

DATA PARTNERSHIP BRIEF

Code Repositories & Workflow / Organisational Data

What we license, why it matters, how your data is protected, and the five steps from first call to offer.

01 Who we are, and why this data

Aptura is a London-based applied AI research company. We build the datasets, evaluations and RL environments that frontier AI labs use to train and test their models on real professional work. Our founding team comes out of Lazard and Partners Group. In every deal we are the licensee and the accountable party: we do not broker, resell, or re-list your data on any marketplace.

AI models are strong on textbook problems and weak on real jobs, and the gap is data. Real repositories and workflow records show how work is actually decomposed, reviewed, blocked and completed inside an organisation. That is exactly what is needed to train and honestly evaluate AI systems doing multi-step professional work. We license your material for that purpose alone.

02 What we are looking for

Code repositories & engineering history

- Private, internal, archived or wound-down codebases with full Git history intact.
- Linked trajectories: ticket → branch → pull request → review → tests → CI → release.
- CI/CD config and build logs, test suites, container artifacts, infrastructure-as-code.
- Incident and resolution records, post-mortems, security-fix histories.

Workflow & organisational data

- **Project trackers:** Jira, Linear, Asana, Monday exports; tickets with status history, comments, sprint and epic structure.
- **Communications:** Slack or Teams workspace exports, email threads tied to projects, where lawfully shareable.
- **Documents & drives:** Google Drive / SharePoint corpora, Confluence and Notion wikis, SOPs, runbooks, meeting notes and decision logs.
- **Business systems:** CRM exports (HubSpot, Salesforce), support desks (Zendesk, Intercom), incident tools (PagerDuty, Opsgenie).
- **Structure:** org charts, role definitions, approval chains and process maps.

What drives value most: linkage. A repository connected to its tickets, reviews and CI history is worth substantially more than the same code in isolation, because it captures the decisions, not just the output. Code never published publicly carries a premium. Active, winding-down and retired assets are all in scope; legacy material is often the most valuable to us and the least valuable to you.

What we will never do with your data

- Resell or redistribute it as a standalone dataset, or publish it in any form.
- Use it to build anything that competes with your business.
- Attempt to re-identify individuals, or profile your employees or customers.
- Retain it beyond the agreed term, or use it outside the agreed purpose.

03 How we protect your data

Your data never leaves a controlled environment, and you stay in control of it at every stage. Three layers of protection:

- **Contractual.** Mutual NDA before anything technical is discussed. The licence has a single permitted purpose; anything outside it is a breach. Onward transfer only to named, contractually bound lab counterparties, on terms no weaker than ours. You retain ownership throughout, with audit rights.
- **Technical.** The controls in the table below, applied to every engagement.
- **Organisational.** A named data custodian as your single point of accountability, trained personnel under written confidentiality obligations, and notification of any security incident within 72 hours.

Control	Implementation
Transfer	Encrypted, over a channel you choose (SFTP, scoped S3/GCS, or physical media).
Storage & access	AES-256 at rest in a UK/EU environment. Named-individual, least-privilege access with MFA, fully logged.
Segregation	A dedicated workspace for your data, never pooled with other partners.
Sanitisation	Automated secret removal (keys, tokens, credentials) plus PII detection and pseudonymisation before any downstream use.
Sample sign-off	You review and approve a sanitised sample before anything is delivered downstream.
Deletion	Certified deletion on request or at end of term, including backups.

04 Built for GDPR from day one

The design principle is simple: **we want as little personal data as possible.** This data is valuable for its structure and reasoning, not for who wrote it. We prefer datasets pseudonymised at source; where they are not, pseudonymisation runs as a first-pass step. We do not seek special category data or customer personal data, and anything embedded is removed during sanitisation. Fully anonymised data falls outside UK/EU GDPR entirely, which is our preferred end state.

Where personal data does remain: roles are confirmed in writing per deal (ordinarily you remain controller of the source and we act as independent controller of the licensed extract); the lawful basis is typically Article 6(1)(f) legitimate interests for scientific and technical research, supported by a documented assessment we share with you; Article 89(1) safeguards apply; data subject rights have a defined route; cross-border transfers use UK IDTA / EU SCCs, and storage can be fixed to the UK/EU on request.

Beyond GDPR, we also check the open-source licence position of any code, confirm no client or supplier NDA blocks the transfer, and confirm contractor IP assignments cover the material.

05 The process: five steps, two to four weeks

- 1 Introductory call** 30 minutes
We explain where the data goes; you describe what you hold. High-level fit only.
- 2 Mutual NDA** typically same week
Our standard mNDA or yours, signed before any specifics are exchanged.
- 3 Data evaluation & Data Profile Form**
A short structured form: volume, format, date range, provenance, rights and chain of title, repo specifics (merge vs. squash history, ticket-linked commits, ever-public status, AI-assistant adoption date), licence position, personal data content, restrictions. Where possible we review a small sample under NDA. Output: a go / no-go and an indicative value range.
- 4 Technical due diligence** one or two sessions
Completeness, extraction effort, secret and PII scanning, open-source check, chain-of-title and GDPR sign-off on both sides.
- 5 Offer**
Price and structure, licence scope, term, delivery, retention and deletion. Negotiated, then executed.

06 Commercial structure

We fit the deal to your situation: a **one-time licence fee** for a defined term (most common), a **perpetual licence** at a premium, an **exclusivity premium** where offered, a **recurring fee** for ongoing feeds, or a **milestone structure** split between signature and validated delivery. Pricing is driven by rarity, depth, cleanliness, licence clarity and cross-system traceability. You get an indicative range at Step 3, not at the end.

07 Common questions

Can we control who the data goes to?

Yes. Recipients can be named, restricted by category, or subject to your approval.

What if we only want to sell part of it?

Normal and encouraged. The Data Profile Form is built for partial scoping.

Our code contains secrets and credentials.

Assume it does; nearly all code does. Secret removal is standard in our pipeline, and you sign off on a sanitised sample before anything moves downstream.

We're mid-wind-down and moving fast.

Tell us. Where the rights position is clean and an administrator can execute, we compress the process considerably.

What if we say no after due diligence?

Nothing happens, the NDA still binds us, and any sample material is deleted with written confirmation.

Next step: a 30-minute call.

No preparation required. If there is no fit, we will tell you on that call.

Nakul Maheshwari · Founding Team, Aptura AI UK Ltd · nakul@aptura.ai



Aptura AI UK Ltd, London, United Kingdom. This document is provided for information only; it is not an offer and does not constitute legal advice. Data protection positions summarised here are indicative and are confirmed contractually on a deal-by-deal basis. © 2026 Aptura AI UK Ltd. Confidential.