# Client-Side Redaction of Swedish Personally Identifiable Information

maskera: deterministic rules and a 43 MB on-device NER model that mask PII before text reaches an LLM, a log, or analytics

Joel Hägvall

`maskera.dev` `github.com/joelhagvall/maskera`

August 14, 2026 · Whitepaper v1.22

**Abstract**

maskera is an open-source (MIT) library that redacts personally identifiable information in Swedish text locally, in the browser or in the caller's own server process, before text is sent anywhere. Structured identifiers (personnummer, organisation numbers, phone numbers, bank account references, and others) are caught by deterministic, format-aware rules with selective checksums. Names, places, organisations, and street addresses in free text are caught by a purpose-built Swedish NER model of 43 MB, small enough to run where the text already is, including inside a browser tab. The input text never leaves the caller's environment, the packages collect no telemetry, and the mapping that restores masked values is held locally by the caller. In the default setup the browser downloads the static assets once from a public host; self-hosting those assets makes zero external requests. In a current, author-coupled comparison on 211 synthetic Swedish names, places, and organisations, the published q4 artifact masked all 211 both with original casing and lowercased; KBLab lower-mix fp32 masked 205 and 187. The corpus was written by Maskera's developer, so the result is directional rather than an independent ranking. maskera is a data-minimisation and data-protection-by-design tool, not a compliance guarantee; this document reports method and limitations with the same candour as the results. Intended readers: data protection officers, security teams, architects, and legal reviewers.

## 1 The problem: free text leaks PII

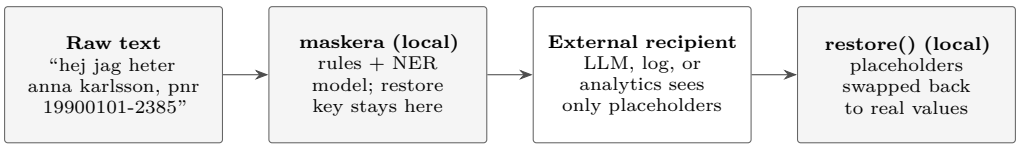
Swedish organisations now send user-generated text to more external systems than ever: large language models behind third-party APIs (frequently hosted outside the EU/EEA), log platforms, analytics pipelines, and support tooling. That text is largely free-form, and free-form text carries personal data: customers type their personnummer into chat windows, case workers mention names and home addresses in ticket notes, patients describe their situation with full identities attached.

Every such flow is a processing operation under data-protection law. Where the recipient is a third-party vendor, questions of processing agreements and third-country transfers follow, and every system that stores the raw text widens the incident surface. The simplest way to reduce all of these risks at once is to not send the personal data at all.

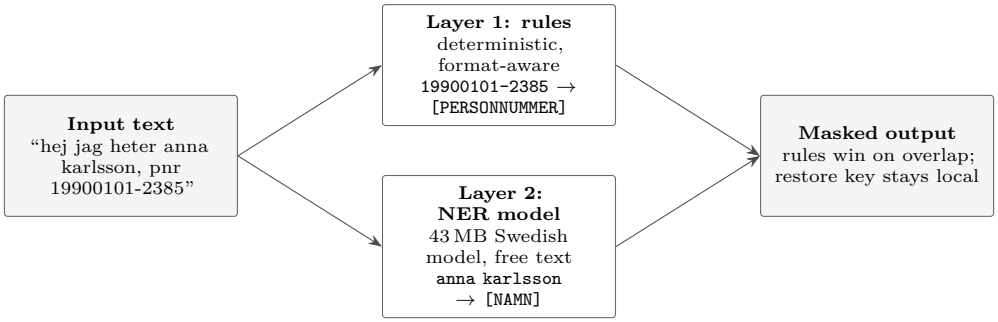
The established technical countermeasures share two weaknesses. Server-side DLP services and redaction proxies only see the text *after* it has left the client, and they are themselves one more system processing raw personal data. The publicly available Swedish NER models, for their part, are on the order of 500 MB; they require a dedicated inference service with a GPU or a strong CPU, which in practice again means a new system that sees the raw text.

maskera takes the opposite approach: redact the text *where it already is*, in the browser before a form is submitted, or inside the caller's own server process before the outbound API call. No new data recipient is introduced.

# 2 Architecture: two layers, rules first




The redaction flow is: raw text → `redact()` → masked text with placeholders → LLM / log / analytics → `restore()` locally. The external recipient only ever sees placeholders such as `[NAMN_1]` and `[PERSONNUMMER_1]`; shaded boxes run inside the caller's environment. Restoration is optional: in many logging and analytics flows it is never performed, and the placeholders are the final stored form.




## 2.1 Layer 1: deterministic rules (`@maskera/core`)

Structured Swedish identifiers are matched by deterministic, format-aware patterns: personnummer, samordningsnummer, organisation numbers, phone numbers, e-mail addresses, postcodes, bankgiro, plusgiro, IBAN, card numbers, IP addresses, and URLs. Checksums are retained where they improve precision; “2019-2024” is not read as a bankgiro number because its checksum fails. Personnummer and samordningsnummer instead require a real date shape but not a valid Luhn digit, so a mistyped control digit remains protected. Strict Luhn validators are still exposed for callers performing validation rather than redaction. The layer is synchronous, has zero dependencies, and can be used entirely on its own, for example for log masking.

## 2.2 Layer 2: a free-text NER model (`maskera`)

What rules can never catch, such as free-text names written without capital letters, is caught by `maskera-sv-ner`, a purpose-built Swedish named-entity-recognition model with four entity types: person, location, organisation, and street address. It handles text the way people actually type it: all lowercase (“hej jag heter anna karlsson”), ALL CAPS, and genitive forms. The model is 43 MB and runs in the browser (WebGPU or WASM) and in Node.

## 2.3 The hybrid rule: the deterministic layer always wins

When both layers run, any model detection that overlaps a rule detection is dropped. A personnummer is therefore always classified by deterministic date-shape logic, never by a probabilistic guess. Within one call, the same value always receives the same placeholder, so a language model can reason about `[NAMN_1]` consistently without ever seeing the underlying value. Restoration happens locally: the placeholder-to-value mapping never leaves the caller's environment, and must never be logged or stored next to the redacted text.

# 3 Privacy model: what leaves the device

**The caller's input text never leaves the caller's environment.** All redaction runs locally, in the browser or in the caller's Node process. The packages contain no telemetry and phone

home to nothing. The claim is verifiable: the code is open, and the browser's developer tools show that typed text triggers no network requests. The hosted demo counts anonymous, cookieless page views via Vercel Analytics; those events never contain text entered by the user.

For completeness, two network calls exist in the default setup, both fetching static assets and never content:

- **Model weights** (∼43 MB), downloaded once from the Hugging Face Hub or from a host of the caller's choosing, then cached.
- **The WASM runtime** for Transformers.js, loaded from a CDN.

Both can be eliminated. The model is a set of static files that can be served from the caller's own origin, and `allowRemoteModels: false` forbids remote fetches outright. In that configuration maskera makes zero external requests, which suits procurement requirements that no new subprocessors be introduced, as well as air-gapped environments. The rule layer has no network capability at all.

### 3.1 Deployment modes

The choices above collapse into four modes:

| Mode | Network requests | New subprocessors | Best for |
|---|---|---|---|
| Default | static assets once (HF Hub, CDN) | possibly HF/CDN | evaluation, demos |
| Self-hosted assets | caller's own origin only | none | production |
| Rules only | none | none | logs, pipelines |
| Air-gapped | none | none | strict environments |

In every mode, including the default, the requests carry no user content: the only things ever fetched are the static model and runtime files. The word *possibly* is deliberate: like any HTTP request, an asset fetch exposes the client's IP address to the host, and in the browser that client is the end user, so a legal review may still want to assess Hugging Face and the CDN. The self-hosted mode exists precisely so that this assessment never arises.

## 4 Model provenance and reproducibility

In a vendor assessment of AI components, two questions dominate: what was the model trained on, and can the answer be checked? For `maskera-sv-ner`, both answers are public.

- **Base model:** `KBLab/bert-base-swedish-cased`, the National Library of Sweden's BERT, released under CC0. It was fine-tuned for entity recognition, then distilled into a smaller student model quantised to a 43 MB ONNX artifact.
- **Published v18 training data:** template-generated Swedish plus documented, openly licensed public Swedish corpora. No customer data was used. The dated sources remain in the v18 model card and training journal.
- **Published privacy-clean v19:** the current release passed every defined release gate using 64,000 generated training rows and 4,760 disjoint validation rows. Customer data, logs, forum posts, news, scraped text, public NER corpora and pseudo-labels are excluded from its task-specific training.
- **v19 identifier exclusion:** a fail-closed scanner rejects every row containing an IBAN or account-shaped string, identity or organisation number, phone, e-mail, URL, public IP, payment identifier, long account/card-like number or postal code. Every ADR span also requires an explicit synthetic marker, so a random street/number pair cannot resolve silently to a plausible real property. Reserved test identifiers remain only in the separate deterministic-rule test suite.
- **No user data:** nothing from people who use maskera enters training; no such collection exists.

- **Attested v19 provenance:** exact train/validation row counts and SHA-256 hashes are bound to the generator code hash. The audit-code hashes and data hashes travel with the published weights in `privacy-attestation.json`. Training, trimming, distillation, ONNX export and publication refuse artifacts without a complete synthetic-only attestation.
- **Reproducibility:** the entire chain, from data generation through training, distillation, ONNX export, and evaluation, exists as runnable scripts in the repository. An outside auditor can follow and re-create every step.

The v19 attestation covers Maskera's task-specific data. It does not assert that the third-party KB-BERT checkpoint was itself pretrained on synthetic-only text. CC0 grants copyright permissions but is not a GDPR determination; that base-model boundary is documented explicitly in the repository's training data protection policy.

This openness supports the caller's own GDPR assessment of the component and their vendor-risk and EU AI Act documentation where applicable; it does not determine which obligations apply.

# 5 Measured quality

The first tables below are the current privacy-clean v19 release snapshot. After an explicit page break, the remaining quality and public-model comparison tables are retained as historical v18 evidence measured on **2026-07-19** with `maskera` version `0.6.4`. They must not be attributed to v19. Those measurements use q4 quantisation from v18 and **exact character spans**, the strict criterion under which both start and end offsets must match. The canonical, always-current document is `docs/BENCHMARKS.md` in the repository; this section is a dated snapshot, and where the two disagree, `BENCHMARKS.md` wins. In a privacy setting the number to weigh most heavily is **leakage**: the share of entities missed entirely and thus left unmasked. The appendix retains category-level failure modes only; raw external sentences were removed from the checkout.

## 5.1 Published v19 release snapshot

On **2026-08-06**, the attested q4 release plus the current packaged runtime passed every defined release gate. The LinkedIn-style row was re-measured from a clean build on **2026-08-10**. The published artifact is 42,705,681 bytes. These synthetic or author-coupled checks are a release battery, not an independent universal quality estimate:

| Set | Precision | Recall | Span $F_1$ | Labeled $F_1$ | Leaks |
|---|---|---|---|---|---|
| Curated (205) | 95.3% | 98.5% | 96.9% | 95.9% | 1 |
| Synthetic ADR (57) | 100.0% | 100.0% | 100.0% | 96.5% | 0 |
| LinkedIn-style (53) | 75.8% | 88.7% | 81.7% | 78.3% | 0 |

All 35 explicitly marked address spans in the revised ADR set are exact and labeled ADDRESS. One gold organisation is covered at the exact span but typed ADDRESS, which makes ADDRESS-only precision 35/36 and explains the lower labeled score. The rare-surname safety gate reaches 96.94% masked recall (285/294), above its historical 94.9% floor. The Hub artifact and source package pin target revision `7a0063375d1baabf66cf9a357dad5f46aea7008e`.

The broader synthetic-gold gate records 93.08% type $F_1$, 94.07% type recall, and 98.31% masked recall. CI gates both an $F_1$ floor and a leakage ceiling; leakage remains the primary privacy metric because it counts values left entirely unmasked.

## 5.2 Current v19 comparison with KBLab lowermix

On **2026-08-11**, the published 43 MB q4 artifact and KBLab's published 496 MB fp32 `bert-base-swedish-lowermix-reallysimple-ner` were run through the same Transformers token-classification pipeline on 121 hand-authored synthetic Swedish texts containing 211 comparable PER/LOC/ORG entities. The same texts were also forced to lowercase. Matching is by overlap; Maskera's product post-processing and its ADDRESS class are excluded.

| Casing | Model | Masked at all | Typed $F_1$ |
|---|---|---|---|
| Original | **Maskera v19 q4** | **100.0% (211/211)** | 87.1% |
| Original | KBLab lowermix fp32 | 97.2% (205/211) | **89.4%** |
| Lowercase | **Maskera v19 q4** | **100.0% (211/211)** | **85.7%** |
| Lowercase | KBLab lowermix fp32 | 88.6% (187/211) | 83.2% |

Maskera covered all 211 entities in both runs. KBLab led typed $F_1$ on original casing by 2.3 percentage points; Maskera led lowercase typed $F_1$ by 2.5 points. This is a **directional, author-coupled comparison**, not an independent universal ranking: Maskera's developer wrote the corpus. Exact model revisions, the pinned Python environment, machine output, and the `pnpm eval:kblab` command are published beside the canonical benchmark.

## 5.3 Current v19 end-product comparison with LogosGuard

On **2026-08-14**, the complete local Maskera v19 pipeline and LogosGuard 2.4.4's Chrome file-redaction flow (Free plan, `Balanced (recommended)`) were run on the same 258 synthetic Swedish domain texts with 952 annotated PII strings. The set includes names, addresses, personnummer, samordningsnummer, phone numbers, banking identifiers, and government, healthcare, and legal language.

Both products were graded with the same strict material-character definition: a full hit removes every letter and digit from an annotated value, a partial leak retains some, and a clear-text miss retains all.

| Product | Fully masked | Partial leaks | Clear-text misses |
|---|---|---|---|
| **Maskera v19 q4** | **98.0% (933/952)** | 8 | 11 |
| LogosGuard 2.4.4 | 63.7% (606/952) | 49 | 297 |

Maskera fully removed 327 more annotated values, a 34.3 percentage-point difference in full-hit rate. This remains **directional, author-coupled evidence**, not an independent ranking: Maskera's developer wrote the synthetic corpus, and the corpus is not exhaustively annotated for precision. LogosGuard's returned files also converted Swedish UTF-8 text into Windows-1252 mojibake; reversible encoding damage was repaired before scoring. Per-document outcomes, capture hashes, product settings, scorers, and checksums are published beside the canonical benchmark.

## 5.4 Historical v18: curated corpus (upper bound, regression tracker)

149 hand-authored Swedish sentences, 205 free-text entities, including hard cases (lowercase, ALL CAPS, genitives, deliberate traps). The corpus shares an author with the training-data generator and should be read as an upper bound, not a universal score.

| Metric | Result | Meaning |
|---|---|---|
| Precision | 99.5% | of the model's detections, how many were correct |
| Recall | 100.0% | of the real entities, how many were found |
| Span $F_1$ | 99.8% | harmonic mean, label-agnostic |
| Leakage | 0.0% | entities missed entirely, 0 of 205 |

### 5.5 Historical v18: external gold corpus (directional floor)

22 verbatim sentences of Swedish Wikipedia prose, 58 entities, written by others and excluded from Maskera's fine-tuning, distillation, and vocabulary selection. A small sample; read it as a directional external floor for the task-specific recipe, not proof of example-level absence from KB-BERT's earlier pretraining corpus.

| Metric | Result | Meaning |
|---|---|---|
| Precision | 96.4% | of the model's detections, how many were correct |
| Recall | 93.1% | of the real entities, how many were found |
| Span $F_1$ | 94.7% | harmonic mean, label-agnostic |
| Leakage | 3.4% | entities missed entirely, 2 of 58 |

### 5.6 Historical v18: street addresses

The two gold sets above cover person, location, and organisation; neither contains street addresses, the fourth class the model emits. Addresses are therefore measured on their own held-out set of 41 sentences (35 address spans, including saint-prefix, free-word-ending, farm and abbreviated forms), with street names chosen to avoid the training generator's stems. Measured **2026-07-19** on the same q4 artifact: redaction recall is **100%** (every address masked) with **zero leaks**, address precision is **100%** (no false ADDRESS flags on the distractor set), and the exact span including the house number is recovered on **35 of 35** addresses. These are historical v18 results. The earlier ordinary street/number surfaces were replaced on 2026-08-06 by explicit synthetic markers, so the historical result does not transfer to the replacement. That replacement has been measured separately in the published v19 snapshot above; it does not relabel this historical v18 table. The previous release's one documented exception (the ordinary word *Festen* over-redacted in one distractor sentence) was fixed in v18, which carries no gate exception at all.

### 5.7 Historical v18: comparison with public Swedish NER models

Same external gold corpus, same harness, overlap matching, labels mapped to person, location, and organisation. "Flagged" is the share of entities marked at all, under any label, i.e. the safety number.

| Model | Size | Flagged | Typed $F_1$ |
|---|---|---|---|
| **maskera** (runs in the browser) | **43 MB** | 0.97 | **0.96** |
| KBLab lowermix reallysimple-ner | ∼475 MB | 1.00 | 0.94 |
| NB AI-Lab scandi-ner | ∼500 MB | 1.00 | 0.94 |
| KBLab neriob (IOB head) | ∼475 MB | 1.00 | 0.94 |
| KB-NER (the SUC classic) | ∼475 MB | 1.00 | 0.92 |
| KBLab reallysimple-ner | ∼475 MB | 0.98 | 0.91 |
| RecordedFuture Swedish-NER | ∼500 MB | 1.00 | 0.88 |

This comparison is an engineering smoke test, not a definitive public leaderboard: its purpose is to confirm that a browser-sized model stays competitive enough to justify local deployment, not to rank the field on 58 entities. The full tables, all-lowercase results, method notes, and known exceptions are in `BENCHMARKS.md`. On text entirely without capital letters the picture is register-dependent and reported honestly there: in the chat/support register maskera targets, the v18 round ties the strongest measured masking (99.3% of rare out-of-training surnames, on a dedicated 294-sentence eval) while typing them correctly far more often (78.2%, versus 71.4% for the previous release), and on lowercased encyclopedic prose the gap to the case-robust KBLab lowermix model has nearly closed (redaction recall 0.95 versus 0.97). None of the alternative models detects street addresses or fits in a web application.

### 5.8 Latency

Measured **2026-07-13** on an Apple M4 Pro laptop against the curated corpus (average sentence 46 characters); warm figures are per sentence, median and 95th percentile. The browser rows run the production demo bundle with the self-hosted model, exactly what maskera.dev ships. A real first visit adds the network download of the ∼43 MB model; returning visitors read it from the browser cache. The WebGPU path is opt-in and compiles shaders on the first inference (∼45 ms, once per visit).

| Environment | Cold start | Warm (median / p95) | Notes |
|---|---|---|---|
| Browser, WASM (demo default) | 0.30 s | 60 / 66 ms | no GPU needed |
| Browser, WebGPU | 0.34 s | 4.0 / 4.8 ms | opt-in |
| Node (native CPU runtime) | 0.20 s | 3.4 / 5.3 ms | server-side |
| Rules only (`@maskera/core`) | none | ∼0.002 ms | no model, no network |

The figures above are per chat-message-sized input. Longer texts are split into overlapping chunks past the model's 512-token window, so latency grows roughly linearly with length: in Node, warm, ∼500 characters take 17 ms, ∼2,000 characters 66 ms, and ∼10,000 characters 0.39 s (medians; p95 and method in `BENCHMARKS.md`). In practice: masking a chat message costs tens of milliseconds in the browser and single-digit milliseconds on a server, a full support ticket stays under half a second, and the deterministic rule layer is effectively free. The measurement scripts live in the repository, under `packages/ner/eval/` and `apps/demo/scripts/`.

### 5.9 Continuous verification

Quality is not a one-off measurement. Every code change in the project re-runs the evaluation against the published model with an $F_1$ floor (0.90) and a leakage ceiling (0.08) that fail the build on regression, and a weekly automated canary re-measures the published model against the latest runtime version and raises an alarm on drift. Both pipelines are public in the repository's CI.

## 6 Regulatory positioning

*This section describes how maskera relates to the regulatory frameworks; it is not legal advice.*

- **Data minimisation (Art. 5(1)(c) GDPR).** maskera's function is that less personal data reaches third parties: what is masked before the outbound call or write is never processed by the recipient.
- **Pseudonymisation (Art. 4(5) GDPR).** Masked text with a locally held restore key is pseudonymisation: the data can no longer be attributed to a person without the additional information the caller keeps separately. Pseudonymised text may still constitute personal data, but the measure is explicitly recognised as a safeguard in Art. 25 (data protection by design) and Art. 32 (security of processing).
- **Third-country transfers (Chapter V GDPR).** To the extent masking removes personal data before it is sent to a vendor outside the EU/EEA, the transfer analysis is correspondingly narrowed. It does not replace the caller's analysis of whatever is still sent.
- **The EU AI Act.** maskera supports the data-protection objectives and is fully documented and reproducible, which eases the caller's vendor and system documentation. The classification of the caller's AI system is not changed by maskera itself.

**maskera is data minimisation, not a compliance guarantee.** Compliance is decided by the caller's whole system: legal basis, contracts, storage, access, and retention. The caller remains the data controller, and maskera makes no legal claims.

# 7 Limitations

A privacy tool that overstates its accuracy is more dangerous than no tool at all, so the limitations are reported as plainly as the results.

- **The model misses things.** The rule layer is deterministic and highly reliable for structured identifiers, but free-text recognition is probabilistic. Unusual names, heavy misspellings, and OCR noise will sometimes slip through, and all-lowercase text still trails cased text (see the appendix). On the external gold text the leakage is low (2 of 58 entities) but not zero.
- **Four entity types, one language.** The model recognises persons, locations, organisations, and street addresses in Swedish text. Data outside those types, and PII in other languages, is not caught by the model (structured identifiers are caught by the rules regardless).
- **The external test set is small.** 22 sentences give a direction, not a final grade. A larger task-training-independent gold set is the project's top measurement priority, and no annotated test set yet exists for the target domains of support, healthcare, and legal text. The staged plan for closing this gap is public (`docs/GOLD_SET_PLAN.md`): 200+ sentences, on the order of 500 entities, across the authority/legal and support/chat registers, annotated before any model output is seen, with a blind second-pass agreement check and per-register reporting.
- **Defence in depth.** maskera is a strong first layer, not the only one. Keep access controls, processing agreements, logging policies, and human review in high-stakes flows.

# 8 Threat model: what maskera does not protect against

Data minimisation has a boundary, and a reviewer should know where it runs. maskera does not protect against:

- **Sensitive content that is not an entity.** maskera masks identifiers and named entities. A sentence can remain deeply revealing without containing either: a health condition, a workplace conflict, or a situation described precisely enough to point at one person all survive redaction intact. Masking makes text less identifying; it does not make it harmless.
- **Mishandling of the restore mapping.** The placeholder-to-value mapping is the key that undoes the pseudonymisation. If the caller logs it, stores it next to the redacted text, or sends it downstream, the protection is reversed outside maskera's control. Keep the mapping in memory, short-lived, and local.
- **Whatever happens after redaction.** maskera controls what the recipient receives, not what the caller's downstream logic or the recipient does with it. An application that re-attaches identifiers after restoration and then logs the result has left the protected path.
- **Re-identification from context.** As noted under regulatory positioning, redacted text may still constitute personal data. Rare combinations of unmasked facts (a job title, a small town, a date) can identify a person even with every entity masked.

None of this is unusual; it is the standard boundary of any redaction layer. It is stated here so that the caller's remaining controls, described in the next section, are chosen deliberately rather than assumed away.

# 9 Recommended deployment

For production use, the project's operations guide (`docs/PRODUCTION.md`) recommends:

- Redact *before* the text reaches the language model, the log, or the analytics pipeline, never after.
- If the model layer fails, fall back to the rule layer, never to raw text.
- Never log the restore key and never store it next to the redacted text; restoration happens

only inside the caller's environment.

- Self-host the model files if the Hugging Face Hub is not an acceptable dependency; zero external requests are then made.
- Evaluate against your own text inside your environment. Keep the customer-specific corpus and missed values there; share only approved aggregate metrics outside it. The metric that matters is leakage.
- Human review remains in place for high-stakes flows (legal, healthcare, financial decisions).

### 9.1 Reviewer checklist

The same recommendations in checklist form, for pasting into an internal review:

- ☐ Redaction runs *before* the LLM call, the log write, and the analytics event.
- ☐ The restore mapping is never logged and never stored next to the redacted text.
- ☐ Model files are self-hosted (or `allowRemoteModels: false` is set); zero external requests confirmed in the network tab.
- ☐ On model failure the system fails closed to the rule layer, never open to raw text.
- ☐ A customer-controlled domain eval runs inside the customer environment, gated on leakage, with no payload exported.
- ☐ Human review remains in place for high-stakes flows.

## 10 Auditing and verification

Everything claimed in this document can be checked without asking us:

- **The code:** MIT-licensed, open in full at `https://github.com/joelhagvall/maskera`.
- **The model:** published openly on Hugging Face (`joelhagvall/maskera-sv-ner`) with a model card. The shipped artifact is `onnx/model_q4.onnx`, SHA-256 `6f4bf061e9af6827e4ffe82bcfcb84709daa84c5f5ed7a05c2083a3e535fda66`; the current hash always lives in `docs/BENCHMARKS.md`.
- **The numbers:** canonical in `docs/BENCHMARKS.md`, dated and reproducible.
- **The training:** the whole pipeline exists as runnable scripts in `training/`.
- **The network claims:** verifiable in the browser's developer tools at `maskera.dev`.

## Appendix: aggregate known-miss categories

In the historical published-v18 snapshot, two gold entities had no overlapping detection (genuine misses) and no ordinary word was over-redacted. The raw external sentences and entity values are deliberately not retained here; only the failure categories needed to interpret the aggregate remain.

| Failure category | Reading |
|---|---|
| Metonymic location | A building used as the name of an institution was not covered as a location. |
| Bare surname | A sentence-initial surname in declarative prose was not covered; the adjacent full-name form was covered. |

The metonymic-location miss had regressed in v13 and remained in v18. The sentence-initial bare-surname miss was new in v18 and was weighed explicitly against the dedicated 294-sentence rare-surname aggregate. Earlier chat-frame, all-caps, sentence-initial organisation and lowercase-encyclopedic categories are documented as aggregate history in `BENCHMARKS.md`. These misses are why leakage is the primary privacy metric, while CI also gates $F_1$: every failure mode is visible and tracked in the public regression corpus.

In short: maskera does not make text harmless, but it materially reduces what leaves the caller's environment, with reproducible evidence and visible failure modes.