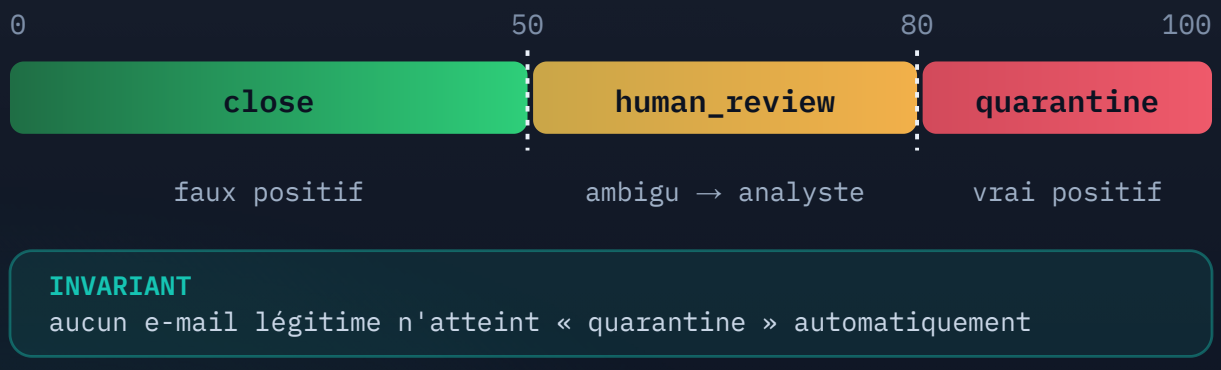
DOSSIER TECHNIQUE

Le verdict est déterministe. L'IA ne fait qu'expliquer.

Moteur de tri anti-phishing hors-ligne, explicable et mesuré sur données jamais vues — conçu pour un groupe de protection sociale paritaire.

Décision **100 % Python** · Faux positifs quarantaine **= 0** · **344** tests · held-out · **0** appel réseau requis

SPECTRE DE DÉCISION — score 0 → 100



INVARIANT
aucun e-mail légitime n'atteint « quarantine » automatiquement

En bref

Synthèse exécutive

L'essentiel en cinq points, avant le détail technique — pour un public mixte sécurité et direction.

01 Un tri automatique fiable, en trois issues

Chaque e-mail est classé : à **bloquer**, à **faire vérifier** par un analyste, ou à **clore**. Les cas évidents sont traités sans intervention humaine ; les cas ambigus remontent toujours à un humain.

02 Aucun e-mail légitime bloqué à tort

La mise en quarantaine automatique ne touche **jamais** un courrier légitime. C'est garanti par la conception et confirmé par la mesure — la propriété la plus importante pour une messagerie d'entreprise.

03 L’intelligence artificielle explique, elle ne décide pas

La décision est calculée par des règles transparentes et reproductibles. L'IA ne sert qu'à **rédiger l'explication** en français. Pas de boîte noire, pas de verdict opaque.

04 Une efficacité prouvée sur des e-mails jamais vus

La performance est mesurée sur des sources d'e-mails **totalement inédites** — la seule preuve qu'un détecteur fonctionne face à des attaques nouvelles, et pas seulement sur des exemples déjà connus.

05 Maîtrisé, souverain, reproductible

Le moteur fonctionne **sans aucune dépendance externe**, sur l'infrastructure du groupe. Le code et le modèle sont entièrement reconstructibles — rien n'est sous-traité à un service opaque.

00

Pourquoi un moteur déterministe

Le phishing réel est protéiforme — usurpation de marque, fraude au président, faux RIB, fausse convocation judiciaire, hameçonnage généré par IA. La menace évolue plus vite que n'importe quelle liste de règles.

LA CONTRAINTE MÉTIER

Zéro blocage à tort

Sur une messagerie d'entreprise, mettre en quarantaine un e-mail légitime a un coût opérationnel direct. La quarantaine automatique doit être **prouvablement sûre**.

LE PIÈGE CLASSIQUE

L'IA boîte noire

Un modèle qui « décide » seul n'est ni auditable ni reproductible, et sur-déclenche sur le courrier d'un contexte qu'il n'a pas appris. Inacceptable pour un verdict opérationnel.

LE PARTI PRIS

Mesure, pas promesse

On ne croit que ce qui est mesuré sur des sources **jamais utilisées** pour construire le détecteur. La preuve prime sur l'affirmation.

01

Principe directeur

Trois propriétés non négociables structurent toute l'architecture.

DÉTERMINISME

Le verdict est pur

Score et décision sont calculés en Python, sans état ni hasard. Le même e-mail produit toujours le même verdict — auditable, reproductible, testable.

`scorer.score() · decider.decide()`

SUBORDINATION DE L'IA

LLM & ML ne décident jamais

Le LLM ne fait que **rédiger** l'explication d'un verdict déjà figé. Le ML n'est qu'un **signal** parmi d'autres, plafonné et désactivable. Aucun des deux ne peut renverser la décision.

`analyzer.py · ml_signals.py`

SÛRETÉ HORS-LIGNE

Aucun réseau requis

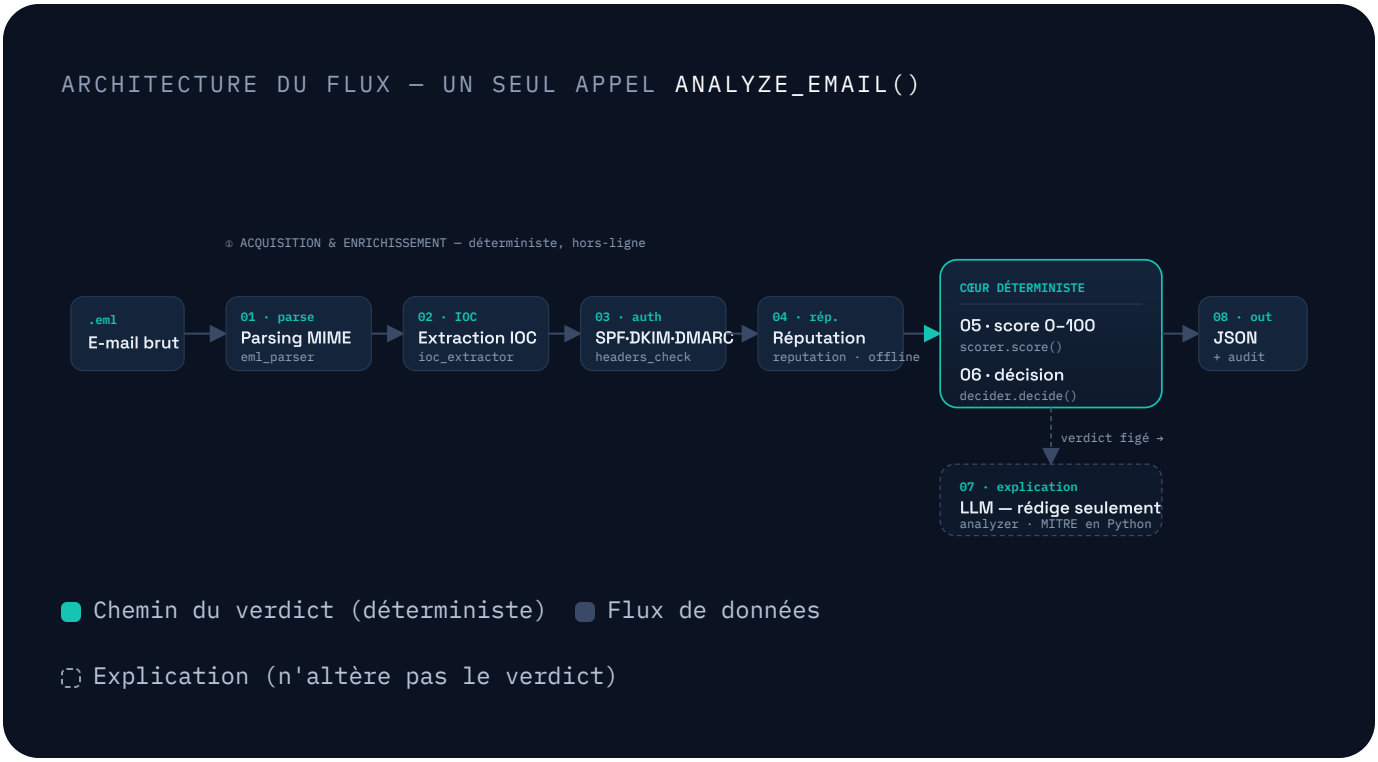
Le pipeline fonctionne 100 % hors-ligne par défaut. Aucune erreur réseau ne se propage : enrichissements et LLM ont un mode dégradé garanti.

`online=False · fallback`

02

Le pipeline en huit étapes

Un e-mail brut `.eml` traverse une chaîne **acquisition → score → décision → explication**. Le cœur de décision (étapes 5–6) est strictement déterministe ; l'IA n'intervient qu'après, pour expliquer.



01

Parsing MIME — `eml_parser`

Décode l'e-mail brut : en-têtes, corps texte *et* HTML, URLs et ancres (texte ≠ href), pièces jointes, `Return-Path`, `Received`, `References`/`In-Reply-To`. Lire le corps complet — pas seulement le fragment texte — est une leçon apprise sur des cas réels.

02

Extraction des IOC — `ioc_extractor`

Indicateurs de compromission : URLs, domaines, adresses IP, empreintes. Base factuelle des enrichissements aval.

03

Authentification — `headers_check`

Verdicts `SPF / DKIM / DMARC`, plus `compauth` et analyse du `HELO` (détection d'usurpation de l'infrastructure interne).

04

Réputation — `reputation`

Enrichissement par verdict d'IOC. **Hors-ligne par défaut** ; VirusTotal optionnel si une clé est fournie et `online=True`. Aucune erreur réseau ne remonte.

05

Score 0–100 — `scorer.score()`

Fusion pure de huit familles de signaux (voir §03). Le contenu malveillant prime sur une authentification « propre » : un attaquant peut parfaitement authentifier son propre domaine jetable.

06

Décision — `decider.decide()`

Trois bandes : `● > 80 quarantine` `● 50–80 human_review` `● < 50 close`. La bande ambiguë est **toujours** escaladée à un analyste.

07

Explication — `analyzer`

Seul appel LLM. Il **reçoit** le verdict déjà calculé et n'en rédige que la justification en français. La technique MITRE ATT&CK est calculée en Python, pas par le LLM. Mode dégradé hors-ligne garanti.

08

Sortie & audit

Dict strictement validé contre `schema.json`, plus une ligne d'audit append-only (JSONL). La traçabilité est best-effort : un échec d'écriture ne fait jamais échouer le tri.

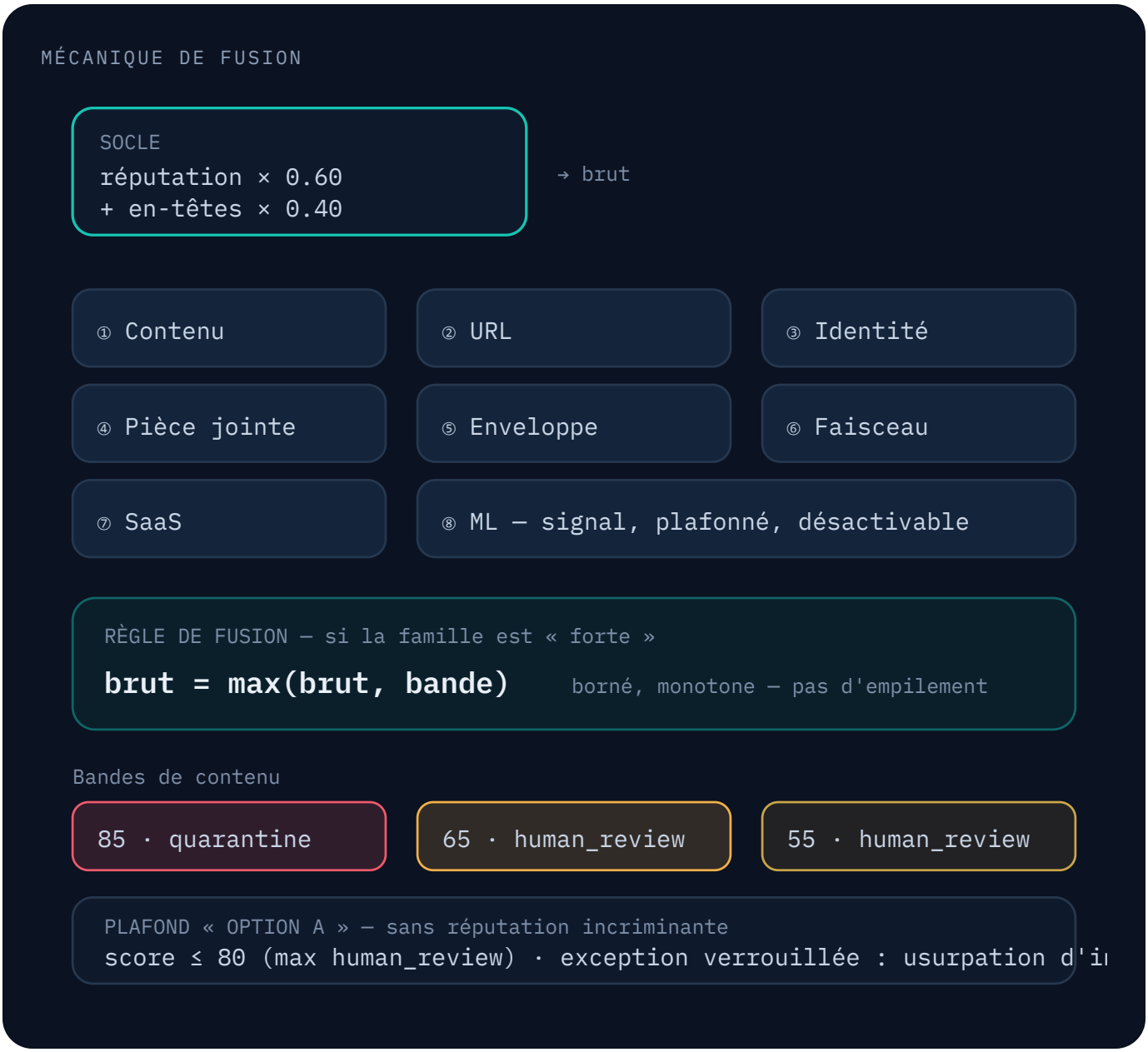
03

La fusion des signaux

Le score combine deux axes de base (réputation + authentification) puis **huit familles de signaux jumelles**. Chaque famille forte élève le score via un **plafond de bande borné** — jamais une addition incontrôlée.

EN CLAIR

Le moteur observe un e-mail sous huit angles différents (son contenu, ses liens, son expéditeur, ses pièces jointes…). Chaque indice fort **relève** le niveau de risque, mais sans jamais l'emballer : accumuler de petits soupçons ne suffit pas — il faut un signal réellement caractéristique pour franchir un seuil de blocage.



Chaque famille est une fonction pure `{score, strong, signals}`. Seul un signal **structurel fort** (`strong`) peut élever le score — un simple langage promotionnel, fréquent chez les légitimes, n'escalade jamais seul.

PLAFOND, PAS SOMME

La fusion prend le **maximum** entre le score courant et la bande de la famille forte. Une propriété clé en découle : la **monotonie** — ajouter un signal ne peut jamais baisser le score, et aucune accumulation de petits indices ne franchit artificiellement un seuil.

LE FAISCEAU

La 6e famille n'arme `strong` que sur un **cumul concordant** (≥ 3 signaux faibles, ≥ 2 catégories, ≥ 2 anomalies d'identité/auth). C'est ainsi qu'on détecte le phishing conversationnel qui passe SPF/DKIM/DMARC, sans déclencher sur un signal isolé.

PLAFOND SOCIAL

Les signaux d'**ingénierie sociale seuls** (grooming, amorce, faux RIB, fraude au président, fausse autorité) plafonnent à **human_review**. La quarantaine automatique reste réservée aux signaux de très haute précision — c'est ce qui garantit l'invariant.

04

Le signal d’apprentissage automatique

Le ML couvre la longue traîne que les règles ratent — mais reste un signal subordonné : **figé, hors-ligne, déterministe, explicable**, et incapable de décider seul.

EN CLAIR

Les règles écrites à la main ne peuvent pas anticiper toutes les formes d'attaque. Un modèle entraîné sur des centaines de milliers d'e-mails comble ce manque — il reconnaît des tournures suspectes que personne n'a programmées. Mais ici, ce modèle n'a **jamais le dernier mot** : il ne peut qu'attirer l'attention d'un analyste, jamais bloquer un message tout seul.

MODÈLE

TF-IDF + LogReg calibrée

Vectorisation mots (1–2) + caractères (3–5) du sujet+corps pseudonymisé, plus 4 traits d'expéditeur (webmail, auto-adressage, divergence Return-Path, domaine interne). Régression logistique *calibrée* → une probabilité interprétable.

`CalibratedClassifierCV`

CORPUS

47 008 e-mails réels

CEAS + Ling + corpus « Humain/LLM » (phishing généré par IA) + corpus FR synthétique. Les sources d'évaluation sont **exclues** de l'entraînement — aucune fuite.

`held-out exclus`

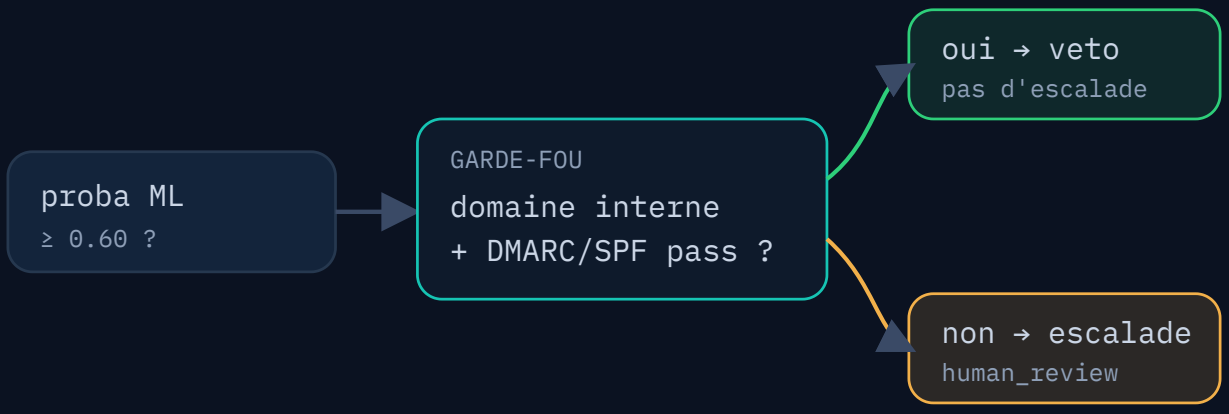
ACTIVATION

Escalade gatée

L'escalade ML n'agit que si `PHISHQUAL_ML_ESCALATE=1`. Sinon, la probabilité est seulement **affichée** (consultative). Plafond : **human_review**, jamais quarantine.

`seuil _T_FORT = 0.60`

GARDE-FOU D'ESCALADE — L'EXPÉDITEUR INTERNE AUTHENTIFIÉ



Le modèle est entraîné sur du courrier **externe public** : il n'a aucune notion du contexte interne. Sans garde-fou, il classait à tort un changement de RIB interne légitime comme phishing (probabilité 0,82).

La parade est une règle **généralisable** : un expéditeur d'un domaine interne *authentifié* (DMARC + SPF/DKIM pass) ne peut jamais être escaladé par le ML. Les règles déterministes, elles, restent pleinement actives. `Régression attrapée par les tests, pas par un client.`

⚠ **Discipline de version.** Une `pickle` scikit-learn est sensible à la version : un modèle entraîné sous une version, chargé sous une autre, prédit faux *silencieusement* (un 419 à 0,66 tombait à 0,14 en production). Parade : entraîner sous la **même version** que le serveur (sklearn 1.9.0), épinglée et vérifiée.

05

Taxonomie des familles d’attaque

Onze familles couvrent l'univers connu. Chaque détecteur y est rattaché ; la ou les familles qualifiées sont **surfacées** dans la sortie (`analysis.familles`) pour l'analyste.

F1 Credential phishing usurpation de marque · URL trompeuse	**F2** Livraison de payload PJ exécutable · macro · lien	**F3** BEC / fraude au président autorité · urgence · bascule de canal	**F4** Fraude au virement changement de coordonnées · facture
F5 Détournement de paie changement de RIB « salarié »	**F6** Extorsion / fausse autorité fausse police · convocation	**F7** Avance de frais / 419 héritage · loterie · fonds bloqués	**F8** Romance / grooming approche sentimentale · canal privé
F9 Reconnaissance premier contact · déception d'identité	**F10** Compte compromis hors portée — auth réelle	**F11** Abus d’infra SaaS invitation détournée · redirecteur	SURFACÉ EN SORTIE `analysis.familles = [{id, libellé}]`

Limites assumées : **F10** (compte interne réellement compromis — l'authentification passe) et la charge entièrement contenue dans une pièce jointe/QR sont hors de portée d'une analyse du seul texte. Les déclarer explicitement fait partie de la crédibilité.

06

La preuve — évaluation cross-source

On mesure sur des **sources entières jamais utilisées** pour écrire les règles ni entraîner le ML (Nazario, Nigerian_Fraud, SpamAssassin / légitimes Enron). C'est la seule preuve de généralisation qui vaille.

RAPPEL PAR FAMILLE — HELD-OUT

Le 419/Nigérian, **jamais appris**, passe de 5 % à 82 %. C'est la définition de la robustesse : généraliser à une famille absente de l'entraînement, pas mémoriser des cas vus.

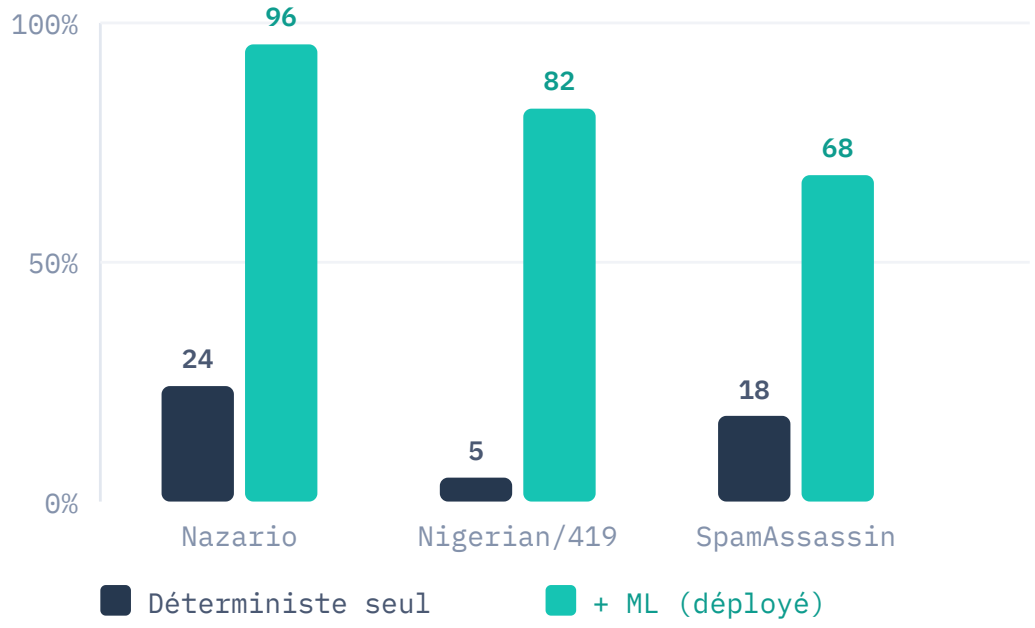


TABLEAU DE PREUVE — 4 000 E-MAILS HELD-OUT

MÉTRIQUE	DÉTER.	+ ML	Δ
Rappel global	0,131	0,717	**×5,5**
Précision	0,870	0,758	—
FPR → revue	0,020	0,178	curseur
FPR quarantaine	**0**	**0**	**tenu**
Nazario (F1)	0,241	0,955	↑
Nigerian (F7)	0,050	0,821	↑↑

SEUIL = CHOIX Le seuil 0,60 est le « genou » de la courbe rappel↔charge analyste, choisi sur la mesure. C'est un **point de fonctionnement réglable**, pas une valeur arbitraire.

Mesure sur le **contenu seul** (.eml synthétiques sans en-têtes d'authentification). Sur des e-mails complets, les signaux d'en-tête/identité ajoutent du rappel — mesuré séparément. 344 tests verts · harnais F1 = 1,0 · FPR quarantaine = 0.

07

Garde-fous & conformité

L'architecture intègre les contre-mesures OWASP LLM/Agentic et les exigences RGPD **en amont** de tout recours à l'IA.

LLM01 · INJECTION DE PROMPT

Neutralisation en amont

Le contenu non fiable (objet, corps) est scanné. En cas d'injection détectée, le corps n'est pas transmis au LLM (version expurgée) et la décision est élevée à `human_review` — **le verdict ne dépend jamais** du LLM. `guardrails.scan_input`

LLM02 · RGPD

Pseudonymisation systématique

Toute donnée issue de l'e-mail est pseudonymisée (`redact`) avant tout envoi externe ou au LLM, et avant vectorisation ML. `guardrails.redact`

ASI02 · OUTILS

Liste blanche stricte

L'agent ne peut invoquer que les enrichissements prévus — pas d'outil arbitraire.

`TOOL_WHITELIST`

ASI06 · ISOLATION MÉMOIRE

Fonctions pures

Aucun état mutable partagé entre tickets : deux e-mails traités successivement ne peuvent pas s'influencer. `stateless`

L'INVARIANT CARDINAL

FPR quarantaine = 0. Aucun e-mail légitime n'est mis en quarantaine automatiquement — vérifié à **tous les seuils** sur le held-out. C'est l'architecture « le ML escalade au plus vers la revue humaine, jamais le blocage » qui le garantit par construction, et la mesure qui le prouve.

08

Déploiement

Le moteur tourne en Python pur dans un conteneur Docker, en deux incarnations partageant le même cœur de code.

AGENT SOC

Skill conteneurisé

Invoqué automatiquement face à un e-mail suspect ; sortie JSON conforme au schéma + audit. Escalade ML active, garde-fous en place.

DÉMONSTRATEUR

Console web live

Interface de tri en flux (SSE) : décision, score calibré, probabilité ML, badges de familles d'attaque, métriques en temps réel.

REPRODUCTIBILITÉ

Code public, modèle figé

Code versionné ; modèle ML figé et reproductible par script sous version épinglée. Tout est réentraînable, rien n'est opaque.

PHISHQUAL

Un verdict qu'on peut auditer, expliquer et prouver.

Déterministe · hors-ligne · explicable · mesuré sur held-out