in v1.2 is wrong.** `mistral4-119b` was replaced by `qwen3.8-27b` on **2026-08-23** (commit eb2a5b1), decided on citation integrity. Every document in this set named the old model until today. The four model identifiers now in service are `gpt-oss-120b`, `qwen3.8-27b`, `gemma4-chat-26b` (since 2026-09-13) and `qwen3.6-chat-35b`. **If a DPA was drafted from v1.2, it covers the wrong synthesis model** — the same failure v1.2 reported against v1.1.
- **A fourteenth agent, Table Repair**, was added on 2026-09-10. It rewrites a single comparison table whose column the writer could not fill, in place, leaving the rest of the report untouched. It runs on `qwen3.8-27b`, so it introduces no new model and no new recipient. §7.1 lists it.
- **The Extractor's token budget** rose from 2,560 to 5,120 (a latency fix: at 2,560 the model exhausted the budget reasoning and every page cost two requests). The Language Guard borrows the Extractor's configuration, so its per-line ceiling moved with it — §7.1 corrected.
- **No change to what data is sent, to whom, or for how long it is kept.**

Changes in v1.2 (2026-08-11) — the engine changed materially since v1.1:

- **Every AI model named in v1.1 was wrong.** `minimax-m2.5-230b` is no longer used anywhere; the pipeline now runs `gpt-oss-120b`, `mistral4-119b` and `qwen3.6-chat-35b`.
[Superseded by v1.3: the synthesis model became `qwen3.8-27b` **on 2026-08-23. Do not take the model list from this entry.]** §7.1 carries the full per-agent routing, and the DPA

scope note in §5/§7.1 now names all three. **If a DPA was drafted from v1.1, it covers the wrong processing.**

- **Two further agents** (Intent Classifier, Outliner) run on an outline-first writing route that is **enabled** in the deployed configuration. §7 and the §4.1 sequence diagram now describe them.
- **Report synthesis moved out of the iteration loop** — the report is written once, after the loop ends. The v1.1 sequence diagram showed it drafted on every iteration. §4.1 corrected.
- **Scraped page text is retained, in PII-filtered form.** v1.1 stated that page text was held "in memory only, never persisted". That was true when written and later stopped being true: a diagnostic feature began storing the extractor input. The PII filter now runs **before** that storage, so what is retained is filtered text, kept for the 30-day run TTL and readable only by an admin who also owns the run. §4.2, §6.1 and §8.2 corrected; the LIA and ZDR statements are corrected in step.
- **Bot-wall interstitial retry disclosed (§8.1).** The retrieval layer now detects Cloudflare-style interstitials and repeats the identical honest-bot request once. This is flagged for counsel's assessment at §12.1 (1a) because it sits near the "no circumvention" line.
- **Per-sub-question coverage verdicts and the "covered well" assessment were removed from reports** on 2026-08-02. §10 previously described report sections that no longer exist.
- **Three data categories were undisclosed** in v1.1 and are now in §4.2: the B2B billing profile, § 356(5) BGB withdrawal-consent records, and run provenance metadata. Two of them raise retention questions we cannot answer ourselves — see §12.1 (3b).
- **AI Act timing.** Annex III high-risk obligations have applied since **2026-08-01**. §6.2 and §12.2 Q7 are re-framed accordingly: the classification question is now overdue rather than forward-looking.
- **Robots.txt scope corrected.** The check runs on CIATA's own direct fetches. In the default configuration Staan retrieves most pages server-side and those do not pass through it.
- **Server-side token revocation** (§9.1) and the **no-administrative-override** property of job-scoped endpoints (§9.2) are now documented — both are safeguards that v1.1 did not claim.

Changes in v1.1 (2026-07-15):

- **Public legal pages are now live and bilingual (DE/EN):** Impressum (/impressum.html, § 5 DDG), AGB / Terms of Service (/terms.html, incl. the statutory Widerrufsbelehrung and Muster-Widerrufsformular) and the Privacy Policy (/privacy.html). This resolves the former open questions §12.11 (Impressum) and §12.12 (ToS), now reframed as review requests.
- The **English AGB and Impressum are convenience translations**; a governing-language clause makes the **German version authoritative** on any discrepancy. Counsel should treat the German text as binding and the English as a translation.

- The defunct **EU ODR platform link was removed** from the Impressum and AGB (the EU Commission discontinued the platform in 2025). The § 36 VSBG non-participation statement is retained, prefaced with a goodwill "contact us first" line.
 - **Processor DPAs are not yet executed / not on file.** DPA status throughout has been corrected from "in place" to "required — not yet executed" (see §5). This is an open pre-launch item.
 - Password reset / forgot-password flow shipped (§2.2 updated).
 - Two engineering changes with compliance relevance: a **content language guard** that prevents leaked non-target-language text in reports (and fails the job with a credit refund if a report is unusable), and a **worker-level account-status re-check** so a deleted account's queued or in-flight research is aborted rather than completed (GDPR Art. 17 erasure). See §6–§7.
-

1. Product Overview

1.1 What is CIATA DeepResearch?

CIATA DeepResearch is a **B2B SaaS product** that provides autonomous, AI-powered deep research reports. A user submits a natural-language research question (e.g., "What are the current UPS/power supply solutions for data centers in Germany?"), and the system autonomously:

1. Decomposes the question into sub-questions
2. Generates targeted search queries
3. Searches the web for relevant sources
4. Reads and extracts factual findings from public web pages
5. Synthesizes a structured, cited research report
6. Delivers the report as Markdown and/or branded PDF

The system is designed as a **research assistant**, not a decision-making system. All outputs are informational. No automated decisions with legal or similarly significant effects are made (Art. 22 GDPR).

1.2 Operating Entity

Company	CIATA GmbH
Address	Arnimallee 7, 14195 Berlin, Germany
Registration	HRB 266084 B, Amtsgericht Charlottenburg
Managing Director	Prof. Dr. Tim Landgraf
Privacy contact (no DPO designated — see DPIA §7 open action)	privacy@ciata.io
Product URL	https://research.ciata.io

1.3 Target Market

B2B customers in the EEA — primarily managers, consultants, and analysts who need comprehensive research reports on topics such as: - Market analysis and competitive intelligence - Regulatory landscape reviews - Technology scouting - Academic and scientific literature overviews

2. MVP Feature Set (Launch Scope)

2.1 Features Included in MVP

Feature	Description
User Accounts	Email + password registration, email verification (Scaleway TEM), JWT authentication
Research Engine	Autonomous multi-iteration research loop with convergence detection
Web UI (SPA)	Single-page app for submitting questions, tracking progress, viewing reports
Real-time Progress	Server-Sent Events (SSE) for live progress updates during research
Report Delivery	Markdown + branded PDF export with AI provenance metadata
Credit System	Per-research-run billing via credit packs (3 free credits on signup)
Payment Processing	Mollie B.V. (EU-native, Amsterdam NL) for credit pack purchases
API Access	REST API with per-user API keys for programmatic access
Data Deletion	GDPR Art. 17 — DELETE endpoint for individual research jobs + account deletion
Admin Panel	User management, system stats, job overview
Privacy Page	Public privacy policy at <code>/privacy.html</code>

2.2 Features Explicitly Excluded from MVP

Feature	Status
C2PA watermarking on PDF reports	Planned post-launch. Not the same as the VSRMark watermark, which is implemented and shipping (\$6.2) — VSRMark is a proprietary Unicode scheme, C2PA is the cross-industry content-credentials standard.
~~Multi-language UI~~	Now partial: all public legal pages (privacy / terms / impressum) are bilingual DE/EN ; app UI is English-first with DE/EN locale support
Team/organization accounts	Single-user accounts only
Custom domain deployment	Single deployment only
On-premise deployment option	SaaS only for MVP

Feature	Status
~~Password reset / "Forgot Password" flow~~	Now implemented — email-based reset flow (no longer excluded)

2.3 Credit Packs & Pricing (MVP)

Pack	Credits	Price (EUR)
Free tier (signup)	3	€0
Starter	10	€29
Professional	50	€99
Enterprise	200	€299

1 credit = 1 research run (question → full report).

3. System Architecture

3.1 Infrastructure Overview

```
graph TB
    subgraph "User Devices"
        Browser["Web Browser / API Client"]
    end

    subgraph "Hetzner Cloud – Germany 
        Caddy["Caddy Reverse Proxy<br/>(TLS termination)"]
        Docker["Docker Container<br/>(non-root user)"]

        subgraph "Application Container"
            FastAPI["FastAPI Application<br/>(Python 3.11)"]
            Worker["Research Worker<br/>(Thread Pool, max 10)"]
            SQLite["SQLite Database<br/>(users, API keys, credits)"]
            RunsDir["Local Filesystem<br/>./runs/ (state, reports)"]
        end
    end

    subgraph "EU Sub-processors"
        AKI["AKI.io LLM API<br/>Frankfurt, Germany "]
        Staan["Staan Search API<br/>EU  (Ecosia + Qwant)"]
        Mollie["Mollie B.V.<br/>Amsterdam, NL "]
        Scaleway["Scaleway TEM<br/>Paris, France "]
    end

    subgraph "Public Internet"
        WebPages["Public Web Pages<br/>(fetched server-side)"]
    end

    Browser -->|"HTTPS"| Caddy
```

```

Caddy -->|"HTTP"| FastAPI
FastAPI --> Worker
FastAPI --> SQLite
Worker --> RunSDir
Worker -->|"LLM inference calls"| AKI
Worker -->|"Search queries"| Staan
Worker -->|"Page fetches"| WebPages
FastAPI -->|"Payment sessions"| Mollie
FastAPI -->|"Transactional emails"| Scaleway

```

3.2 Technology Stack

Layer	Technology	Notes
Runtime	Python 3.11 on Docker (slim)	Non-root user (UID 1000)
Web Framework	FastAPI + Uvicorn	Async ASGI server
Database	SQLite (file-based)	No cloud DB; local file on Hetzner
Reverse Proxy	Caddy	Auto-TLS via Let's Encrypt
LLM API Client	OpenAI Python SDK	Points to AKI.io base URL
Search	Staan API (default)	Also supports DuckDuckGo, Google CSE
HTML Parsing	BeautifulSoup4	Extracts text from web pages
PDF Export	WeasyPrint	Markdown → HTML → PDF
Payments	Mollie Python SDK	EU-native payment provider
Email	Scaleway TEM API (email_service.py)	Transactional email (verification, password reset, purchase confirmation, report-ready notification); Brevo retained as rollback (EMAIL_PROVIDER=brevo)
Auth	JWT (python-jose) + bcrypt	Access + refresh token pattern
Containerization	Docker + Docker Compose	Single-container deployment

3.3 Deployment Details

- **Hosting:** Hetzner Cloud, Germany (ISO 27001 certified datacenter)
- **Disk encryption at rest: Not enabled** (confirmed 2026-07-15). The production `./runs` and `./data` Docker volumes are **not** encrypted at rest (no LUKS/dm-crypt or Hetzner volume-level encryption). This is a documented, accepted gap — do **not** state "encrypted at rest" anywhere.
- **Backups: Encrypted** off-site backups, stored geo-redundantly within the EU (Hetzner Frankfurt → Helsinki).
- **Compensating controls for the at-rest gap:** bcrypt-hashed passwords, TLS in transit, strict per-user application-level access isolation, automated 30-day retention cleanup, and the encrypted EU backups above. Enabling volume-level (LUKS/dm-crypt) at-rest

encryption before the hot-production cutover is a recommended hardening step (DPIA action A1).

- **TLS:** Caddy handles automatic TLS certificate provisioning
 - **Container security:** Runs as non-root user (UID 1000)
 - **Health checks:** Internal GET /v1/health every 30 seconds
 - **Persistence:** Two Docker volumes — ./runs (research data) and ./data (SQLite DB)
-

4. Detailed Data Flow

4.1 End-to-End Research Flow

```
sequenceDiagram
    participant U as User (Browser)
    participant API as FastAPI Server<br/>(Hetzner DE )
    participant DB as SQLite DB<br/>(Hetzner DE )
    participant Search as Staan Search<br/>(EU )
    participant Web as Public Web Pages
    participant LLM as AKI.io LLM<br/>(Frankfurt DE )
    participant Mollie as Mollie B.V.<br/>(NL )

    Note over U,API: Authentication Phase
    U->>API: POST /v1/auth/login {email, password}
    API->>DB: Verify credentials (bcrypt)
    API-->>U: JWT access + refresh tokens

    Note over U,API: Research Submission
    U->>API: POST /v1/research {question}
    API->>DB: Check credit balance
    API->>DB: Deduct 1 credit
    API-->>U: {job_id, status: "running"}

    Note over API,LLM: Research Engine Loop (Background Thread)

    rect rgb(240, 248, 255)
        Note over API,LLM: Step 1: Question Decomposition
        API->>LLM: "Decompose this question into sub-questions"
        LLM-->>API: List of 4-8 sub-questions
    end

    loop For each iteration (up to 10)
        rect rgb(255, 248, 240)
            Note over API,LLM: Step 2: Query Generation
            API->>LLM: "Generate search queries for gaps"
            LLM-->>API: 3-6 search queries
        end

        rect rgb(240, 255, 240)
            Note over API,Search: Step 3: Web Search
            API->>Search: Search queries (derived, not original question)
            Search-->>API: URLs + titles + snippets
        end
    end
```

```

    rect rgb(255, 240, 255)
        Note over API,Web: Step 4: Page Fetching
        Note over API: robots.txt check (respect Disallow)<br/>direct CIATA fetches
only – not Staan full_content
        Note over API: Per-domain rate limit (1 req/sec)
        API->>Web: HTTP GET (honest bot User-Agent)
        Web-->>API: HTML content
        Note over API: HTML → text extraction
        Note over API: Content truncation (max 16,000 chars in prod)
        Note over API: PII regex filter applied
    end

    rect rgb(240, 240, 255)
        Note over API,LLM: Step 5: Finding Extraction
        API->>LLM: "Extract facts from this page" (PII-scrubbed text)
        LLM-->>API: Structured findings (fact, relevance, confidence)
        Note over API: Quality gate: min relevance 0.4, min confidence 0.3
        Note over API: Deduplication: Jaccard similarity (0.80 threshold)
    end

    rect rgb(255, 255, 240)
        Note over API,LLM: Step 6: Coverage Evaluation
        API->>LLM: "Evaluate coverage of sub-questions"
        LLM-->>API: Coverage score per sub-question
    end

    Note over API: Check termination: coverage ≥ 85%? Stall? Token budget? IGR
saturation?
    end

    rect rgb(255, 255, 240)
        Note over API,LLM: Step 7 (post-loop): Report Synthesis – runs ONCE
        Note over API: The report is written once, after the loop ends.<br/>Per-iteration
drafting is opt-in and off in production.
        API->>LLM: Classify question intent (intent_classifier)
        LLM-->>API: Report shape
        API->>LLM: Propose topic sections, then file each finding under one (outliner)
        LLM-->>API: Outline + finding assignment
        API->>LLM: Merge entity names into one name each (glossary)
        LLM-->>API: Glossary
        Note over API: Figure sheet (figures) – disabled by default since v1.7;<br/>runs
only with research.figure_sheet set.
        API->>LLM: Write each topic section in its own call, then title, summary and gaps
(section_writer)
        LLM-->>API: Report (Markdown)
        Note over API: With research.write_per_section off, one synthesizer call<br/
>writes the whole outline instead.
        Note over API: Transcription guard – if the writer copied findings<br/>rather
than writing prose, regenerate. After repeated<br/>failures, fall back to the template
writer.<br/>state.synthesis_route records which writer produced the report.
    end

    rect rgb(240, 255, 248)
        Note over API,LLM: Step 8 (post-loop): Verification

```

```

API->>LLM: Check every cited paragraph against its citations and the finding pool
(paragraph_checker)
LLM-->>API: Issues (if any)
API->>LLM: Rewrite each section holding a flagged line, one call per section,
spliced back (report_patcher, only if statements were flagged)
LLM-->>API: Corrected passages
API->>LLM: Scope analysis
LLM-->>API: Topics missing or covered shallowly
Note over API: Language guard – repair stray non-target-language lines.<br/>If
unrepairable, fail the job and refund the credit
Note over API: Embed VSRMark watermark + AI provenance block
end

Note over API: Save final report to disk (./runs/{job_id}/)
API-->>U: SSE event: "done" + final report available

Note over U,API: Report Retrieval
U->>API: GET /v1/research/{job_id}/report
API-->>U: Markdown report with AI provenance metadata
U->>API: GET /v1/research/{job_id}/report.pdf
API-->>U: Branded PDF with AI disclosure + data processing footer

```

4.2 Data Categories & Retention

Data Category	Contents	Storage Location	Retention	Legal Basis
Account data	Email, bcrypt-hashed password, name	SQLite DB (Hetzner DE)	Until account deletion	Art. 6(1)(b) — Contract
Research questions	User-submitted text (max 4,000 chars)	./runs/{job_id}/state.json	30 days (configurable TTL), user-deletable	Art. 6(1)(b) — Contract
Generated reports	Markdown text with citations	./runs/{job_id}/report.md	30 days (configurable TTL), user-deletable	Art. 6(1)(b) — Contract
Research state	Sub-questions, findings, coverage scores	./runs/{job_id}/state.json	30 days (configurable TTL), user-deletable	Art. 6(1)(b) — Contract
Scraped web page text	Clean text extracted from public HTML, PII-filtered , capped at 16,000 chars per source. This is the exact text handed to the extractor; it is retained so that a zero-yield source can be diagnosed. Raw, unfiltered page text is never stored. Readable only through the	./runs/{job_id}/state.json (per source)	30 days (configurable TTL), user-deletable	Art. 6(1)(f) — Legitimate Interest

Data Category	Contents	Storage Location	Retention	Legal Basis
	report inspector, which requires both the admin role and ownership of the run — there is no administrative override, so no member of staff can read another user's stored page text through this path (see §9.2).			
Extracted findings	Facts with source URLs (PII-filtered)	<code>./runs/{job_id}/state.json</code>	30 days (configurable TTL)	Art. 6(1)(b) — Contract
B2B billing profile	Organisation name, given/family name, email, street and number, postal code, city, country, VAT-IdNr., commercial register number	SQLite DB (Hetzner DE), <code>company_profiles</code> table	While the account exists, then 90 days after account deletion (#193) — see the § 147 AO question below	Art. 6(1)(b) — Contract + Art. 6(1)(c) — Legal obligation (invoicing)
Withdrawal-consent records	<code>job_id</code> , <code>user_id</code> , consent version, locale, consent timestamp	SQLite DB (Hetzner DE), <code>withdrawal_consents</code> table	Not deleted with the 30-day run TTL — these evidence a § 356(5) BGB waiver and must outlive the run. Required minimum period to be confirmed by counsel.	Art. 6(1)(c) — Legal obligation (§ 356(5) BGB); Art. 6(1)(f) for defence of legal claims
Run provenance metadata	<code>schema_version</code> , <code>code_revision</code> (short git SHA), <code>agent_models</code> (per-agent model IDs), <code>config_digest</code>	<code>./runs/{job_id}/state.json</code>	30 days (with the run)	Operational metadata — not personal data; Art. 6(1)(b) as part of the run record
Payment data	Mollie payment IDs, credit pack metadata	Mollie B.V. (NL) + transaction refs in SQLite	10 years (§ 147 AO)	Art. 6(1)(b) + (c)
Technical logs	Request metadata, error traces	Container log, then a compressed archive on the same server (<code>/var/log/deepresearch/</code> , root-only)	At most 90 days: kept until the next deploy, then deleted 90 days after the container	Art. 6(1)(f) — Legitimate Interest

Data Category	Contents	Storage Location	Retention	Legal Basis
			started (issue #161)	
AI-generated content watermark	VSRMark invisible Unicode watermark embedded in report text (producer id, version, ai_generated flag, job_id, UTC timestamp)	./runs/{job_id}/report.md (encoded in carrier text)	30 days (same as report), user-deletable	Art. 6(1)(c) — Legal obligation (EU AI Act Art. 50(2))
JWT tokens	Short-lived access tokens in browser localStorage	Client-side only	Access: 60 min, Refresh: 7 days	Art. 6(1)(b) — Contract
API keys	Per-user keys (bcrypt-hashed prefix dr_)	SQLite DB (Hetzner DE)	Until revoked	Art. 6(1)(b) — Contract
LLM prompts/responses	Research prompts, LLM completions	Zero retention at AKI.io (asserted; DPA not yet executed — see §5)	Transient (request duration)	Art. 6(1)(b) — Contract
Search queries	Derived queries (not original question)	Staan, OVH France (Staan DPA on file — see §5)	25 months at Staas (Staan DPA, Appendix 1)	Art. 6(1)(b) — Contract

4.3 Automated Retention Enforcement

- **Background timer:** Runs every 24 hours, deletes run directories older than DATA_RETENTION_DAYS (default: 30)
- **On-demand deletion:** DELETE /v1/research/{job_id} — immediate deletion with ownership verification
- **Account deletion:** Soft-delete via DELETE /v1/auth/me (requires password confirmation) — clears PII, deactivates account
- **Expired unverified accounts:** Background cleanup every hour — deletes accounts where email verification expired (72h TTL)

5. Sub-Processors

Sub-Processor	Role	Location	Data Sent	DPA
AKI.io GmbH	LLM inference (AI model hosting)	Frankfurt, Germany DE	Research prompts, PII-filtered page excerpts	⚠ Required — not yet executed / not on file
European Perspective SAS (Staan)	Web search index (Ecosia + Qwant)	EU EU	Derived search queries (never original question)	⚠ Required — not yet executed / not on file

Sub-Processor	Role	Location	Data Sent	DPA
Hetzner Online GmbH	Infrastructure hosting (server, storage, encrypted EU backups Frankfurt → Helsinki)	Germany [D][E]	All application data (physical datacenter controls; disk not encrypted at rest — see §3.3)	⚠ Required — not yet executed / not on file
Mollie B.V.	Payment processing	Amsterdam, Netherlands [N][L]	Payment amounts, user references, billing address (if provided)	⚠ Required — not yet executed / not on file
Billie GmbH	B2B invoice / Net-30 BNPL (embedded in Mollie checkout)	Berlin, Germany [D][E]	Billing name, billing address, VAT-IdNr., Handelsregister number (required for underwriting); only transmitted when buyer selects Billie as payment method	Independent controller (BNPL) — no Art. 28 DPA; Art. 13/14 disclosure
Qonto	Invoicing — client invoice generated for every completed purchase	Paris, France [F][R]	Buyer name, email address, billing address, VAT-IdNr. (if provided), purchase amount	Independent controller (licensed PI) — no Art. 28 DPA; Art. 13/14 disclosure
Scaleway SAS	Transactional email (verification, password reset, purchase confirmation, report-ready notification)	Paris, France [F][R]	Email address; name (if given); verification link / reset link / purchase links; purchase details (credits, pack label, price, new balance, Mollie payment reference, purchase date); report title, report summary (up to ~500 chars of executive summary), report link, deletion date. ⚠ TO CONFIRM: retention period of TEM sending logs at Scaleway	⚠ Pending — DPA "Version of 1 June 2024" to be accepted and filed
Brevo SAS (formerly Sendinblue) — former processor	Transactional email until the Scaleway switch (2026-09-26); rollback only thereafter. Brevo holds sending logs of mails it delivered before the switch. Wind-down (account closure, log-deletion request) is a planned open item.	Paris, France [F][R]	Until the switch: email address, name, verification and reset links, purchase details; for report-ready mails sent before the switch also report title, summary, link and deletion date	⚠ DPA never executed; wind-down planned

△ **TO CONFIRM (classification): Billie GmbH** is reached via Mollie's checkout and makes its own BNPL credit decision → treated as an independent controller; no Art. 28 DPA required, Art. 13/14 disclosure provided in the privacy policy. Confirm classification with counsel.

△ **TO CONFIRM (classification): Qonto** is a licensed payment institution acting in its own name → treated as an independent controller for the invoices it issues; no Art. 28 DPA required, Art. 13/14 disclosure provided. Qonto is called for every completed purchase. Optionally confirm its role via dpo@qonto.com.

△ **OPEN ITEM (DPAs not executed):** None of the current processor DPAs (AKI.io, Staas, Hetzner, Mollie, Scaleway) is on file as executed. For Scaleway, the standard DPA is the "Version of 1 June 2024" (https://www-uploads.scaleway.com/DPA_2024_ENG_b0abb5cc26.pdf); it must be formally accepted and filed. Brevo's DPA was also never executed; close the Brevo account and obtain deletion confirmation of its sending logs as part of the wind-down. For each current processor, locate the signed AVV/DPA — or execute it if it was never formally accepted — and record the type, counterparty version, and execution date. Attach copies as annexes to this briefing. Until then, treat every "zero retention / DPA" statement in this document as an intended contractual basis, not a confirmed one.

Note on Plausible Analytics: The system uses a **self-hosted** Plausible instance at trc.ciata.io, running on CIATA-controlled infrastructure. Plausible is cookie-free and does not collect personal data. Because CIATA operates the analytics server itself, Plausible is **not** a third-party sub-processor and is not listed here. Disclosure appears in the public privacy policy at </privacy.html> §11.

PLANNED — operational telemetry via OpenObserve (not yet enabled). The codebase contains a metrics exporter (`src/perf.py`) that ships operational telemetry to an OpenObserve instance. It activates only when `OPENOBSERVE_API_KEY`, `OPENOBSERVE_ORG` and `OPENOBSERVE_STREAM` are all set; they are not set today, so **nothing is currently transmitted**. We intend to enable it, and are disclosing it here in advance so it can be assessed before it goes live rather than after.

What would be transmitted. Two record types, both structured, neither containing report or page content: - Performance records: function name, duration in milliseconds, success/failure status, and an error string where a call failed. - Counter records: a fixed metric name from a closed set defined in the source (e.g. `synthesis_findings_cited`, `interstitial_detected_url`, `job_failed_quality_gate`), a numeric value, and optionally the internal `job_id`, a locale code, a line count, or a short fixed reason code.

What would not be transmitted. No research question, no report text, no scraped page content, no source URLs, no email address, no user id, no IP address. Metric names ending in `_url` are labels only — no URL is sent with them.

Two points for counsel. (1) The `job_id` is an opaque per-run identifier that can be correlated back to a user through our own database, which we think makes these records pseudonymous personal data rather than anonymous — we would like that confirmed, and

told whether the intended purpose (service reliability and quality monitoring) is properly grounded in Art. 6(1)(f). (2) The performance record includes a raw error string taken from the exception. We do not deliberately put user data in exception messages, but we cannot categorically exclude that an error text quotes a URL or a fragment of processed content. Should we scrub or drop that field before enabling the exporter?

Hosting is undecided and it changes the answer. If the instance is self-hosted on CIATA infrastructure (the code default points at localhost:5080), OpenObserve is not a third-party sub-processor — the Plausible position above. If a hosted OpenObserve service is used instead, it becomes a processor requiring an Art. 28 DPA and, depending on the provider's region, a transfer assessment. **Action before enabling:** decide the hosting model, execute a DPA if third-party, add the disclosure to /privacy.html, and add the entry to the Art. 30 record and the sub-processor register.

[!IMPORTANT] **No data transfers outside the EEA.** All sub-processors operate exclusively within EU/EEA jurisdiction. When Staan is the configured search backend (default), the entire data pipeline is EU-contained.

[!WARNING] **Alternative search backends (DuckDuckGo / Google CSE):** The system also supports DuckDuckGo (US-based, privacy-first) and Google Custom Search (US-based) as search backends. For the MVP, **Staan is the default**. If a deployment operator chooses to configure DuckDuckGo or Google, this constitutes a data transfer to US servers. This is disclosed in the privacy policy and the in-app GDPR information panel.

6. Technical Compliance Measures

6.1 GDPR Technical Measures

Measure	Implementation	Code Reference
PII filtering	Regex-based scrubbing of emails, phone numbers, IBANs, tax IDs, social security numbers, credit cards, IP addresses, German postal addresses before content is sent to the LLM	pii_filter.py
robots.txt compliance	Every direct page fetch by CIATA checks the target site's robots.txt first; if crawling is disallowed the page is skipped. 1-hour cache TTL per domain. Scope limit: in the production-default configuration Staan retrieves most page text server-side (full_content), and CIATA's own fetch — and therefore this check — does not run for those results; Staan applies its own crawling policy. CIATA fetches directly only when Staan returns no content for a result.	page_reader.py L64-L113
Input sanitisation	User research questions are stripped of VSRMark variation selectors (U+FE00–FE0E, U+E0100–E01EF) and zero-width formatting characters (U+200B–200D, U+2060, U+FEFF) at API ingestion, before the question reaches the	watermark.py – sanitize_input() , applied in api/

Measure	Implementation	Code Reference
	engine or any LLM prompt. This prevents a caller from injecting watermark-class characters that would corrupt or forge the Art. 50(2) provenance marking of the resulting report.	models.py and engine.py
Honest bot identification	User-Agent: CIATA-DeepResearch/1.0 (+https://research.ciata.io; bot; contact: privacy@ciata.io) — identifies as bot with contact info, not spoofed as browser	page_reader.py L27
Per-domain rate limiting	Minimum 1 second between requests to the same domain (thread-safe)	page_reader.py L116-L129
Content truncation	Max 16,000 characters per page in production (config.yaml override; code default 8,000, avoids processing entire sites)	page_reader.py L194-L195
No raw content persistence	The PII filter runs before anything is written to disk. What is persisted is the filtered text actually sent to the extractor (capped at 16,000 chars/source), alongside the extracted findings. Unfiltered page text is never written to disk and exists only for the duration of the fetch.	engine.py — scrub_pii() applied before source.page_text is assigned
HTTPS/TLS	Caddy reverse proxy with automatic certificate provisioning	Infrastructure config
Encryption at rest	Not enabled (confirmed 2026-07-15) — production disks are not encrypted at rest. Accepted gap; compensating controls: bcrypt, TLS in transit, per-user access isolation, 30-day retention, and encrypted EU backups (Frankfurt → Helsinki). See §3.3.	Hetzner infrastructure config; DPIA action A1
Backups	Encrypted off-site backups, geo-redundant within the EU (Hetzner Frankfurt → Helsinki).	Hetzner Storage / backup service
Password hashing	bcrypt (via bcrypt library)	auth_users.py
JWT authentication	HS256 tokens, configurable TTL (access: 60 min, refresh: 7 days)	auth_users.py L122-L144
Ownership checks	Every data endpoint verifies the requesting user owns the resource	app.py L198-L210
Rate limiting	Sliding-window per-key rate limiter (30 req/60s default)	auth.py L122-L178
No cookies/tracking/analytics	No Google Analytics, no Matomo, no tracking pixels. Session via JWT in localStorage only. Cookie-free analytics via Plausible (self-hosted at trc.ciata.io — GDPR-compliant, no cookies, no personal data).	privacy.html §11
Data deletion endpoint	DELETE /v1/research/{job_id} — GDPR Art. 17 implementation	app.py L445-L468
Account soft-delete	DELETE /v1/auth/me — requires password, clears PII, deactivates	auth_users.py L393-L419
		worker.py , engine.py

Measure	Implementation	Code Reference
Account-status enforcement across job lifecycle	A deleted/deactivated account's queued or in-flight research is aborted rather than completed — the worker re-checks <code>is_active</code> before a run starts and at each iteration (supports Art. 17 erasure; a submission-time check alone would let an already-queued job finish for a deleted account)	
Automated data retention	30-day TTL with background cleanup timer	job_store.py
Non-root container	Docker runs as UID 1000	Dockerfile L50
Request IDs	Every request gets a unique ID for audit trail	middleware.py L40-L62

6.2 EU AI Act Technical Measures

Requirement	Implementation	Details
Invisible machine-detectable watermark — Art. 50(2) (PRIMARY)	VSRMark Unicode variation-selector watermark — a proprietary steganographic scheme, not governed by any EU implementing act, ISO standard or the C2PA specification as at 2026-08-11; we ask counsel whether a non-standardised marking discharges Art. 50(2) embedded invisibly in every report's carrier text	Encodes JSON payload: producer, version, <code>ai_generated: true</code> , <code>job_id</code> , UTC timestamp. Applied at report finalization in <code>engine.py</code> (<code>_finalize</code>) and as fallback in <code>job_store.py</code> . Source: watermark.py
AI provenance metadata — Art. 50(2)	Machine-readable HTML comment block in every report	<code><!-- AI-PROVENANCE generator: CIATA DeepResearch/1.0 ... eu_ai_act: Article 50(2) transparency disclosure --></code> — report.py L150-L164
Visible AI disclosure — Art. 50(2)	Every report contains: "🤖 This report was generated by AI (CIATA DeepResearch). It may contain errors or omissions. Verify critical facts independently."	report.py L194-L197
PDF AI disclosure	Branded banner in PDF header + data processing footer on every PDF	pdf_export.py L171-L204
No automated decision-making	Explicitly documented: no Art. 22 GDPR decisions	privacy.html §15
Risk classification	System intended as limited risk (informational AI tool, not high-risk per Annex III)	🕒 Assessment now overdue, not forward-planning. Annex III high-risk obligations have applied since 2026-08-01 (24 months after entry into force on 2024-08-01) — i.e. since ten days before the date of this document. The Art. 50(2)

Requirement	Implementation	Details
		transparency obligations, which we believe are the ones that bind us, have applied since 2025-08-01 and are addressed by the three measures above. If counsel concludes the system does fall within Annex III, CIATA is already subject to conformity assessment, registration and human-oversight duties and needs immediate remediation advice. See §12.2 Q7.
C2PA watermarking (PLANNED — NOT YET IMPLEMENTED)	C2PA content credentials on PDF output are planned for post-launch. Not implemented in MVP.	DPIA.md L103

6.3 PII Filter — Detailed Pattern Coverage

The PII filter ([pii_filter.py](#)) scrubs the following categories from scraped web page text **before it is sent to the LLM and before any of it is written to disk**:

Pattern	Replacement Tag	Description
Email addresses	[EMAIL]	RFC 5322 simplified pattern
Phone numbers	[TELEFON]	International + DACH formats, false-positive guards for years
German tax ID	[STEUER-ID]	11-digit Steuerliche Identifikationsnummer
German SSN	[SOZIALVERS-NR]	12-char Sozialversicherungsnummer
Swiss AHV number	[AHV-NR]	756.xxxx.xxxx.xx format
IBAN	[IBAN]	International bank account numbers
Credit card numbers	[KREDITKARTE]	13-19 digits with Luhn validation
IPv4 addresses	[IP-ADRESSE]	Standard dotted-quad format
German postal addresses	[ADRESSE]	Street name patterns (Straße, Allee, Platz, etc.) + house number

Limitations acknowledged: Regex-based filtering cannot catch all personal data (e.g., names mentioned in running text without structural markers). This is documented as a residual risk in the DPIA and accepted as proportionate under Legitimate Interest.

7. LLM Agent Architecture

The research engine uses **18 specialized LLM "agents"** — each is a single-purpose prompt template with specific temperature, token and model settings. No agents have memory across runs. No fine-tuned models are used. The thirteen core agents are listed first; the two agents of the outline-writing route follow.

Agent	Purpose	Model	Temperature	Output cap
Decomposer	Breaks research question into 4-8 sub-questions	gpt-oss-120b	0.5	4096
Re-decomposer	Refines sub-questions when coverage plateaus	gpt-oss-120b	0.6	4096
Query Generator	Creates search queries targeting uncovered gaps	gpt-oss-120b	0.7	4096
Relevance Gate	Binary yes/no filter on search result relevance	gpt-oss-120b	0.0	2048
Extractor	Extracts structured facts from page text	gemma4-chat-26b	0.3	8192
Evaluator	Assesses coverage score per sub-question	gpt-oss-120b	0.0	8192
Synthesizer	Combines all findings into a coherent report	qwen3.8-27b	0.4	32768
Table Repair	Rewrites a single comparison table whose column the writer left unfilled, in place; the rest of the report is untouched. Invoked only when a table has a column carrying at most one value across three or more rows, or a row that does not match its header	qwen3.8-27b	0.4	8192
Paragraph Checker	Post-synthesis verification: every cited paragraph is judged against its cited findings and the whole finding pool, five paragraphs per call; flags misattributed and unsupported statements	qwen3.8-27b	0.2	24576
Consistency Checker	Post-synthesis cross-check: framing sections (executive summary, decision sections, research gaps) checked statement by statement against the topic sections; lines carrying numbers, dates or article/section references compared for the same quantity; runs in parallel with the Paragraph Checker; flags go to the Report Patcher	qwen3.8-27b	0.2	24576
Entity Extractor	Identifies named entities for search seeding	gpt-oss-120b	0.2	2048
Scope Analyzer	Post-loop assessment of what the research missed	gpt-oss-120b	0.2	4096

Agent	Purpose	Model	Temperature	Output cap
Language Guard	Repairs stray non-target-language text (e.g. leaked CJK) in a finished report, line-by-line; if a repair call fails the leaked characters are removed in code; if a report is too corrupted to repair (more than 25 CJK lines), the job fails and the credit is refunded	gemma4-chat-26b	0.0	8192 (per line)
Report Patcher	Rewrites each report section that holds a flagged line, one call per section; each result is spliced back line by line so that only flagged lines can change, and a rejected rewrite appends a Verification Notice instead; invoked only when statements were flagged. A report without section headings is rewritten whole	qwen3.8-27b	0.2	24576

Five further agents run on the "outline-first" writing route, which is enabled in the deployed configuration (`config.yaml`, `research.use_outline: true`; the code default is off and the setting can be reverted via `USE_OUTLINE` in the environment without a redeploy). The last three belong to per-section writing (`research.write_per_section: true` since 2026-09-15, revertible via `WRITE_PER_SECTION`):

Agent	Purpose	Model	Temperature	Output cap
Intent Classifier	Classifies the research question into a report shape, so the same question always yields the same structure	gpt-oss-120b	0.0	2048
Outliner	Proposes the report's topic sections, then files each finding under one of them	gpt-oss-120b	0.3	8192
Section Writer	Writes one topic section per call from the findings filed under it, then the title, summary and research gaps from the finished sections	qwen3.8-27b	0.4	16384
Glossary	Merges the names the findings use for one entity into one name and acronym, so sections written apart name it the same way	gpt-oss-120b	0.3	8192
Figure Sheet	Disabled by default (<code>research.figure_sheet: false</code>). When enabled: labels quantities that two different findings state with different values; every value is checked against the finding it is attributed to, and tiers, years, ranges and figures of different kinds are dropped	gpt-oss-120b	0.3	8192

On this route the report is written into the resulting outline instead of into a fixed template: by the Section Writer when per-section writing is on, otherwise by the Synthesizer in one call. It processes the same categories of data as the template route and no others. If the outline stage fails, or if the writer is caught transcribing findings rather than writing prose, the engine falls

back to the template writer; `state.synthesis_route` records which writer actually produced each report.

7.1 LLM Provider Details

Property	Value
Provider	AKI.io GmbH
Location	Frankfurt, Germany 
Data Center	ISO 27001 certified
Reasoning Model	gpt-oss-120b — decomposer, re-decomposer, query generator, evaluator, entity extractor, scope analyzer, intent classifier, outliner, glossary, figure sheet (disabled by default)
Synthesis Model	qwen3.8-27b — synthesizer, section writer (the production writer since 2026-09-15), table repair and, since 2026-09-12, the paragraph checker and the report patcher; since 2026-09-27, the consistency checker (issue #172). Replaced mistral4-119b on 2026-08-23 (commit eb2a5b1) on citation integrity: over 8 prompt pools × 3 repeats, mistral produced malformed [ref:] digests in 6 of 24 reports and hallucinated digests in 4 of 24; qwen was clean in 72. Runs with <code>reasoning_effort=medium</code> at per-agent caps of 16,384 to 32,768 tokens (\$7).
Extraction Model	gemma4-chat-26b (Gemma 4 26B, reasoning disabled) — the extractor, the pipeline's highest-volume agent (one call per page read). Since 2026-09-13: 5x faster than qwen3.6-chat-35b with more and better-grounded findings on 40 real pages. The <code>relevance_gate</code> moved to gpt-oss-120b the same day (same verdicts, 4.5x faster).
Language-Guard Model	gemma4-chat-26b (Gemma 4 26B, reasoning disabled) — the <code>language_guard</code> only (repair of stray non-target-language lines). Moved from qwen3.6-chat-35b on 2026-09-27 (issue #167): qwen3.6-chat-35b was using its token budget on reasoning, leaving no tokens for the repair output; gemma4-chat-26b with reasoning disabled repaired 39 of 39 lines in a live test (the old setting repaired 2 of 10), at 0.4 s median per line.
API Protocol	OpenAI-compatible chat completion API
Connection	Streaming for reasoning agents; non-streaming for the mechanical agents (extractor, <code>relevance_gate</code> , <code>language_guard</code>)
Data Retention	Zero retention is the intended basis (no storage, no training use); pending DPA execution — see §5
DPA	 Not yet executed with AKI.io GmbH (see §5). When executed, must explicitly cover the three model IDs currently in use (gpt-oss-120b, qwen3.8-27b, gemma4-chat-26b) and past processing on qwen3.6-chat-35b (in use until 2026-09-27).

 **TO CONFIRM:** The AKI.io DPA must explicitly cover the **three model identifiers currently in use** (gpt-oss-120b, qwen3.8-27b, gemma4-chat-26b) and must also cover past processing on qwen3.6-chat-35b (in use until 2026-09-27). Note that the synthesis model changed on 2026-08-23; a DPA drafted against mistral4-119b would not cover the model that writes the reports. Please verify that the DPA's scope of processing lists all three current models and the retired qwen3.6-chat-35b, or obtain a written confirmation from AKI.io that all are covered.

Model change since the previous version of this briefing (2026-07-15). That version named `minimax-m2.5-230b` as the primary model. **MiniMax is no longer used anywhere in the research pipeline** — AKI.io is retiring the model, and it was removed from the last research agent on 2026-07-26. Every agent names its model explicitly in `config.yaml`, so no call falls through to a default. The identifier `minimax-m2.5-230b` survives only as an unreachable fallback value in `src/config.py` and `.env.example` — dead configuration, not a live code path; removing it is queued as a cleanup. Counsel should treat the three models above as the complete set of models that process customer data. Model routing is a configuration value we revise as models are retired or outperformed — see question §12.2 (9) on whether such changes carry any external duty.

Model change since v1.3 (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. Measured before the change on seven retained reports: the previous checks were 2% precise, and the previous patcher could not return a long report, so every patch shipped the Verification Notice instead of a correction.

Model change since v1.4 (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).

Change since v1.11 (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.

Model change since v1.10 (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. `qwen3.6-chat-35b` was consuming its entire token budget on reasoning, leaving no tokens for the repair output; `gemma4-chat-26b` with reasoning disabled repaired 39 of 39 lines in a live test (the old setting repaired 2 of 10). When a repair call fails, the leaked characters are now removed in code. The DPA scope must cover the three models in use plus past processing on `qwen3.6-chat-35b`. Same data categories, same recipient (AKI.io GmbH, Frankfurt).

7.2 What is Sent to the LLM

For each LLM call, the following is sent:

1. **System prompt** — A task-specific instruction template (hardcoded in source code, no user data)
2. **User context** — Varies by agent:
3. Decomposer/Re-decomposer: The research question text
4. Query Generator: Question + current sub-questions + coverage gaps + existing queries
5. Extractor: **PII-scrubbed page text** (max 16,000 chars in prod; 8,000 code default) + research question
6. Evaluator: Sub-questions + all findings + source count + previous coverage score (**not** the report draft — the evaluator never receives report text)
7. Intent Classifier: The research question only
8. Outliner: The research question + all findings (to propose sections and file each finding under one)
9. Glossary: The research question + the entity names the findings are about, with acronyms and counts
10. Figure Sheet (only when enabled): The research question + the findings that state a figure, grouped by entity (at most 8 per entity)
11. Section Writer: The research question + the outline + the findings filed under one section, with the glossary and, when enabled, the figure sheet; for the final call, the written sections + the findings filed under no section
12. Synthesizer: All findings + sub-questions + coverage map (+ the outline, on the outline route)
13. Paragraph Checker: The draft report's cited paragraphs + all findings (a batch whose call fails is sent once more)
14. Report Patcher: Draft report + the flagged statements with the checker's reasons
15. Scope Analyzer: Sub-questions + draft report + coverage data

[!IMPORTANT] **PII filter is applied between page fetching and everything downstream.** Fetched page text is scrubbed for personal data patterns before it is sent to any LLM **and before it is written to disk**. Neither the LLM nor the stored run state ever contains raw, unfiltered page content.

8. Web Scraping — Legal Considerations

8.1 What the System Scrapes

- **Only publicly accessible web pages** — no login bypass, no paywall circumvention, no authentication
- **Only HTML pages** — binary files (PDF, images, etc.) are skipped via Content-Type check
- **Text extraction only** — no images, no JavaScript execution, no screenshots

[!WARNING] **Bot-wall interstitials — please assess (added 2026-08-06, after the previous version of this briefing)**. Some sites answer our request with a Cloudflare-style holding page rather than the article. The system recognises four such interstitials by their title/body text ("Checking your browser", "Just a moment...", "Enable JavaScript and cookies to continue", "Attention Required! | Cloudflare") and, on detecting one, **waits 6 seconds and repeats the request once** to the same URL.

What the retry does **not** do: it solves no CAPTCHA or JavaScript challenge, executes no JavaScript, presents no cookie, token or session, and does not change or disguise its User-Agent — the second request is byte-for-byte the same honest bot request as the first. If the interstitial is still returned, the URL is abandoned and never processed. In effect it recovers pages where the wall was a transient rate-limit response, and nothing else.

We flag this explicitly because it sits close to the "no circumvention" line above and we would rather counsel rule on it than not see it. See question §12.1 (1a).

8.2 Compliance Measures for Web Scraping

Measure	Implementation
robots.txt compliance	Checked before every direct CIATA page fetch; cached 1 hour per domain. Not applied to pages Staan retrieves server-side (the production default) — see §6.1.
Honest bot identification	User-Agent identifies as bot with contact email
Per-domain rate limiting	1 second minimum between requests to the same domain
Content truncation	Max 16,000 characters per page in production (8,000 code default)
No content republication	Only extracted factual findings with attribution appear in the report; the retained page text is an internal diagnostic record and is never republished to users
PII filtering	Personal data patterns removed before AI processing and before storage
Bounded retention of page text	The PII-filtered extractor input (max 16,000 chars/source) is kept in the run's state. json for the 30-day run TTL and is deleted with the run. Unfiltered page content exists only in memory during the fetch.
Cloudflare / bot-wall handling	If a fetched page turns out to be a Cloudflare-style interstitial ("Checking your browser", "Just a moment...", "Enable JavaScript and cookies to continue", "Attention Required! Cloudflare"), the system waits 6 seconds and issues one further request to the same URL with the same honest bot User-Agent. It solves no challenge, executes no JavaScript, presents no cookie or token, and does not alter its identification. If the interstitial persists, the URL is abandoned. See §8.1 and question §12.1 (1a).

8.3 Legitimate Interest Assessment

A full Legitimate Interest Assessment (LIA) has been conducted and documented: [LIA.md](#)

Key conclusion: The legitimate interests of CIATA and its users outweigh the minimal impact on data subjects because: 1. Only publicly accessible content is processed 2. Publisher

restrictions (robots.txt) are respected on CIATA's own fetches 3. Personal data is filtered before AI processing 4. Content is not re-published; only factual summaries with attribution 5. Processing is transient with configurable data retention

9. Authentication & Authorization Model

9.1 Authentication Methods (checked in order)

1. **JWT Bearer token** — Authorization: Bearer <jwt> header
2. **JWT query parameter** — ?token=<jwt> (for SSE/EventSource which doesn't support custom headers)
3. **Per-user API key** — X-API-Key: dr_<key> → maps to a specific user account
4. **Legacy shared API key** — X-API-Key: <key> → returns the sentinel user id __legacy_api_key__ (backward compat)

Server-side token revocation. JWTs are not merely time-limited. Every JWT-authenticated request re-checks a token_version counter held against the user record (database schema v12); changing or resetting a password increments it, which invalidates **all** outstanding access and refresh tokens for that user immediately, independently of their remaining TTL. This closes the window in which a stolen token would otherwise stay valid for up to its full lifetime after the user takes remedial action.

9.2 Authorization Model

- Every data endpoint (GET/DELETE research jobs) verifies **ownership** — users can only access their own data
 - Admin users can access all data
 - Payment verification checks that the payment metadata user_id matches the requesting user
 - Rate limiting: 30 requests per 60 seconds per API key (configurable)
-

10. Report Output — What Users Receive

Every generated report contains:

1. **AI provenance metadata** — Machine-readable HTML comment block (EU AI Act Art. 50(2))
2. **Visible AI disclosure** — Text notice that the report is AI-generated
3. **Research question** — The original question
4. **Generated metadata** — Date, coverage score, iterations, sources, findings, token usage, model name
5. **Main report body** — Structured markdown with citations (e.g., [1], [2])

6. **Research Gaps block — warnings only.** The report states what it can say against itself and nothing in its own favour: sub-questions with almost no evidence behind them, and topics the scope analyzer flagged as missing or shallow. It deliberately does **not** issue per-sub-question verdicts (answered / partial / unanswered), does not assert what was "covered well", and reports no completeness percentage. Those claims were removed on 2026-08-02 after the scope analyzer was observed reporting five topics as uncovered while each had its own section in the same report: a false gap costs the reader a moment, a false assurance of coverage costs them their trust in the document. If the gaps block is empty, no reassurance is printed in its place.
7. **Numbered references** — All source URLs with titles, linked
8. **Methodology section** — How the research was conducted (iterations, stop reason, IGR trajectory)
9. **Source diversity stats** — Domain entropy, dominant domain share

PDF reports additionally contain: - CIATA branded header with logo - AI disclosure banner - Data Processing Summary footer (AI Model, Search, Storage, Training info + privacy@ciata.io contact)

11. Existing Compliance Documentation

The following documents have already been drafted and are available in the repository:

Document	Path	Status
GDPR Compliance & Data Flow Statement	gdpr-compliance.md	⚠ Superseded — see DPIA/LIA/ZDR (do not send)
Data Protection Impact Assessment (DPIA)	DPIA.md	Draft v1.2 (revised 2026-08-11)
Legitimate Interest Assessment (LIA)	LIA.md	Draft v1.2 (revised 2026-08-11)
Zero Data Retention Statement	ZDR_STATEMENT.md	Draft v2.1 (revised 2026-08-11)
Records of Processing Activities (Art. 30)	ROPA.md	Draft v1.0 (new, 2026-08-11)
Technical & Organisational Measures (Art. 32)	TOM.md	Draft v1.0 (new, 2026-08-11)
Personal Data Breach Response Procedure (Art. 33/34)	INCIDENT_RESPONSE.md	Draft v1.0 (new, 2026-08-11)
Sub-Processor / DPA Register	DPA_REGISTER.md	Draft v1.0 (new, 2026-08-11)

Document	Path	Status
Privacy Policy (public)	privacy.html	Live — bilingual DE/EN (revised 2026-08-11)
Impressum (public)	impressum.html	Live — bilingual DE/EN (§ 5 DDG)
Terms of Service / AGB (public)	terms.html	Live — bilingual DE/EN (incl. Widerrufsbelehrung)
API Reference	api-reference.md	Current — regenerated against the code (2026-08-11). Developer reference with no compliance content: §4.1 and §9 of this briefing, not this file, describe the API surface for the purposes of your assessment.

12. Open Questions for the Law Firm

[!IMPORTANT] We need the law firm's assessment on the following topics:

12.1 GDPR Questions

1. **Web scraping legality:** Is our Legitimate Interest Assessment sufficient for automated web page reading? Do we need additional safeguards or a different legal basis?

1a. Bot-wall interstitial retry — specific assessment requested. As described in §8.1, when a fetch returns a Cloudflare-style interstitial the system waits 6 seconds and repeats the identical honest-bot request once, then gives up. It solves no challenge and presents no credential, cookie or token. Does a single delayed repeat of an unmodified request amount to circumventing an effective technical protection measure — under § 202a StGB, § 95a UrhG, the DSM Directive Art. 4 TDM opt-out regime, or the operator's terms of use? Is the "honest bot, no challenge solved" character enough to keep it on the right side of the line, or should we remove the retry before launch? We would also welcome a view on whether a robots.txt-style machine-readable opt-out or an interstitial should be treated as an equivalent reservation of rights for TDM purposes (see also question 10).

1. **Incidental PII in LLM prompts:** Our regex PII filter catches common patterns but cannot catch names mentioned in running text. Is the current approach (regex + documented residual risk) sufficient, or do we need an LLM-based NER pre-filter?
2. **Data retention policy:** Is the 30-day default TTL appropriate? Should we offer shorter retention or "process-and-delete" mode where data is deleted immediately after report delivery?

3a. Retained extractor input. Within the 30-day run record we keep, per source, the PII-filtered page text that was handed to the extractor (capped at 16,000 characters), so that a source which produced no findings can be diagnosed. It is deleted with the run, and an admin-only endpoint can display it. Is retaining filtered third-party page text for this diagnostic purpose proportionate under Art. 6(1)(f), or should it be shortened, sampled, or dropped?

3b. Two retention periods we cannot set ourselves. (i) The **B2B billing profile**

(`company_profiles`) is kept while the account exists and deleted 90 days after account deletion (since 2026-10-08, #193); the invoices issued from it stay with Qonto. We think the invoices, not our profile, carry the § 147 AO retention. Is that right, and is 90 days after deletion an appropriate period for our copy? (ii) **Withdrawal-consent records** (`withdrawal_consent`s) evidence the customer's § 356(5) BGB request for immediate performance. They must outlive the 30-day run TTL, but we have not set a period. What minimum should we hold them for, given the relevant limitation periods?

1. **DuckDuckGo/Google fallback:** If a deployment operator configures a US-based search backend, what additional safeguards or disclosures are needed? Standard Contractual Clauses? Explicit user consent?
2. **Scaleway for transactional emails.** Scaleway SAS (a French subsidiary of the Iliad Group) publishes a standard DPA dated 1 June 2024. Three specific points for counsel: (a) Scaleway's sub-processor list includes data-centre colocation operators headquartered in the US (Equinix EMEA, Iron Mountain, Digital Realty), each listed with EU-only processing locations; no transfer outside the EU is claimed for TEM. Please confirm whether the US parent-company nexus of these colocation operators creates any residual transfer risk or requires additional safeguards. (b) The report-ready email (issue #159) carries the report title and a short summary of the report (up to about 500 characters of the executive summary). Is this research content in a different data category from verification and purchase mails for transfer-assessment or data-category purposes? (c) Professional-secrecy holders subject to § 203 StGB (law firms, medical practices, tax advisors) can turn the report-ready mail off in their account settings. Is opt-out a sufficient default for this user group, or should the report-ready mail be opt-in for professional-secrecy holders?

5a. Operational telemetry before we switch it on (§5, OpenObserve). We plan to enable an operational metrics exporter and would rather have your view before it goes live. Records would carry a fixed metric name, a number, and an opaque per-run `job_id` — no question text, report text, page content, URLs, email addresses or IP addresses. Three questions: (i) does the `job_id`, which we can correlate to a user through our own database, make these records pseudonymous personal data requiring disclosure and an Art. 6(1)(f) basis? (ii) the performance record includes a raw exception error string that we cannot categorically guarantee is free of incidental content — scrub it, drop it, or accept it? (iii) if we self-host the instance on our own infrastructure, do you agree it is not a third-party sub-processor, on the same reasoning as our self-hosted Plausible?

5b. Brevo wind-down. Brevo SAS was the email processor until 2026-09-26. Its Art. 28 DPA was never executed. Please advise on the steps required to wind down the Brevo account and

obtain deletion confirmation of its sending logs, and whether the absence of an executed DPA creates any notification or remediation obligation before closure.

1. **User research questions as personal data:** Can a research question itself constitute personal data? (e.g., "What treatment options exist for condition X in patient Y?") Should we implement content moderation?

12.2 EU AI Act Questions

1. **Risk classification — time-critical.** We assess DeepResearch as **limited risk** (informational AI tool, not high-risk per Annex III). Does the law firm agree? Are there edge cases where it could be considered higher risk? Please note this is **no longer a forward-planning question**: Annex III high-risk obligations have applied since **2026-08-01**, ten days before the date of this briefing. If your assessment puts us inside Annex III — or leaves it genuinely arguable — we need to know what we are already non-compliant with and in what order to fix it. A specific edge case we would like tested: our customers are consultants and analysts who use these reports as an input to their own advice to third parties. Does that downstream use pull the tool toward any Annex III use case, or does the 'informational research assistant' characterisation hold?
2. **Art. 50(2) transparency obligations:** Our current implementation includes: (a) a visible "AI-generated" disclosure label in every report, (b) a machine-readable HTML provenance comment, and (c) an invisible VSRMark Unicode variation-selector watermark encoding producer id, version, ai_generated flag, job_id, and UTC timestamp. C2PA content credentials are planned post-launch but not yet implemented. Is the combination of (a)+(b)+(c) sufficient to satisfy Art. 50(2)?
3. **General-purpose AI model rules:** AKI.io hosts gpt-oss-120b (reasoning agents, relevance gate), qwen3.8-27b (synthesis, table repair, paragraph check, patch) and gemma4-chat-26b (extraction and language guard) — all three are open-weight models, and none is fine-tuned by us. qwen3.6-chat-35b (formerly the language guard) was retired on 2026-09-27. Do the GPAI provider obligations (Art. 53) flow through to us as deployers, or are they solely AKI.io's responsibility? A related question: we change which model serves which agent as models are retired or outperformed (six such changes since 2026-07-15). Does swapping an open-weight model behind an unchanged product feature carry any notification, documentation or re-assessment duty toward customers or authorities?
4. **Copyright and TDM:** Does our web scraping fall under the Text and Data Mining exception (Art. 4 DSM Directive / § 44b UrhG)? How do opt-out signals (robots.txt, TDM-specific headers) interact with our system?

12.3 General Questions

1. **Imprint / Impressum:** We now publish a **bilingual (DE/EN) Impressum** at /impressum.html per § 5 DDG (the DDG replaced the TMG in May 2024; the page previously cited § 5 TMG and was corrected on 2026-08-11) (company, HRB 266084 B, VAT-IdNr., Geschäftsführer, contact, § 18 (2) MStV content responsibility, and a § 36 VSBG dispute-

resolution statement). Please review it for completeness and correctness for a B2B SaaS and confirm whether any further mandatory particulars are missing.

2. **Terms of Service / AGB:** We now publish a **bilingual (DE/EN) AGB** at `/terms.html` covering scope, subject matter, registration, the credit/payment model (Mollie), an AI-content and liability disclaimer, rights of use, availability/liability limits, the statutory **Widerrufsbelehrung + Muster-Widerrufsformular** with § 356(5) / § 327 BGB early-expiry logic for digital services, data protection, termination, and final provisions. Please review. Two specific points: **(a)** the **English version is a convenience translation** with a governing-language clause making the **German authoritative** — is that acceptable, or should the English bind for EN-language consumers? **(b)** Is the § 356(5) / § 327 BGB early-expiry construction correctly worded for our per-report credit-consumption model (credit redeemed → immediate performance requested → withdrawal right for that report expires)?
3. **Insurance:** Should we carry professional liability / E&O insurance given that users rely on AI-generated research for business decisions?

13a. Four documents drafted for the first time — please review them as drafts, not as settled positions. These did not exist before this version and are included in the package: - `ROPA.md` — **Art. 30 Records of Processing Activities**. Our first attempt at a controller-side record. Is the activity breakdown the right one, and are the required Art. 30(1) fields complete? - `TOM.md` — **Art. 32 technical and organisational measures**. Drafted as the annex an enterprise customer would attach to a DPA. Every measure is verified against the source code, and measures we could not verify were left out rather than softened. Is it adequate to hand to a customer's DPO? - `INCIDENT_RESPONSE.md` — **Art. 33/34 breach response**. Written for a company of our actual size: one managing director, no DPO, no on-call rota, a single production host. It names the points where we have a single point of failure rather than papering over them. Is the triage threshold right, and is the 72-hour workflow defensible? - `DPA_REGISTER.md` — **sub-processor and DPA register**. Doubles as the tracker for getting the outstanding DPAs signed. Please check the Billie and Qonto independent-controller classifications in particular.

Two related questions: does the absence of a DPO remain tenable given the DPIA's own Art. 35(1) high-risk conclusion and § 38(1) sentence 2 BDSG? And which of these four do you need finalised before launch, as opposed to shortly after?

4. **~~Google Fonts CDN:~~ RESOLVED** (2026-07-10) — Google Fonts and the jsDelivr CDN (`marked.js`) have been replaced with self-hosted assets. All font and JS files are now served from CIATA's own infrastructure (see `frontend/css/fonts.css`). No further action required on this point.
5. **Plausible analytics — disclosure sufficiency:** We use a self-hosted Plausible instance (`trc.ciata.io`) on CIATA-controlled infrastructure. Plausible is cookie-free and does not collect personal data; CIATA is not a processor of third-party data here. The instance is now disclosed in the public privacy policy at `/privacy.html` §11. Is the current disclosure language sufficient, or does it need any amendment?

13. File Index

For the law firm's reference, here is the complete source code structure:

```
auto_deep_research/
├── src/
│   ├── engine.py           # Core research loop (1,592 lines)
│   ├── llm_client.py       # AKI.io LLM client wrapper, truncation guard (583 lines)
│   ├── searcher.py         # Web search backends: Staan, DDG, Google (376 lines)
│   └── page_reader.py       # Page fetching: robots.txt, rate limiting, bot-wall
detection (367 lines)
├── pii_filter.py           # PII regex scrubber (199 lines)
├── state.py                # Research state data model + run provenance (508 lines)
├── report.py               # Markdown report renderer (533 lines)
├── pdf_export.py           # PDF export with branding (504 lines)
├── outline.py              # Outline data model for the outline-writing route (305
lines)
├── citations.py            # Citation token resolution (214 lines)
├── subquestions.py         # Sub-question resolution (59 lines)
├── watermark.py            # VSRMark watermark embed/extract + input sanitising (403
lines)
├── version.py              # Run provenance stamping – code revision, models, config
digest (130 lines)
├── url_validation.py       # URL/SSRF validation (64 lines)
├── perf.py                 # Performance instrumentation, optional OpenObserve export
(273 lines)
├── config.py               # Centralized configuration (561 lines)
├── cli.py                  # CLI interface (not used in SaaS) (210 lines)
├── agents/
│   ├── decomposer.py       # Question decomposition (233 lines)
│   ├── query_generator.py   # Search query generation (263 lines)
│   ├── extractor.py         # Finding extraction from pages (145 lines)
│   ├── evaluator.py         # Coverage evaluation (208 lines)
│   └── synthesizer.py       # Report synthesis: template, outline and per-section
routes (622 lines)
│   ├── glossary.py          # One name per entity for the per-section writer (235
lines)
│   ├── figures.py           # Figure disagreements for the per-section writer,
disabled by default (326 lines)
│   ├── paragraph_checker.py # Post-synthesis verification, per cited paragraph (256
lines)
│   ├── report_patcher.py    # One whole-report rewrite of the flagged lines,
spliced back (314 lines)
│   ├── language_guard.py     # Repairs stray non-target-language text (239 lines)
│   ├── relevance_gate.py     # Search result filtering (172 lines)
│   ├── scope_analyzer.py    # Post-loop scope assessment (191 lines)
│   └── intent_classifier.py  # Report-shape classification, outline route (148
lines)
├── outliner.py             # Outline proposal + finding assignment (327 lines)
├── report_metrics/         # Deterministic report-quality metrics (internal QA)
└── report_judge/           # LLM-judge scoring dimensions (internal QA)
```

```

├── api/
│   ├── app.py                # FastAPI application + endpoints (1,253 lines)
│   ├── auth.py               # Authentication middleware (289 lines)
│   ├── auth_users.py         # User auth endpoints, token revocation (679 lines)
│   ├── database.py           # SQLite layer: users, keys, credits, profiles,
consents (1,539 lines)
│   ├── job_store.py          # Research job management + retention enforcement (629
lines)
│   ├── worker.py             # Background worker thread pool (638 lines)
│   ├── models.py             # API request/response models (531 lines)
│   ├── billing.py            # Credit checking (44 lines)
│   ├── mollie_payments.py    # Payment processing (790 lines)
│   ├── qonto_invoicing.py    # Invoice generation (357 lines)
│   ├── vat.py                # VAT handling (77 lines)
│   ├── email_service.py      # Transactional email: shared message builder + Brevo
and Scaleway senders (896 lines)
│   ├── report_forensics.py   # Admin report forensics – own runs only (464 lines)
│   ├── report_diagnostics.py # Admin report diagnostics – own runs only (138 lines)
│   ├── health.py             # Health/readiness endpoints (173 lines)
│   ├── rate_limit.py         # Rate limiting (73 lines)
│   ├── i18n.py               # API message localisation (158 lines)
│   ├── middleware.py         # Request IDs, access logging (147 lines)
│   └── dependencies.py       # FastAPI dependency injection (15 lines)
├── frontend/
│   ├── index.html            # Main SPA
│   ├── admin.html            # Admin panel
│   ├── inspector.html        # Admin report inspector (own runs only)
│   ├── privacy.html          # Privacy policy (bilingual DE/EN)
│   ├── terms.html            # AGB / Terms of Service (bilingual DE/EN)
│   ├── impressum.html        # Imprint (bilingual DE/EN)
│   ├── locales/              # UI translations (de.json, en.json)
│   ├── css/                  # Stylesheets
│   └── js/                   # JavaScript
├── docs/
│   ├── api-reference.md      # REST API documentation
│   ├── engine_flow.md        # Research pipeline flow diagram
│   ├── gdpr-compliance.md    # SUPERSEDED – retained for history only, do not send
│   └── compliance/
│       ├── LAW_FIRM_BRIEFING.md # This document
│       ├── DPIA.md            # Data Protection Impact Assessment
│       ├── LIA.md             # Legitimate Interest Assessment
│       ├── ZDR_STATEMENT.md   # Data Retention & Sub-processor Statement
│       ├── ROPA.md            # Records of Processing Activities (Art. 30)
│       ├── TOM.md             # Technical & Organisational Measures (Art. 32)
│       ├── INCIDENT_RESPONSE.md # Breach response procedure (Art. 33/34)
│       └── DPA_REGISTER.md     # Sub-processor / DPA register
├── tests/                    # 2,231 tests
├── Dockerfile                # Multi-stage Docker build
├── docker-compose.yml        # Container orchestration
├── pyproject.toml            # Python project config
└── .env.example              # Environment variable template

```