LinkedIn Data Risk Checklist for Founders and Buyers

How to use this checklist

This checklist is for founders, GTM operators, RevOps leaders, and procurement teams evaluating any vendor, workflow, or internal project that touches LinkedIn-derived data. The point is not abstract compliance theater. The point is to surface whether LinkedIn data is anywhere in your systems, vendors, models, browser tooling, or customer workflows before it becomes a lawsuit, subpoena, takedown, or ugly diligence surprise.

Red flags to audit immediately

1. Any employee uses a Chrome extension that overlays or extracts data on linkedin.com.
2. Any internal spreadsheet, CRM field, or enrichment output stores LinkedIn URLs, member IDs, or copied profile text.
3. Any vendor contract describes data as public web, professional profiles, enrichment, contact intelligence, or verified contacts without naming provenance.
4. Any model, prompt corpus, fine-tuning set, or evaluation fixture includes LinkedIn-derived text or attributes.
5. Any archived S3 bucket, database backup, warehouse table, or Git blob still contains LinkedIn-sourced rows.
6. Your company has or had a LinkedIn Company Page or employee admins who clicked through LinkedIn terms.
7. Internal Slack or email contains phrases like scrape, bypass, Sales Navigator, browser automation, or proxy rotation tied to LinkedIn.

Vendor diligence questions

Ask every vendor these questions in writing:
- Does any data in your product originate from LinkedIn, directly or indirectly?
- Do you or any upstream provider use browser extensions, automated browsers, bots, or scripts on linkedin.com?
- Have you received a cease-and-desist, account restriction, page removal, or lawsuit from LinkedIn?
- Can you warrant non-use of LinkedIn data in production, staging, development, training data, prompts, and model evaluation?
- If challenged, can you document source provenance for each major data class?
- Will you indemnify us for claims tied to your data provenance?
- What happens if an upstream source is found to be LinkedIn-derived?

Internal clean-room standard

A real clean-room standard is stricter than most teams think. Aim for all of the following:
- No LinkedIn account for the company, no Company Page, no employee-admin usage for the data business itself.
- No LinkedIn URLs, member IDs, profile text, screenshots, or copied fields in any environment.
- No LinkedIn-derived data in product databases, staging, sandboxes, notebooks, fixtures, or examples.

- No LinkedIn data in Git history, deleted branches, or reachable blobs.
- No browser extensions touching linkedin.com on company-managed machines used for data operations.
- No vendors whose vendors touched LinkedIn. Two hops still counts as provenance risk.
- Written provenance documentation for every dataset you ship or buy.

Board and diligence questions

If you are a founder, assume these will get asked in diligence eventually:
- Can you state under oath whether the company has ever possessed LinkedIn-derived data?
- Can you prove your models were not trained on LinkedIn-derived inputs?
- Can you identify every customer exposed to disputed data?
- Can you destroy downstream aggregates, inferred features, and synthesized records if ordered?
- Can you survive a customer notification requirement?
- Can you survive six months of discovery without your runway imploding?

Decision rule

If the honest answer is maybe, probably, or indirectly, treat it as yes. If LinkedIn data touched the building, you have work to do. The cheapest time to fix provenance is before purchase, before ingest, before model training, and definitely before a complaint lands.