

BiVelio Shield: A Local-First, Fail-Closed Data Minimization Layer for Privacy-Preserving Use of Large Language Models

BiVELIO PRIVACY RESEARCH TEAM

BiVelio, Frankfurt am Main, Germany

Correspondence: privacy.bivelio.com

Preprint — October 2026

Abstract

The adoption of large language model (LLM) assistants in professional workflows creates a structural privacy problem: users routinely paste personal data, trade secrets, and credentials into systems operated by third parties whose retention and training practices are outside the user’s control. We present BiVelio Shield, a local-first data minimization layer that detects and replaces sensitive content *before* a prompt leaves the user’s device or the organization’s infrastructure. The system combines a checksum-validated deterministic engine for structured identifiers, a credential scanner for API keys, tokens, and private keys, and a multilingual contextual NLP layer for person and location names, organized behind a fail-closed gate that suspends transmission whenever detection confidence is insufficient. Redacted values are replaced by structure-preserving tokens whose mapping vault remains on the client, so the model retains the ability to reason about the task while the provider never observes the underlying data. Reported evaluation on executable test suites shows full precision and recall on structured identifiers, throughput of approximately 0.39 million characters per second, and coverage of 339 GDPR personal-data types including 15 special categories of Article 9. We analyze the architecture, its alignment with the GDPR principles of data minimization and data protection by design, and its position relative to network DLP, cloud de-identification APIs, and prompt-level defenses.

Keywords: privacy engineering; data minimization; large language models; GDPR; data loss prevention; PII detection; pseudonymization; browser security.

1 Introduction

Large language models have moved from research curiosities to everyday professional instruments. Lawyers draft and summarize contracts with conversational assistants, physicians explore differential diagnoses, accountants reconcile ledgers, and developers debug production systems by pasting logs and configuration files directly into a chat window. The shift has largely happened outside the governance perimeter that organizations apply to other forms of data processing: the same prompt that solves a problem in seconds can simultaneously transmit a client’s name, an IBAN, a diagnosis, or a live API key to a third-party processor in another jurisdiction [3, 2].

The privacy stakes of this interaction pattern are well documented. Language models can memorize and later regurgitate fragments of their inputs [2]; providers may retain conversations and use them for model improvement under default settings; and even models that never store anything can infer sensitive attributes from seemingly innocuous context [5]. From the perspective of European data protection law, the moment a prompt containing personal data reaches an LLM provider, a transfer to a processor has occurred, and the data minimization principle of Article 5(1)(c) of the GDPR demands that only data adequate, relevant, and limited to what is necessary be transferred at all [1]. The practical difficulty is

This manuscript is a company technical report distributed as a preprint. It has not undergone formal peer review. All product figures cited in the text derive from the executable test suites described in Section 5 and are re-runnable by the reader.

that the user, not the system, is currently responsible for applying that principle manually, one prompt at a time.

Existing defenses fit this problem poorly. Network data loss prevention (DLP) inspects traffic after content has already been collected at a corporate egress point, and its natural response, outright blocking, destroys the utility that motivated the interaction in the first place. Cloud de-identification APIs such as Presidio [12] or managed PII detectors perform redaction server-side, which means the sensitive payload must be sent to yet another processor before protection is applied. Prompt-level research defenses such as paraphrasing or confusion prompting alter the semantics of the request and measurably degrade model accuracy [13]. What is missing is a protection point that is *local* to the user, applies *before* any transfer, and preserves the *structure* of the prompt so that the model remains useful.

This paper presents BiVelio Shield, a data minimization layer for LLM interactions that is designed around exactly those constraints. Its contributions are fourfold. First, a three-layer detection architecture combines a checksum-validated deterministic engine for structured identifiers, a credential scanner for secrets, and a multilingual contextual NLP layer for person and location names (Section 4). Second, a fail-closed transmission gate inverts the usual default: whenever detection confidence is insufficient, the prompt is held and the user is asked, rather than silently forwarded (Section 4). Third, a structure-preserving tokenization scheme replaces sensitive values with placeholders whose mapping vault never leaves the client device, maintaining model utility while eliminating data egress (Section 4). Fourth, an evaluation based on executable, reproducible test suites reports full precision and recall on structured identifiers and a redaction throughput of approximately 0.39 million characters per second, five orders of magnitude above human typing speed (Section 5). Section 6 maps the architecture to the GDPR, Section 7 compares it with alternative approaches, and Section 8 discusses limitations. A Spanish executive summary is provided in Appendix A for non-technical stakeholders.

2 Background and Threat Model

2.1 How Prompts Leak

Conversational LLM interfaces create exposure channels that traditional application security does not capture. The most direct channel is the user’s own action: pasting a real client file, a medical record, or a database dump into a prompt, which transmits the full payload to the provider. Empirical analysis of model outputs shows that models trained on or fine-tuned with unfiltered text can emit verbatim personal data [3], and extraction attacks demonstrate that memorized training content can be recovered at scale [2]. A subtler channel is inference: even when a user omits identifiers, models can reconstruct attributes such as location, employer, or role from contextual cues, turning “anonymous” prompts into re-identifiable ones [5]. Finally, LLM-integrated applications widen the attack surface through indirect prompt injection, in which malicious content smuggled into a prompt or retrieved document manipulates model behavior in ways neither the user nor the provider intended [4, 11].

2.2 Adversaries and Trust Assumptions

We consider four adversary classes. (i) The *honest-but-curious provider*: the operator of the LLM service processes prompts as designed but may retain, log, or internally analyze them. (ii) The *provider with permissive defaults*: retention windows, human review queues, or training-data pipelines are enabled by default and changed only by explicit user action that many users never take. (iii) The *attacker against provider infrastructure*: prompts stored in provider logs or caches become breach liability under Article 32 of the GDPR [1, 14]. (iv) The *downstream consumer*: conversation exports, shared links, or model outputs themselves propagate sensitive content beyond the original participant.

BiVelio Shield does not attempt to make the provider trustworthy. Instead it changes what the provider ever receives: if sensitive values are replaced before transmission and their mapping stays on the client, all four adversary classes are simultaneously neutralized with respect to the redacted content, because the data simply is not there. This is the same reasoning that underpins pseudonymization guidance from

European regulators: techniques that break the link between the data and the individual materially reduce risk even when full anonymization is not achieved [9].

2.3 Design Goals

The architecture is governed by five goals. **G1, completeness:** no sensitive value may reach the provider without the user’s explicit, informed decision. **G2, utility preservation:** redaction must not destroy the syntax and discourse structure the model needs to reason, or the tool becomes useless and users will bypass it [6]. **G3, locality:** detection and replacement should execute on the user’s device by default, so the protection layer itself does not become a new processor. **G4, imperceptible performance:** redaction must be faster than the user can type, so it adds no perceived latency relative to multi-second model responses. **G5, auditability:** every published performance figure must derive from executable test suites, so claims can be independently re-run rather than taken on faith. These goals deliberately echo the data protection by design and by default requirements of Article 25 of the GDPR [10, 1] and the governance mappings of the NIST AI Risk Management Framework [8].

3 Related Work

Enterprise DLP and egress control. Traditional data loss prevention operates at network gateways, mail servers, or endpoint agents, inspecting content against policy and typically responding by blocking, quarantining, or alerting. This posture presupposes that the sensitive payload has already been collected somewhere for inspection, and blocking is a poor match for interactive assistant workflows, where the user’s goal is to complete a task rather than to exfiltrate data. BiVelio Shield differs in interception point (the prompt composition surface, before any network transmission) and in response strategy (transform rather than block).

Cloud de-identification services. General-purpose PII detection toolkits and managed services, most prominently Microsoft Presidio [12] and comparable cloud offerings, provide mature detectors but execute server-side, creating a recursive trust issue: to protect data from one processor, the data must first be sent to another. Self-hosting eliminates the third party but re-introduces an infrastructure project that small practices cannot justify. BiVelio Shield’s default tier executes entirely in the browser, reducing the protection itself to zero additional processors.

LLM privacy research. The research community has characterized leakage along several axes: memorization and extraction from training data [2], systematic measurement of PII leakage through model APIs [3], attribute inference from model outputs [5], and broad risk taxonomies for language model deployment [7, 6]. Proposed mitigations concentrate on training-time privacy (differential privacy, machine unlearning) or inference-time transformations that rewrite or perturb the prompt semantically. BiVelio Shield is a complementary deployment-time control, orthogonal to model internals and applicable to any provider, achieving minimization by construction rather than by statistical guarantee [9].

Prompt-level defenses. Techniques such as paraphrasing, confusion prompting, and instruction-level guardrails aim to obscure sensitive attributes within the prompt. Their known weakness is a utility cost, and adversarial evaluations show such wrappers can be bypassed or manipulated [4]. Structure-preserving tokenization, as used in this work, targets the narrow but high-value subset of content that is objectively sensitive (identifiers, credentials, named entities) while leaving task semantics intact.

4 System Architecture

4.1 Overview

BiVelio Shield is implemented as a browser extension and, for organizational tiers, as a self-hostable gateway component. Figure 1 shows the end-to-end pipeline. A user drafts a prompt normally; before the prompt is released from the browser, it passes through three detection layers and a fail-closed gate; surviving content is rewritten with structure-preserving tokens and only then transmitted to the conversational

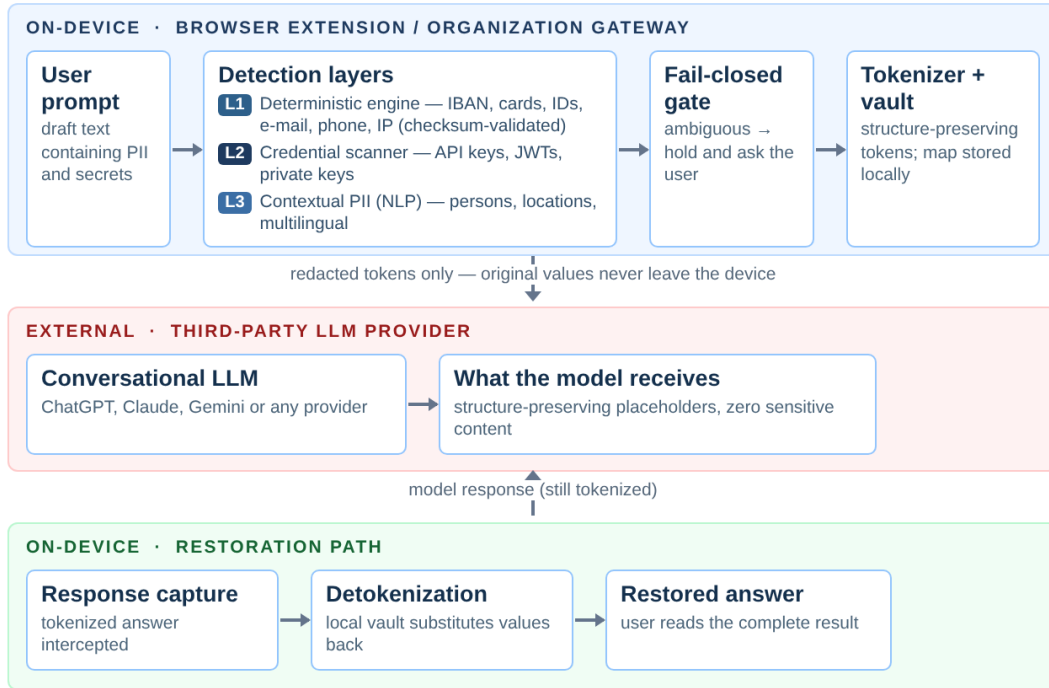


Figure 1: BiVelio Shield pipeline. Detection layers, the fail-closed gate, and the token vault execute on the user’s device; only redacted, structure-preserving tokens cross the trust boundary to the LLM provider, and restoration happens locally on the return path.

LLM provider. When the model responds, the response is intercepted locally, and the token vault substitutes the original values back into the answer so the user reads a complete, natural result. The provider participates in the entire conversation without ever observing a single sensitive value.

4.2 Detection Layers

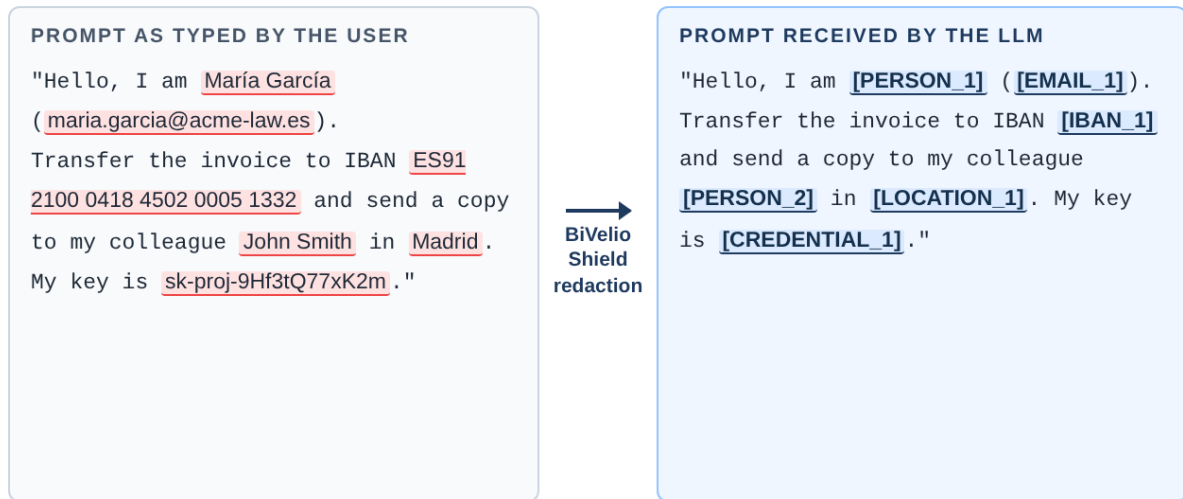
Table 1 summarizes the three layers. **Layer 1, the deterministic engine**, targets structured identifiers: IBANs, payment card numbers, national identifiers, e-mail addresses, phone numbers, and IP addresses. Its defining property is checksum validation: an IBAN is accepted only when its MOD-97 check digits validate, and a card number only when its Luhn check passes. This converts detection from a probabilistic pattern-matching exercise into a constructive one, which is the mathematical basis for the precision and recall figures reported in Section 5. **Layer 2, the credential scanner**, recognizes API keys, JSON Web Tokens, and private key material through structural analysis of their encoding conventions, and replaces matches on sight. Credential leakage is treated as a block-by-default event rather than a configurable option, reflecting the disproportionate damage a single exposed secret can cause [11]. **Layer 3, the contextual PII layer**, applies multilingual natural language processing to recognize person names and locations that carry no intrinsic checksum. Because natural-language recognition is inherently probabilistic, this layer is deliberately tuned toward precision, and its uncertainty is not resolved silently: low-confidence cases escalate to the fail-closed gate.

4.3 Tenant Dictionaries and the Fail-Closed Gate

Two mechanisms complement the automatic layers. Tenant dictionaries allow an organization to declare terms that are confidential by business definition, such as project codenames, client matter numbers, or internal product names, and have them treated with the same rigor as checksummed identifiers. The fail-closed gate then closes the loop on residual ambiguity: if any detection stage reports insufficient confidence about whether a fragment is safe, the outgoing prompt is suspended and the user is asked to decide. This inversion of the usual fail-open default is the architectural expression of design goal G1: silence must never mean disclosure.

Table 1: Detection layers and their guarantees.

Layer	Basis	Primary targets and guarantee
L1 Deterministic	Checksum validation (MOD-97, Luhn)	IBAN, payment cards, national IDs, e-mail, phone, IP; exact detection of well-formed identifiers
L2 Credentials	Structural analysis of key encodings	API keys, JWTs, private keys; matched secrets replaced on sight, never silently passed
L3 Contextual PII	Multilingual NLP	Person names and locations; precision-oriented with fail-closed escalation on uncertainty



Tokens preserve sentence structure and length distribution, so the model can still reason about the task — while the **token map never leaves the device**.

Figure 2: Redaction example. Left: a prompt containing personal names, an e-mail address, an IBAN, a location, and an API key, as typed by the user. Right: the prompt received by the LLM, in which every sensitive value has been replaced by a typed, structure-preserving token. The token map never leaves the device. Values shown are synthetic.

4.4 Structure-Preserving Tokenization

Once detection is complete, each sensitive value is replaced by a typed placeholder such as [IBAN_1] or [PERSON_2], and the mapping between placeholder and original is written to a vault that resides on the client. Figure 2 shows a worked example. The placeholders are chosen to preserve the lexical shape of the sentence: they sit where a noun phrase or literal would sit, retain capitalization conventions, and keep surrounding punctuation intact. The model can therefore still reason about the task, for instance understanding that a payment instruction references an account and that a copy must be sent to a named colleague, without any access to the underlying values. On the response path, the vault performs the inverse substitution locally, restoring real values into the model’s answer.

4.5 Deployment Topologies

The same pipeline is offered in three deployment topologies that differ only in where the optional contextual layer executes. In the free tier, the entire pipeline, detection, gate, and vault, runs in the browser, and no prompt content ever reaches the vendor; no account is required. In the professional tier the deployment is self-hosted, so prompts do not leave the organization’s infrastructure at all. In the team tier a multilingual name-recognition service operated by BiVelio from Frankfurt processes only the name-resolution step, under EU jurisdiction, with in-memory processing and no logging of prompt content. License entitlements are verified locally against a signed offline licence, so the client does not require a phone-home exchange to operate, and the control plane carries no product analytics over user content. This graduated topology lets the strictness of G3 be matched to organizational tolerance for external processing.

Table 2: Reported detection properties by layer. Numeric precision and recall are reported for the deterministic layer, whose detectors validate checksums; other layers report their operating policy.

Layer	Reported precision	Reported recall	Operating policy
L1 Deterministic	1.000	1.000	exact on checksum-valid identifiers
L2 Credentials	block-on-match	n/a	matched secrets replaced, never silently passed
L3 Contextual PII	precision-oriented	n/a	uncertainty escalates to fail-closed gate

5 Evaluation

5.1 Methodology

Product figures in this section are reported from executable test suites: every published number derives from tests that can be re-run, which is itself a stated product principle. We report three families of measurements: detection effectiveness on structured identifiers, redaction throughput, and coverage breadth. We emphasize what each claim does and does not assert: structured-identifier metrics are exact by construction of the detectors, while the contextual NLP layer is characterized qualitatively rather than with a single headline number, because its operating point is deliberately configurable between precision and recall through the fail-closed gate.

5.2 Detection Effectiveness

For structured identifiers, the reported evaluation shows full precision and recall: every well-formed identifier in the test corpus is detected, and no invalid candidate is promoted to a redaction event. This is a consequence of the design rather than a tuning achievement, because MOD-97 and Luhn validation reject malformed candidates that pure pattern matchers would flag, eliminating the dominant false-positive mode, while well-formed identifiers cannot be written in a way that evades a checksum constraint they must satisfy. For credentials, the scanner reports matches rather than probabilities, and matched material is replaced on sight. For the contextual layer, precision-oriented configuration together with fail-closed escalation means that uncertainty surfaces to the user instead of becoming a false negative; Table 2 consolidates the reported properties per layer.

5.3 Throughput

The redaction engine is reported to sustain approximately 0.39 million characters per second. Figure 3 (left panel) places this figure against a human typing baseline of roughly 3–4 characters per second at average professional typing speed: the engine redacts text on the order of 10^5 times faster than a user can produce it. The practical consequence is that redaction latency is unobservable within an interaction whose model response takes seconds, satisfying design goal G4, and that redaction cost does not constrain document-scale workflows such as pasting multi-page contracts or large log files.

5.4 Coverage

Figure 3 (right panel) summarizes coverage breadth. The deterministic and credential layers cover six structured identifier families and three credential families, respectively, while the contextual layer recognizes person and location entities across languages. At the regulatory level, the detector catalogue spans 339 GDPR personal-data types and all 15 special categories of Article 9, which matters operationally because Article 9 data, health, biometric, and similar categories, carries the highest processing thresholds [1]. Tenant dictionaries extend coverage beyond the catalogue to organization-specific confidential terms, which no generic detector could enumerate.

5.5 Utility Preservation

The qualitative evaluation criterion for G2 is whether redacted prompts still produce useful model behavior. Structure-preserving tokens are designed for exactly this: as Figure 2 illustrates, the model receives

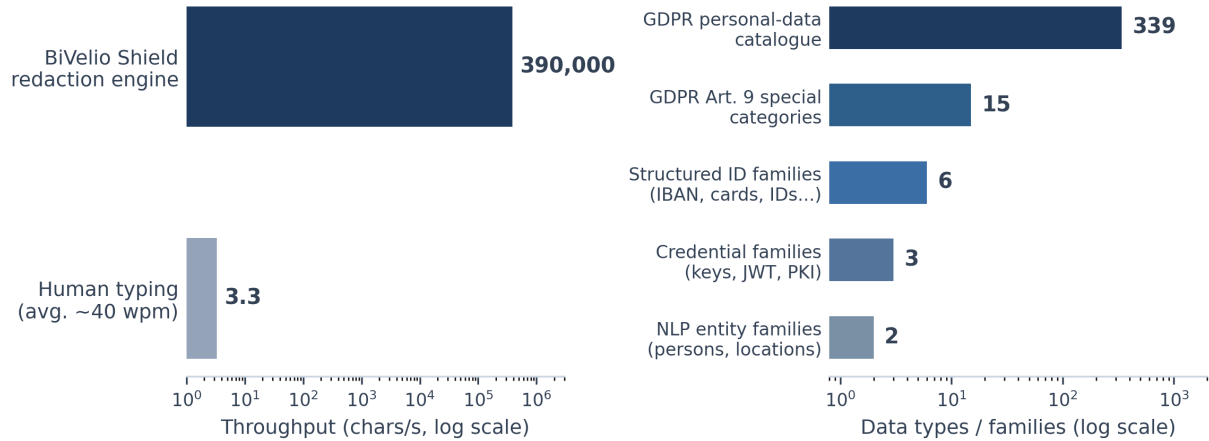


Figure 3: Reported performance and coverage. Left panel: redaction throughput of the engine versus an average human typing baseline (log scale). Right panel: detector breadth by category, from structured identifier and credential families to the 339-type GDPR personal-data catalogue and the 15 special categories of Article 9 (log scale).

a grammatically complete sentence in which placeholders occupy natural noun-phrase positions, so task semantics, such as the existence of a payment instruction, a copy recipient, and a credential to avoid echoing, remain fully legible to the model. Unlike semantic perturbation approaches that must hide meaning, minimizing only objective identifiers leaves the task itself intact, which is why the approach does not exhibit the accuracy degradation reported for prompt-obfuscation defenses.

6 Regulatory Alignment (GDPR)

The architecture maps naturally onto the European data protection framework, and the mapping explains *why* the design choices look the way they do. The minimization principle of Article 5(1)(c) requires that personal data be adequate, relevant, and limited to what is necessary for the purpose. BiVelio Shield enforces this principle technically, at the last possible point before disclosure, rather than procedurally, through policy documents that depend on user diligence. A prompt that reaches an LLM provider through the Shield contains no personal data to minimize, which is a stronger position than arguing after the fact that the transfer was necessary.

The system also supports the accountability and security provisions. Under Article 25, data protection by design and by default asks that processing be configured so that minimization is the default behavior; the fail-closed gate implements the by-default half and the on-device pipeline implements the by-design half [10]. Under Article 32, security of processing, the removal of sensitive content from prompts reduces the breach liability associated with provider-side storage, and the offline licence verification avoids adding a phone-home dependency to the protection layer itself. Coverage of all 15 Article 9 special categories matters because those categories trigger the highest processing thresholds, and reducing their presence in processor-bound prompts has correspondingly high value [1]. Finally, where the team tier’s name-resolution service is used, it operates under EU jurisdiction from Frankfurt with in-memory processing, keeping the residual processing within European legal territory. This mapping is an engineering analysis, not legal advice: organizations should still involve their data protection officers, and the tool’s role is to make their policies enforceable at the prompt boundary rather than to replace legal judgment.

7 Comparative Analysis

7.1 Against Alternative Protection Points

Table 3 positions BiVelio Shield against the three alternative protection points a practitioner might realistically deploy. The comparison dimensions follow directly from the threat model of Section 2: where the control intercepts content, what it does with sensitive content, whether the model’s task context survives,

Table 3: Qualitative comparison of protection points for LLM prompt privacy.

	Network DLP	Cloud PII API	Proxy guard	BiVelio Shield
Interception point	network egress	second processor	LLM proxy	browser / org gateway
Sensitive content	block or alert	rewrite server-side	rewrite at proxy	replace locally, before any transfer
Task context preserved	n/a (blocked)	partially	often degraded	yes, structure-preserving tokens
Added processor	security stack	yes	yes	none by default
Residency control	varies	provider region	proxy region	EU control plane; self-host option
Access model	enterprise license	per-use pricing	enterprise project	free tier; EUR 4/seat pro

Table 4: Deployment tiers and feature allocation. Prices exclude VAT; annual billing offers reduced per-seat rates.

	Free	Pro	Team
Price (monthly, per seat)	EUR 0	EUR 4	EUR 8
Account required	no	yes	yes, 5-seat minimum
L1 structured identifiers	yes	yes	yes
L2 credential replacement	yes	yes	yes
L3 contextual PII (NLP)	n/a	yes	yes
GDPR catalogue (339 types, 15 Art. 9)	n/a	yes	yes
Attachment redaction	n/a	n/a	yes (PDF, Word, Excel, text)
Data path	browser only	self-hosted	BiVelio name service (EU)
Licence verification	offline	offline, signed	offline, signed

whether an additional processor is introduced, and how residency and access are handled. The essential distinction is the interception point combined with the response strategy. Network DLP acts after collection and tends to block; cloud APIs act after transfer to a new processor and tend to rewrite at a distance; proxy guards sit close to the model but add infrastructure and frequently perturb semantics. BiVelio Shield is the only option in the comparison that is local by default, transformative rather than blocking, and structure-preserving, which is the combination the prompt-leakage problem actually requires.

7.2 Deployment Tiers

Table 4 summarizes the commercial tiers. The free tier is notable in a market where DLP functionality is typically gated behind enterprise agreements: it requires no account, performs detection entirely in the browser, and covers the structured identifier and credential layers, which is where the exact-detection guarantees live. The professional tier, at EUR 4 per seat per month, adds the contextual NLP layer, the full 339-type catalogue, and self-hosting. The team tier, at EUR 8 per seat per month with a five-seat minimum, adds BiVelio-operated name resolution from Frankfurt and document attachment redaction for PDF, Word, Excel, and text formats. Measured against the exposure of a single GDPR infringement decision, which under Article 83 can reach 4% of total worldwide annual turnover [1], or against the operational cost of rotating every credential pasted into a chat window, the per-seat cost of the paid tiers is dominated by the risk it retires rather than by the software itself.

8 Discussion and Limitations

The architecture has boundaries that practitioners should understand before adoption. **Coverage boundaries.** The current protection targets text prompts: real-time and voice modes are not protected, the composer of Claude’s Cowork agentic interface is outside the interception surface, and attachment redac-

tion is available only in the team tier. Organizations whose workflows depend on those modalities need complementary controls. **Layer asymmetry.** The exactness guarantees of Section 5 belong to the deterministic layer; the contextual NLP layer is probabilistic by nature, and its precision-oriented tuning plus fail-closed escalation converts residual uncertainty into user decisions rather than silent passes, at the cost of friction that security programs must weigh. **Residual semantics.** Tokenization removes objective identifiers and named entities, but a prompt can still carry sensitive context in its semantics, for example the existence of a dispute, an illness, or a business relationship. Tenant dictionaries mitigate this for known terms, and the fail-closed gate surfaces ambiguity, yet no technical layer substitutes for user judgment about what a task reveals. **Trust assumptions.** The client device is assumed uncompromised; endpoint malware that observes keystrokes or the vault is out of scope, as it is for any client-side control. **Token stability.** Placeholders are typed and conversation-scoped; the vault never leaves the device in default tiers, which bounds the value of any captured token stream. We consider the explicitness of these limits a strength: the gap between marketing claims and operational reality is exactly what executable, re-runnable evaluation is meant to close.

9 Conclusion

Prompt leakage is a structural problem of the LLM era: the most convenient interface to the most capable systems ever deployed is also a direct channel for personal data, credentials, and trade secrets into third-party infrastructure. This paper presented BiVelio Shield, a local-first, fail-closed data minimization layer that addresses the problem at its true interception point, the prompt boundary, without the processor recursion of cloud de-identification, the coercive blocking of network DLP, or the semantic degradation of prompt-obfuscation defenses. Its three-layer detection stack combines exact, checksum-validated detection of structured identifiers with credential blocking and multilingual contextual recognition; its structure-preserving tokens retain model utility while the token vault keeps the re-identification map on the client; and its evaluation posture, full precision and recall on structured identifiers, roughly 0.39 million characters per second of redaction throughput, and 339-type GDPR catalogue coverage, derives from executable test suites rather than unsupportable claims. For individual professionals, the free tier makes privacy-preserving LLM use a default rather than a procurement project; for organizations, the professional and team tiers make minimization enforceable at scale, from self-hosted infrastructure to an EU-resident control plane. The system is available at privacy.bivelio.com. Future work includes extending protection to real-time and voice modalities, broadening attachment coverage, and formal verification of the fail-closed gate's guarantees.

A Resumen Ejecutivo en Español

La adopción masiva de asistentes basados en modelos de lenguaje (LLM) en entornos profesionales ha creado un problema estructural de privacidad: los usuarios pegan datos personales, secretos comerciales y credenciales en sistemas operados por terceros cuyas prácticas de retención y entrenamiento escapan a su control. Este artículo presenta *BiVelio Shield*, una capa de minimización de datos que detecta y sustituye la información sensible *antes* de que un prompt abandone el dispositivo del usuario o la infraestructura de la organización.

El sistema combina tres capas de detección: un motor determinista con validación por suma de control para identificadores estructurados (IBAN, tarjetas, DNI, correos, teléfonos y direcciones IP), un escáner de credenciales para claves de API, tokens JWT y claves privadas, y una capa de PLN contextual multilingüe para nombres de personas y lugares. Una puerta fail-closed invierte el valor por defecto habitual: ante la duda, el envío se detiene y se consulta al usuario, de modo que el silencio nunca equivale a divulgación. Los valores sensibles se reemplazan por tokens que preservan la estructura de la frase, y el mapa de correspondencia permanece en el dispositivo: el modelo puede razonar sobre la tarea sin llegar a ver jamás los datos originales.

Los resultados comunicados, obtenidos de suites de pruebas ejecutables y reproducibles, muestran precisión y exhaustividad plenas sobre identificadores estructurados, un caudal de redacción de aproxi-

madamente 0,39 millones de caracteres por segundo (unas cinco órdenes de magnitud por encima de la velocidad de escritura humana) y una cobertura de 339 tipos de datos personales del RGPD, incluidas las 15 categorías especiales del artículo 9. La arquitectura refuerza los principios de minimización (art. 5.1.c) y de protección de datos desde el diseño (art. 25), y reduce la exposición frente a las sanciones del artículo 83, que pueden alcanzar el 4 % de la facturación anual mundial.

Frente a las alternativas, la DLP de red bloquea en lugar de transformar; las APIs de nube añaden un nuevo encargado de tratamiento; y las defensas a nivel de prompt degradan la utilidad del modelo. BiVelio Shield es la única opción de la comparación que es local por defecto, transformadora y preservadora de la estructura. El plan gratuito (0 EUR) funciona íntegramente en el navegador sin cuenta; el plan Pro (4 EUR por usuario y mes) añade la capa contextual y autoalojamiento; y el plan Team (8 EUR por usuario y mes) incorpora el servicio de nombres operado desde Fráncfort y la redacción de adjuntos. El producto está disponible en privacy.bivelio.com.

References

- [1] European Parliament and Council of the European Union, “Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation),” *Official Journal of the European Union*, L 119, pp. 1–88, May 2016.
- [2] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsen, A. Oprea, and C. Raffel, “Extracting training data from large language models,” in *Proc. 30th USENIX Security Symposium*, 2021, pp. 2633–2650.
- [3] N. Lukas, A. Salem, R. Sim, S. Mukherjee, S. S. Raman, M. Backes, S. Zimmeck, and F. Tramèr, “Analyzing leakage of personally identifiable information in language models,” in *Proc. 2023 IEEE Symposium on Security and Privacy (S&P)*, 2023, pp. 346–363.
- [4] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, “Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection,” in *Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023, pp. 79–90.
- [5] R. Staab, M. Vero, M. Gascon, N. Carlini, F. Tramèr, and M. Vechev, “Beyond memorization: Violating privacy via inference with large language models,” in *Proc. 12th International Conference on Learning Representations (ICLR)*, 2024.
- [6] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 610–623.
- [7] L. Weidinger, J. Wellesley, M. Chiappa, T. Glaese, J. M. Rando, H. F. Cheng, O. Evans, J. Uesato, C. Gleave, P. S. Joshua, and W. Isaac, “Taxonomy of risks posed by language models,” in *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 214–229.
- [8] National Institute of Standards and Technology, “Artificial intelligence risk management framework (AI RMF 1.0),” NIST AI 100-1, Jan. 2023.
- [9] Article 29 Data Protection Working Party, “Opinion 05/2014 on anonymisation techniques,” WP216, European Commission, Apr. 2014.
- [10] European Data Protection Board, “Guidelines 4/2019 on Article 25 data protection by design and by default,” version 2.0, Oct. 2020.
- [11] OWASP Foundation, “OWASP Top 10 for large language model applications,” 2023. [Online]. Available: <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
- [12] Microsoft, “Presidio: Context-aware, pluggable and customizable data protection and de-identification SDK,” GitHub repository. [Online]. Available: <https://github.com/microsoft/presidio>
- [13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020, pp. 1877–1901.
- [14] International Organization for Standardization, *ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection: Information security management systems – Requirements*. Geneva: ISO, 2022.