BiVelio Shield

Architecture, Deployment Boundaries, and Reproducible Detection Benchmarks

BiVelio Privacy Research Team

BiVelio Privacy Research

Correspondence and project information: `https://privacy.bivelio.com`

Revised technical report — October 2026
Repository revision `cafd425c`

Abstract

Large language model (LLM) assistants can receive personal data, confidential business information, and credentials through prompts, attachments, retrieved context, and tool calls. BiVelio Shield is a browser-extension and self-hosted gateway system intended to reduce that exposure through deterministic detection, credential handling, optional contextual natural-language processing (NLP), policy-based transformation, and reversible pseudonymization. This report describes the implemented architecture at repository revision `cafd425c` and corrects the boundaries of its guarantees. The browser and gateway do not share one universal pipeline: in the Team configuration, raw text is sent to a names plane in Falkenstein, Germany, before local substitution; in Pro, contextual NLP requires a configured customer-operated gateway; and an external LLM receives a sanitized prompt, not no prompt. Fail-closed behavior applies to specified gateway failures and routes, whereas some browser names-plane failures degrade to deterministic-only processing with a warning.

The frozen catalogue contains 339 entries, of which 204 are credential types classified as secrets; it is therefore not a catalogue of 339 GDPR personal-data types. Fifteen catalogue entries carry an `art9` label, representing eight general Article 9(1) classes plus seven health-related identifier types, not fifteen legal categories. On the cooperative deterministic benchmark (1,000 synthetic documents, seed 1234), the published precision, recall, and F1 are each 1.000 over 4,522 entities: 2,689 across nine structured types and 1,833 across six credential types. Residual verbatim leakage is 0/4,522 and pseudonymization round-trip is 2,689/2,689. Published adversarial results are lower and seed-dependent. Contextual benchmarks with Presidio and GLiNER are reported separately, including non-zero leakage for the deterministic-plus-Presidio stack. Throughput measures detection and scoring in the Python core, not end-to-end redaction. No task-utility benefit is demonstrated. The evidence supports a reproducible engineering baseline with explicit residual risks, not complete privacy protection.

Keywords: large language models; data minimization; pseudonymization; data loss prevention; contextual PII; reproducible benchmarks; GDPR.

Status and scope. This is a corrected engineering report, not a doctoral dissertation, a peer-reviewed article, an independent security assessment, or legal advice. It consolidates measurements already present in the repository and distinguishes observed benchmark results from design intent and open research questions. No new utility, superiority, compliance, or production-security result is claimed.

1. Introduction

Professional use of hosted LLMs creates a data-governance problem at the moment a user submits content. A prompt may contain direct identifiers, health information, business secrets, or reusable credentials; attachments and tool calls add further paths. Training-time memorization and extraction are important but distinct risks [7, 13]. This report concerns the narrower deployment-time problem of what crosses from a user or customer environment to an external model provider.

BiVelio Shield applies detection and policy actions near that boundary. Its deterministic engine recognizes selected structured identifiers and credentials. Optional contextual models recognize names, locations, and other free-text entities. Depending on policy and deployment, the system can pseudonymize, redact, block, or route content locally. Reversible mappings allow later restoration in supported response paths. These mechanisms can reduce exposure for detected spans, but they do not make text anonymous, prove that all sensitive content was found, or establish that a model's answer remains useful.

The purpose of this revised report is threefold: (i) to state the actual trust boundaries and degradation behavior of each deployment; (ii) to publish cooperative, adversarial, and contextual results together with their denominators and limitations; and (iii) to separate regulatory taxonomy from implementation coverage. Prices, purchase arguments, and comparative marketing claims are outside the scientific and engineering scope.

1.1. Contributions and non-claims

The report provides:

- a versioned description of the browser, customer-gateway, and BiVelio-operated names-plane boundaries, including restoration and licensing dependencies;
- a consolidated reproduction protocol for existing deterministic and NLP benchmarks, with cooperative and adversarial outcomes reported side by side;
- a corrected treatment of catalogue size, Article 9, checksum-based detection, Presidio, channel coverage, and fail-closed semantics; and
- an explicit list of unmeasured properties: downstream task utility, comparative superiority, end-to-end latency, production recall, adaptive security, and safe restoration against a hostile page.

Nothing in this report should be read as proof of GDPR compliance, complete prevention of data egress, universal detection, anonymization, or a new doctoral research result.

2. Threat Model and Security Claims

2.1. Assets, boundaries, and adversaries

Assets include personal data in prompts and documents, organization-specific confidential terms, credentials, and the reversible token-to-value mapping. Relevant adversaries include an honest-but-curious provider, a provider or downstream system with broader retention than the user expects, an attacker who later accesses provider-side records, and a hostile script in the provider's web page. Endpoint compromise, malicious software with access to keystrokes or browser storage, and direct network paths that bypass the gateway remain outside the protection guaranteed by the inspected components.

The intended security property is conditional: if all external traffic is forced through an enforcing gateway, the route and content are supported, and a modeled processing failure occurs, the gateway can deny rather than forward the original buffer. This is a control-flow property under those premises. It is not a theorem about detector completeness, browser coverage, contextual inference, or all implementations of the extension.

2.2. What pseudonymization does and does not achieve

Replacing a detected value with a typed marker reduces verbatim disclosure of that value to the external provider. It can also reduce the sensitivity of provider-side logs if neither the original nor the mapping reaches them. The original mapping nevertheless remains a high-value asset, and stable markers can support linkage within their scope. Context may reveal a person, diagnosis, employer, dispute, or transaction even when direct identifiers are removed. European guidance treats pseudonymization as a risk-reduction measure within data processing, not as anonymization [9, 10].

Accordingly, this report uses *sanitized* to mean transformed by the configured detectors and policy, not "free of personal data." Model output creates an additional channel: a model can generate sensitive or unsafe content independently of the input markers. Output scanning and tool-call controls have their own coverage limits and are not evaluated by the input-detection benchmark.

3. Related Work

3.1. Leakage, inference, and application-level attacks

Carlini et al. demonstrated extraction of memorized training data from large language models [7], while Lukas et al. measured personally identifiable information leakage in language-model settings [13]. Staab et al. showed that inference can violate privacy beyond verbatim memorization [18]. Greshake et al. analyzed indirect prompt injection in LLM-integrated applications [11]. These works motivate separating input minimization from model memorization, attribute inference, and agent/tool security; one control does not solve all four.

Broader risk taxonomies and governance frameworks provide useful context [1, 15, 20], but they do not constitute evaluations of BiVelio Shield. Likewise, OWASP guidance motivates credential and prompt-injection controls without establishing the effectiveness of this implementation [16].

3.2. Local de-identification and deployment controls

Microsoft Presidio is a context-aware, pluggable de-identification framework that can run locally in Python or in containers and can also be exposed as a service [14]. It should not be characterized as inherently a cloud API requiring transfer to another processor. BiVelio itself uses Presidio as an optional local NLP component. Meaningful comparisons must therefore hold deployment and detector configuration constant: local deterministic rules, local Presidio, local GLiNER, an ensemble, and any hosted service are distinct conditions.

Network DLP and application gateways also differ by interception point and available actions. Blocking can be appropriate for unsupported content or reusable secrets, while pseudonymization may retain more task context. Neither location nor action alone establishes privacy, utility, or security. This report offers no quantitative head-to-head baseline comparison.

3.3. Prompt pseudonymization and utility evaluation

Local replacement with later back-mapping is not, by itself, a novel research contribution. Pasch and Cha evaluated masking, context-enriched masking, and pseudonymization on personal writing tasks, reporting both incomplete masking and response-quality effects [17]. Hou et al. proposed a general pseudonymization framework for cloud-based LLMs that includes detection, controlled substitute generation, replacement, and task evaluation [12]. Sun et al.'s DePrompt combines prompt desensitization with privacy and utility evaluation [19]. Together, these studies show why marker placement and grammatical fluency are not substitutes for measuring utility and residual disclosure.

The earlier manuscript cited Brown et al.'s *Language Models are Few-Shot Learners* for degradation caused by paraphrasing or confusion prompting. That paper does not evaluate those defenses and the attribution is removed. No equivalent causal claim is made here. BiVelio's present results concern detection, leakage of gold values, round-trip fidelity, and limited timing—not answer quality or task performance.

4. System Architecture and Actual Trust Boundaries

4.1. Detection and policy components

Table 1 separates implementation mechanisms from what the available evidence supports [2, 5].

Checksum gating is useful for rejecting malformed candidates and reducing some false positives. It does not imply recall 1.0 outside a corpus: candidate generation, normalization, token boundaries, overlap resolution, and unsupported formats remain decisive. A Luhn-valid number can also be a non-card identifier. Exact-span evaluation and overlap matching are different claims; the published core benchmark uses overlap matching.

4.2. Three deployment topologies—not one universal pipeline

Table 2 records the deployment-specific processing, storage, and network dependencies [6].

The Team order is material and must not be reversed:

> raw text in browser → names plane (Falkenstein) → span positions
> → local substitution/redaction → sanitized prompt → external LLM.

Thus, "local redaction before any sensitive text leaves the device" is false for the Team names-resolution step. The service is documented as stateless for content, but transient raw-text processing is still processing across a trust boundary. Deterministic minimization applied before that request can reduce what the names plane receives; it does not change the fact that unredacted residual text is submitted for inspection.

For Pro self-hosting, the accurate boundary is also narrower than "prompts never leave the organization." Originals and the vault remain in the customer environment, while the external provider receives the sanitized prompt. If a local model is selected by policy, the routing decision and its actual endpoint must be evaluated separately.

Table 1: Components, functions, and bounded interpretations at `cafd425c`.

Component	Function	Evidence supported	Important boundary
Deterministic engine	Structured patterns, format rules, and checksum gates.	Measured on labelled synthetic core and adversarial corpora.	A checksum validates a delimited candidate; it does not find every candidate or prove exact boundaries.
Credential engine	API keys, tokens, private keys, and assignment forms.	Six credential types in the cooperative benchmark; 204 secret types in the wider catalogue.	Unrecognized formats and bare high-entropy strings can pass. Browser substitution and gateway blocking are different policies.
Contextual NLP	Optional Presidio, spaCy models, and GLiNER-related paths.	Separate synthetic PERSON/LOCATION measurements for specified stacks.	Probabilistic recall, language routing, and configuration matter; Article 9 effectiveness is not benchmarked.
Tenant dictionaries	Organization-defined confidential and matter-specific terms.	A policy mechanism for known terms.	Does not infer unknown confidential semantics or eliminate contextual leakage.
Transformer and vault	Pseudonymize, redact, block, or route locally; maintain reversible mappings.	Round-trip measured for pseudonymized values in the core corpus.	Vault location and restoration depend on deployment; irreversible redaction has no round-trip.
Adapters and walkers	Supported request, response, tool, stream, and document surfaces.	Gateway denial of unknown routes and specified unsupported content under policy.	Protocol drift and uninspected browser transports can bypass coverage.

Table 2: Deployment-specific processing, storage, and network dependencies.

Property	Browser deterministic (Free)	Pro with customer gateway	Team names plane
Contextual NLP	Not supplied by browser-only deterministic configuration.	Requires a deployed gateway and explicit URL/configuration; a licence alone does not activate it.	Configured BiVelio names service unless disabled or replaced by managed policy.
Text crossing before LLM sanitization	No names-plane request; local deterministic processing.	Raw content processed inside customer gateway perimeter.	Raw text sent over TLS to names plane; returned spans applied locally.
External LLM input	Locally transformed prompt on supported paths.	Sanitized prompt forwarded by customer gateway; originals and vault remain in that perimeter.	Locally transformed prompt after names-plane spans and deterministic substitutions.
Vault and restoration	Browser vault; supported responses restored into provider page DOM.	Gateway vault; response restoration in gateway path.	Vault and substitution in browser; restoration writes into page DOM.
Names-plane location	Not applicable.	Customer-selected infrastructure.	BiVelio-operated Hetzner service in Falkenstein (fsn1), Germany—not Frankfurt.
Network dependency	Provider access; no contextual inspection service.	Customer gateway, provider, and licensing/seat mechanisms as applicable.	Names plane, provider, and licensing/seat mechanisms; contextual failures can degrade.

Table 3: Four concepts that must not be conflated.

Concept	Implemented or measurable meaning	What it does not establish
Closure on modeled gateway failures	Unknown endpoints, specified unsupported content, processing errors, and BLOCK decisions denied when traffic traverses the gateway.	Coverage of direct egress, every browser route, every future protocol, or every implementation fault.
Contextual degradation	Some names-plane errors, timeouts, over-length text, or partial responses continue deterministic-only with warning.	Suspension on every uncertainty; a high-confidence false negative is not converted into abstention.
Selective abstention	Possible confidence/risk policy; current low-confidence actions concern detected spans, not universal non-detection.	Calibrated risk–coverage performance; no such evaluation is presented.
Detection completeness	Empirical property measured against specified gold corpus and adversary model.	Universal guarantee derived from checksums, a gate, or zero observed leak.

4.3. Failure semantics

Explicit user consent to send without a protection step is a user decision, not an automatic safety property. Likewise, default gateway blocking of recognized secrets must not be generalized to the extension, which substitutes recognized secrets on its browser paths, or to secrets that no detector recognizes.

4.4. Vault, DOM restoration, and MAIN/ISOLATED worlds

In the browser extension, network interception runs partly in the page's MAIN world, while detection, vault access, and the bridge run in the extension's ISOLATED world. This separation reduces direct exposure of vault contents to page JavaScript, but it is not a complete security boundary. The current design uses a private bridge channel rather than a general window restore operation, and restoration is scoped to conversation-related containers and selected vaults. Those controls narrow earlier bulk cross-conversation oracle risks.

A residual oracle remains by construction: restoration writes original values into a DOM controlled by the provider's page. A hostile page can place a valid marker in an answered container, navigate or link to a conversation, and observe restored text one conversation at a time. Selector scoping does not protect against the page that owns those containers. Consequently, vault confidentiality cannot be inferred solely from ISOLATED-world storage, and restoration against a malicious provider page is not demonstrated safe. Values repeated from another conversation may also remain as markers, which is a restoration-fidelity limitation.

The gateway has a different restoration boundary: its vault and response transformation run in gateway infrastructure, not necessarily on the user's device. Browser round-trip numbers and gateway round-trip behavior should therefore not be presented as measurements of one universal client-side vault.

4.5. Channel and attachment coverage

Table 4 is a coverage matrix from static documentation review, not a claim of exhaustive end-to-end verification [4, 6].

This matrix is intentionally a set of boundaries rather than a single green "protected browser" claim. Each channel requires its own fault-injection and end-to-end tests before a fail-closed property can be asserted.

4.6. Licence verification and check-ins

Offline Ed25519 signature verification is a real mechanism, but it does not imply that all paid extension functions operate without network check-ins. Seat and revocation checks contact the control plane. Certain paid capabilities are disabled when a seat is over cap, revoked, or cannot refresh beyond the documented staleness grace (72 hours in the referenced implementation). Names-plane availability also introduces an interval-dependent heartbeat. These checks carry licensing material and value-free identifiers/counters rather than prompt contents according to the design documentation; they are nevertheless operational network dependencies and must be included in availability analyses.

Table 4: Channel and attachment coverage, degradation, and residual limitations.

Surface	Actual protection or behavior	Degradation / consent	Residual limitation
Modelled browser prompt routes	Deterministic substitution; contextual spans when configured and available.	Explicit choices may permit unprotected sending.	Detection omissions, adapter drift, and route-specific behavior remain.
Names-plane calls	Contextual inspection precedes local substitution in Team/configured paid paths.	Errors, timeout, long text, or partial answers can continue deterministic-only with warning.	Raw residual text reached names plane; names may reach LLM after degradation.
Unknown or guarded browser routes	Some route-blind guards detect selected content; some paths warn rather than substitute.	Behavior depends on route and content class.	Guarded routes can release original bytes; no browser-wide inspected-boundary claim.
Attachments in extension	Team document handling applies local deterministic detection and contextual names inspection when configured.	Free/Pro file flows may pass originals after warning and user decision.	**Team extracted text can also receive configured contextual NLP after local deterministic substitution; scans, images, and unsupported formats remain outside.**
Gateway documents and requests	Supported extraction and recursive content processing; specified unsupported/error/block conditions denied.	Policy and licensed features affect behavior.	Only forced gateway traffic is covered; scans and provider-side file references need separate treatment.
Other modes and transports	Real-time/voice and some agentic, settings, memory, Canvas, Gemini, worker, or transport surfaces limited or outside interception.	Some surfaces only warn or have no active control.	Workers, service workers, new protocols, and direct egress can bypass extension.

5. Catalogue and Regulatory Interpretation

5.1. The number 339 is not a count of GDPR personal-data types

At `cafd425c`, the combined policy catalogue contains 339 entries. Of these, 204 are credential detector types with category `secret`. The residual 135 entries include personal-data-related detectors and researched/roadmap entries; the total is not an official GDPR taxonomy and must not be described as "339 GDPR personal-data types." Credentials may protect organizations and individuals, but they are a separate security class in the implementation.

Table 5: Denominators that answer different questions.

Denominator	Value	Interpretation
Catalogue entries	339	Combined catalogue size at frozen revision; includes secrets and does not imply operational coverage.
Credential types	204	Secret-category entries; excluded from GDPR personal-data category counts.
Other catalogue entries	135	Not a count of enabled detectors, evaluated types, or statutory classes.
Article 9-labelled entries	15	Eight general classes plus seven health identifiers; not fifteen legal categories.
Cooperative benchmark types	$15 = 9 + 6$	Nine structured and six credential types in the 4,522-entity core benchmark.
Contextual targets	PERSON, LOCATION	Two classes in the 800-entity NLP corpus; not an Article 9 evaluation.

Catalogued, implemented, enabled for a particular licence/configuration, and evaluated are four different states. Live and roadmap entries must be distinguished, as must browser and gateway implementations. This report does not infer per-type operational coverage from the catalogue total. A future artifact should generate an inventory by type, jurisdiction, engine, status, plan/configuration, and corpus support directly from frozen code.

5.2. Article 9: eight general classes and fifteen catalogue entries

Article 9(1) lists eight general classes in the terminology used here: (1) racial or ethnic origin; (2) political opinions; (3) religious or philosophical beliefs; (4) trade-union membership; (5) genetic data; (6) biometric data for the purpose of uniquely identifying a natural person; (7) health data; and (8) data concerning a natural person's sex life or sexual orientation [10]. The wording and scope of biometric data matter; not every biometric datum is covered identically.

The implementation maps these to eight general entries: `RACIAL_ETHNIC_ORIGIN`, `POLITICAL_OPINIONS`, `RELIGIOUS_BELIEFS`, `TRADE_UNION`, `GENETIC_DATA`, `BIOMETRIC_DATA`, `HEALTH_DATA`, and `SEX_LIFE_ORIENTATION`. Seven additional health-related identifiers carry the `art9` label: `ES_HEALTH_CARD`, `ES_MEDICAL_RECORD_NO`, `FR_CARTE_VITALE`, `US_MBI`, `PT_SNS_UTENTE`, `BR_CNS`, and `DE_KVNR`. These are concrete national identifier types associated with health processing, not additional general statutory classes. Criminal convictions and offences belong to Article 10, not Article 9, and are not included in this count.

The NLP benchmark does not evaluate these Article 9 classes or the fifteen `art9`-labelled entries. Consequently, neither catalogue breadth nor PERSON/LOCATION results demonstrate effective special-category detection.

5.3. GDPR alignment as engineering analysis

Selective transformation at a prompt or gateway boundary can support Article 5(1)(c) data minimization and Article 25 data protection by design and by default [8, 10]. Reducing original values in external logs may also contribute to an Article 32 risk-management program. These are design mappings, not a legal determination. Lawfulness, purpose limitation, controller/processor contracts, transparency, retention, international transfers, and Article 9 exceptions require organizational and legal analysis.

Pseudonymized prompts can remain personal data, particularly where mappings, contextual attributes, or linkage possibilities persist. The Team names plane processes raw residual text before local redaction and therefore requires its own processor and transfer analysis. A German location does not by itself settle all legal questions. No claim is made that the provider receives no personal data or that using the system establishes compliance.

6. Evaluation

6.1. Frozen revision, data, and reproduction commands

All repository-derived descriptions and numbers in this section refer to `origin/main` commit `cafd425c` (full identifier `cafd425ca95adcf2c8479c65beae223defc4cc19`), which is also the checked-out baseline used for this revision. The benchmark documentation is `docs/BENCHMARKS.md` at that revision [3]. No hidden corpus, newly annotated real-world dataset, or externally validated result is introduced.

The deterministic corpus is seeded and synthetic, with Spanish, Catalan, French, and English connective text, valid positive values, and unlabelled hard negatives. The reported metrics are span-level precision/recall/F1 with overlap matching; residual leak is the fraction of gold values still present verbatim after sanitization; the partial-leak surface counts gold values not fully covered; and round-trip applies to pseudonymized rather than irreversibly redacted values.

For true positives TP, false positives FP, and false negatives FN under the stated overlap criterion,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$\text{F1} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{2}$$

For G gold values, L values leaked verbatim, P pseudonymized values, and R correctly restored pseudonymized values,

$$\text{Residual leak} = \frac{L}{G}, \qquad \text{Round}-\text{trip} = \frac{R}{P}. \tag{3}$$

Throughput is reported as $\Theta_C = C/t$ characters per second and $\Theta_D = D/t$ documents per second, where C is corpus characters, D documents, and t the timed detection-and-scoring interval.

Documented commands, run from the repository root with the project environment installed, are:

```
make benchmark-doc
bpg benchmark --markdown --size 1000
```

```
bpg benchmark --size 1000 --seed 1234 --track core
bpg benchmark --size 1000 --seed 1234 --track adversarial
bpg benchmark --size 1000 --seed 42 --track adversarial
bpg benchmark --size 1000 --seed 7 --track adversarial
make nlp-models
bpg benchmark-nlp --size 400 --seed 2026
make benchmark-nlp
scripts/measure_gliner_gain.py
```

The audit also reproduced the core block with the environment-qualified command:

```
PYTHONPATH=src:. .venv/bin/python -m bivelio_privacy_gateway.cli \
  benchmark --size 1000 --seed 1234 --track core
```

A valid reproduction must record the commit, dependency manifest, model versions, licence context, hardware, load, command, corpus size, seed, and track. The documented NLP command returns no result in a Community context; that constraint is part of reproducibility, not a result. More seeds of the same generator do not create external validity.

6.2. Cooperative deterministic and credential benchmark

The published reference machine for the deterministic block is an Apple M4 Pro (14 cores, 48 GB), macOS 15.7.4, CPython 3.13.7, single-threaded, without NLP models. For size 1,000 and seed 1234, the corpus contains 443,724 characters and 4,522 labelled entities. The total combines nine structured types (2,689 entities) and six credential types (1,833 entities); it is not a count of 4,522 structured identifiers [3].

Table 6: Published cooperative results by evaluated type (size 1,000, seed 1234).

Class	Entity type	Support	Precision	Recall	F1
Structured	CREDIT_CARD	305	1.000	1.000	1.000
Structured	EMAIL	314	1.000	1.000	1.000
Structured	IBAN	300	1.000	1.000	1.000
Structured	IP_ADDRESS	289	1.000	1.000	1.000
Structured	PHONE_AD	283	1.000	1.000	1.000
Structured	PHONE_ES	277	1.000	1.000	1.000
Structured	SPANISH_CIF	338	1.000	1.000	1.000
Structured	SPANISH_DNI	305	1.000	1.000	1.000
Structured	SPANISH_NIE	278	1.000	1.000	1.000
Structured subtotal (9 types)		2,689	—	—	—
Credential	ANTHROPIC_API_KEY	307	1.000	1.000	1.000
Credential	AWS_ACCESS_KEY_ID	270	1.000	1.000	1.000
Credential	GITHUB_TOKEN	326	1.000	1.000	1.000
Credential	JWT	307	1.000	1.000	1.000
Credential	OPENAI_API_KEY	315	1.000	1.000	1.000
Credential	PRIVATE_KEY	308	1.000	1.000	1.000
Credential subtotal (6 types)		1,833	—	—	—

Micro precision, recall, and F1 are each 1.000 (TP = 4,522, FP = 0, FN = 0). Residual verbatim leakage is 0.0000% (0/4,522), the partial-leak surface is 0/4,522, pseudonymization round-trip is 100.0000% (2,689/2,689), and irreversible-redaction leaks are zero. These are corpus-specific observations. They do not prove exact detection of every well-formed identifier, performance on real text, or coverage of all 339 catalogue entries.

The gold generator is intentionally independent of detector checksum code and is anchored to external worked examples. Nevertheless, the corpus mints valid values and selected hard negatives; it does not represent every malformed adversarial input. A documented live defect illustrates the gap: in both Python and TypeScript, "BTW NL123456789B01 on file" can produce an over-wide validated IBAN span that consumes six characters of following prose. This is over-redaction rather than a gold-value leak, so the headline leak metrics remain zero and do not detect the defect [3].

6.3. Throughput is detection plus scoring

The published cooperative block reports 381,110 characters/s (859 documents/s) on its reference run. Across twelve runs on the reference machine, the documentation reports 350–399 kchar/s (788–900 documents/s), with 192 kchar/s under contention. The audit's separate reproduction recorded 328,285.6 characters/s. These values are machine- and load-specific observations, not a product constant.

Most importantly, the timer covers `engine.detect` and `_score_document`. Transformation, vault operations, network I/O, model inference, document extraction, browser work, and restoration occur outside that interval. The metric is therefore Python-core detection-and-scoring throughput, not end-to-end redaction throughput. It cannot establish unobservable interaction latency or document-scale end-to-end performance. Comparisons with human typing speed are omitted because they do not answer that question.

6.4. Adversarial track

The adversarial track changes placement and representation: fullwidth and Arabic-Indic digits, zero-width splits, compatibility prefixes, neighbouring combining marks, and adjacent digits. Published measurements from 9 August 2026 use size 1,000 and seeds 1234, 42, and 7. Reporting only cooperative seed 1234 would select the most favorable condition [3].

Table 7: Published core and adversarial results across three seeds.

Track	Seed	Precision	Recall	Residual leak	Partial	Round-trip
core	1234	1.0000	1.0000	0.00%	0	1.000000
core	42	1.0000	1.0000	0.00%	0	1.000000
core	7	1.0000	1.0000	0.00%	0	1.000000
adversarial	1234	0.9940	1.0000	0.00%	0	1.000000
adversarial	42	0.9874	0.9989	0.00%	0	0.999641
adversarial	7	0.9894	1.0000	0.00%	0	1.000000

Across these draws, adversarial precision ranges from 0.9874 to 0.9940, minimum recall is 0.9989, and minimum round-trip is 0.999641. The observed residual and partial-leak lines remain zero, but zero observed events in this synthetic track are not a probability-zero guarantee. Per-placement precision is lower in some cases: 0.9514–0.9540 for a leading digit and 0.9818–0.9840

for zero-width split or interior combining-mark placements. No confidence intervals or independent adversarial dataset are reported.

6.5. Contextual NLP results

The separate Presidio benchmark uses 400 synthetic documents, 800 PERSON/LOCATION gold entities, seed 2026, and medium spaCy models for Spanish, Catalan, French, English, Portuguese, and German. Oracle mode supplies the correct language; auto mode uses the built-in heuristic. Results were re-measured on 9 August 2026 under CPython 3.13.7 and reproduced identically on two runs [3].

Table 8: Published contextual NLP results; stacks and modes are not interchangeable.

Measurement	Configuration	Result	Scope / caveat
Detection	Presidio, oracle language	P 0.9061; R 0.9888; F1 0.9456	400 synthetic documents; 800 PERSON/LOCATION entities.
Detection	Presidio, automatic language	P 0.9257; R 0.9500; F1 0.9377	Heuristic language selection included.
Language routing	Dependency-free heuristic	Accuracy 0.86	Six clean, short synthetic languages; not code-switching.
Residual leak	deterministic + Presidio	4.875% (39/800)	Gold PERSON/LOCATION values verbatim after sanitization.
Residual leak control	deterministic + null NER	100.000% (800/800)	Confirms active leakage measurement.
Residual leak	deterministic + Presidio + GLiNER	0.000% (0/800)	Same synthetic corpus/seed; measured 12 September 2026.
Wall time	null / Presidio / +GLiNER	0.1 s / 3.3 s / 7.1 s	Development laptop; GLiNER approximately 2.2× Presidio stack.
NLP throughput	Presidio benchmark	160–181 documents/s	Single-thread, three contended-machine runs; load-dependent.

The 4.875% figure belongs to deterministic + Presidio, not to the GLiNER-augmented names plane. Conversely, 0/800 for deterministic + Presidio + GLiNER is an observation on one synthetic PERSON/LOCATION corpus, not a production guarantee. It says nothing about real notarial prose, addresses, Article 9 classes, misspellings, out-of-distribution languages, or adaptive attacks. The documented timing for the NLP scoring loop excludes extra detectors, so wall time—not `NLPReport.seconds`—is the relevant comparison for the GLiNER stack.

The deterministic and NLP timing figures are not directly comparable workloads. Their orders of magnitude nevertheless show why a contextual stack needs its own end-to-end latency and resource evaluation. No p50/p95/p99 browser or gateway latency is claimed.

6.6. Metric limitations and absent evaluations

- Overlap matching can credit a prediction without requiring exact span boundaries; additional prose consumed by a span is not always penalized by the scorer.
- Whole-value residual leakage does not detect every partial disclosure; the separate partial-coverage metric is necessary and remains corpus-specific.
- Round-trip fidelity in a benchmark does not establish robust restoration under streaming, paraphrased markers, cross-conversation references, or hostile DOM behavior.
- Synthetic in-distribution corpora and multiple seeds do not establish performance on independent, multilingual, real-world data.
- No quantitative baseline establishes superiority to local Presidio, GLiNER alone, another gateway, or a no-redaction condition under matched utility.
- No task benchmark measures answer quality, reasoning, writing, extraction, translation, or user effort after redaction and restoration. Utility preservation is a hypothesis, not a demonstrated result.
- No per-channel fault-injection campaign establishes fail-closed behavior for every browser, gateway, streaming, tool, and attachment path.

7. Discussion, Validity, and Research Agenda

7.1. What the current evidence supports

The repository contains an unusually explicit set of executable checks and published failure modes. The cooperative benchmark is reproducible at its stated denominator, and the adversarial and NLP sections expose degradation that a single 1.000 headline would conceal. The architecture also makes a useful engineering distinction between deterministic identifiers, credentials, contextual entities, and organization-defined terms.

The strongest defensible conclusion is therefore narrow: BiVelio Shield is a testable data-minimization implementation whose deterministic core performs perfectly on the specified cooperative synthetic corpus, whose performance changes under adversarial placements and contextual models, and whose deployment-specific trust boundaries materially affect privacy. It is not evidence of complete sensitive-data detection, anonymous provider-bound text, universal fail-closed transmission, or preserved downstream utility.

7.2. Threats to validity

Construct validity.
Overlap F1, verbatim leakage, and catalogue counts measure different constructs and cannot substitute for exact-span correctness, residual inference risk, or operational coverage.

Internal validity.
Shared generators, configuration, and language heuristics affect results; timing is load-sensitive.

External validity.
Synthetic names, locations, identifiers, and credentials do not represent real professional text or adaptive adversaries.

Security validity.
Static control-flow reasoning and unit tests do not prove that all external traffic traverses the boundary or that a hostile page cannot exploit restoration.

Legal validity.
Engineering labels do not determine statutory categories or compliance.

No statistical inference beyond the documented observations is attempted. Future comparative studies should pre-specify endpoints, use independent test sets, report confidence intervals and paired analyses, and separate exploratory from confirmatory results.

7.3. Research required before stronger claims

A credible research program should evaluate policies that choose among forwarding, pseudonymization, redaction, abstention, and local routing under a specified threat model. Direct leakage, linkage, inference, task-utility loss, latency, cost, over-redaction, and user overrides need separate measures. Baselines should include local deterministic rules, local Presidio, GLiNER, the ensemble without policy, and the complete system under matched conditions.

Security experiments should mutate endpoints and schemas, cross streaming chunk boundaries, inject timeouts and corrupt policies, test packaged browser builds, and verify that no error path emits the original buffer. Restoration experiments should reproduce DOM-oracle attacks, test tenant/session/tab isolation, and measure broken or paraphrased markers. Utility experiments should compare original, masked, opaque-token, realistic-pseudonym, suppressed, and local-model conditions across tasks and frozen model versions. These are proposed experiments, not results reported here.

8. Conclusion

BiVelio Shield provides concrete mechanisms for reducing prompt-borne exposure before content reaches an external LLM on supported paths. Its engineering evidence is strongest for a frozen deterministic benchmark and weakest where broad claims previously obscured deployment differences, contextual recall, channel coverage, and restoration risk. Correcting those claims does not diminish the reproducible results; it makes their meaning usable.

At `cafd425c`, the accurate summary is: 339 catalogue entries include 204 credential types; fifteen `art9`-labelled entries represent eight general legal classes plus seven health-related identifiers; the 4,522-entity cooperative result comprises nine structured and six credential types; adversarial and NLP measurements show non-trivial limitations; and throughput covers detection plus scoring only. Team sends raw text to a Falkenstein names plane before local redaction, Pro contextual processing requires a configured gateway, and external providers receive sanitized rather than absent prompts. Fail-closed statements are limited to modeled paths and failures, with contextual degradation documented elsewhere. No task-utility benefit, production completeness, anonymization, or universal security property has been demonstrated.

A. Resumen en español

Objeto y alcance. Este informe técnico describe BiVelio Shield en la revisión `cafd425c` del repositorio y corrige las afirmaciones del manuscrito anterior. No es una tesis doctoral, un artículo revisado por pares, una auditoría independiente ni un dictamen jurídico. El sistema combina detección determinista, detección de credenciales, PLN contextual opcional, políticas de transformación y seudonimización reversible para reducir la exposición de datos en interacciones con LLM. La reducción de exposición para fragmentos detectados no equivale a anonimización ni garantiza que todo dato sensible sea detectado.

Arquitectura real. No existe una única canalización con idénticas garantías en todos los despliegues. En Team, el texto crudo residual se envía al plano de nombres ubicado en Falkenstein (Alemania), el servicio devuelve posiciones de entidades y la sustitución se realiza después localmente; por tanto, la redacción no precede a toda salida de texto del dispositivo. En Pro, la capa contextual requiere un *gateway* desplegado y configurado; la licencia por sí sola no la activa. El proveedor externo recibe el *prompt* sanitizado: los originales y la bóveda permanecen en el perímetro correspondiente, pero el *prompt* sanitizado sí sale. En navegador, la bóveda reside en la extensión y la restauración escribe en el DOM; en *gateway*, ambas operaciones pertenecen a la infraestructura del *gateway*.

Fallos y cobertura. El cierre *fail-closed* se limita a rutas y fallos modelados, especialmente cuando todo el tráfico atraviesa el *gateway*. Algunos errores, tiempos de espera, textos largos o respuestas parciales del plano de nombres degradan el envío a detección determinista con aviso, en lugar de suspenderlo. Esto no es abstención selectiva ni prueba de detección completa. La cobertura varía por ruta, transporte, modo y adjunto. Los adjuntos Team aplican detección determinista local y, cuando está configurado, también análisis contextual de nombres del texto extraído; otras rutas pueden limitarse a avisar o permitir originales tras consentimiento. Los escaneos, las imágenes y los formatos no compatibles quedan fuera. La separación MAIN/ISOLATED reduce exposición directa, pero la página hostil todavía puede explotar la restauración en el DOM como oráculo dentro de su alcance. La verificación Ed25519 *offline* coexiste con *check-ins* de asientos y revocación y con desactivación tras el período de gracia.

Taxonomía corregida. El catálogo contiene 339 entradas, pero 204 son tipos de credenciales clasificados como secretos; no existen 339 "tipos de datos personales GDPR" oficiales en esa cifra. El artículo 9(1) contiene ocho clases generales. Las 15 entradas etiquetadas `art9` corresponden a esas ocho clases más siete identificadores sanitarios concretos, no a quince categorías jurídicas. El *benchmark* no evalúa la eficacia sobre esas clases.

Resultados existentes. En el corpus cooperativo sintético de 1.000 documentos y semilla 1234 hay 4.522 entidades: 2.689 de nueve tipos estructurados y 1.833 de seis tipos de credenciales. Se publican precisión, *recall* y F1 de 1,000; fuga *verbatim* residual de 0/4,522; y restauración de 2,689/2,689 valores seudonimizados. Estas cifras no se derivan universalmente de MOD-97 o Luhn. En el *track* adversarial, la precisión publicada varía entre 0,9874 y 0,9940, el *recall* mínimo es 0,9989 y la restauración mínima 0,999641 según la semilla. En PLN, Presidio obtiene $P = 0{,}9061$ y $R = 0{,}9888$ con idioma *oracle*, y $P = 0{,}9257$ y $R = 0{,}9500$ en modo automático; la exactitud del detector de idioma es 0,86. La fuga de deterministic + Presidio es 39/800 (4,875 %); al añadir GLiNER se observa 0/800 en ese corpus sintético de PERSON/LOCATION, sin evidencia equivalente para texto real ni artículo 9.

Interpretación. El caudal publicado mide detección y puntuación del núcleo Python, no redacción de extremo a extremo. No se ha demostrado utilidad de las respuestas, superioridad frente a *baselines*, latencia extremo a extremo, seguridad universal de restauración ni cobertura efectiva de todos los canales. Presidio puede ejecutarse localmente y no debe presentarse como una API de nube intrínseca. La seudonimización reduce ciertos riesgos, pero el contexto residual puede seguir revelando atributos. Estas conclusiones limitadas sustituyen las afirmaciones absolutas y promocionales del original.

B. Declarations

Conflict of interest

BiVelio develops and distributes the system described in this report. The report is company-authored and BiVelio has a commercial interest in the system. No independent validation or absence of bias is claimed. Readers should treat the results as an internal engineering evaluation and reproduce them independently before relying on them.

Data and artifact availability

The evidence sources are the repository files and executable benchmark generators at commit `cafd425ca95adcf2c8479c65beae223defc4cc19`, especially `docs/BENCHMARKS.md`, `docs/ARCHITECTURE.md`, `docs/THREAT_MODEL.md`, `docs/DEPLOYMENT-TOPOLOGY.md`, and `packages/extension/docs/KNOWN-LIMITS.md`. All reported benchmark corpora are synthetic and generated by the documented commands; no new personal-data dataset was collected for this revision. Access to licensed NLP components and models is required for some commands, and repository or artefact availability may be restricted by the project's distribution terms. This report does not claim that every component is open source.

Funding

The work is presented as internal BiVelio engineering documentation. No external funding is claimed; detailed funding information was not supplied for this revision.

Authoring and AI-assistance statement

Authorship is attributed at team level because no verified individual author list was supplied. AI-assisted language editing and document generation were used for this revision. Technical claims and numbers were checked against the cited repository evidence; no AI system is designated as an author. Any venue-specific disclosure requirements remain the responsibility of the submitting authors.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, 2021. doi: 10.1145/3442188.3445922.

[2] BiVelio. *Architecture*. BiVelio, 2026. `docs/ARCHITECTURE.md`, repository revision `cafd425c`.

[3] BiVelio. *BV-PRGA Benchmarks*. BiVelio, 2026. `docs/BENCHMARKS.md`, repository revision `cafd425c`.

[4] BiVelio. *BiVelio Shield—Known Detection Limits*. BiVelio, 2026. `packages/extension/docs/KNOWN-LIMITS.md`, repository revision `cafd425c`.

[5] BiVelio. *Threat Model*. BiVelio, 2026. `docs/THREAT_MODEL.md`, repository revision `cafd425c`.

[6] BiVelio. *Deployment Topology*. BiVelio, 2026. `docs/DEPLOYMENT-TOPOLOGY.md`, repository revision `cafd425c`.

[7] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650. USENIX Association, 2021. URL `https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting`.

[8] European Data Protection Board. Guidelines 4/2019 on article 25 data protection by design and by default. Version 2.0, 2020. URL `https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-42019-article-25-data-protection-design-and_en`.

[9] European Data Protection Board. Guidelines 01/2025 on pseudonymisation, 2025. URL `https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf`.

[10] European Parliament and Council of the European Union. Regulation (eu) 2016/679 (general data protection regulation), 2016. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[11] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90. ACM, 2023. doi: 10.1145/3605764.3623985.

[12] Shilong Hou, Ruilin Shang, Zi Long, Xianghua Fu, and Yin Chen. A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation, 2025. URL `https://arxiv.org/abs/2502.15233`.

[13] Nils Lukas, Ahmed Salem, Robert Sim, Suman Mukherjee, S. S. Raman, Michael Backes, Sebastian Zimmeck, and Florian Tramèr. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy*, pages 346–363. IEEE, 2023. doi: 10.1109/SP46215.2023.10179300.

[14] Microsoft. Presidio: Data protection and de-identification sdk, 2026. URL `https://microsoft.github.io/presidio/`. Official documentation and source repository; accessed October 2026.

[15] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023.

[16] OWASP Foundation. Owasp top 10 for large language model applications, 2023. URL `https://owasp.org/www-project-top-10-for-large-language-model-applications/`.

[17] Simon Pasch and Min Chen Cha. Balancing privacy and utility in personal llm writing tasks: An automated pipeline for evaluating anonymizations. In *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pages 32–41. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.privatenlp-main.3.

[18] Robin Staab, Mark Vero, Mar Gascon, Nicholas Carlini, Florian Tramèr, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In

The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p.

[19] Xiongtao Sun, Gan Liu, Zhipeng He, Hui Li, and Xiaoguang Li. Deprompt: Desensitization and evaluation of personal identifiable information in large language model prompts, 2024. URL https://arxiv.org/abs/2408.08930.

[20] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229. ACM, 2022. doi: 10.1145/3531146.3533088.