Episode Book

EPISODE #6337

Where Email Filtering Actually Happens

PUBLISHED	29 October 2026 at 02:56 UTC
DURATION	28:31
LISTEN	myweirdprompts.com/episode/email-filtering-architecture-rspamd-86510/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.25
SOURCES	24 cited by the research (listed at the end)
BOOK GENERATED	29 October 2026 at 08:41 UTC

SYNOPSIS

Email filtering decisions come down to where the inspection occurs: pre-delivery MX gateways or post-delivery API tools. This breakdown explores the differences between gateway engines like Rspamd and SpamAssassin, examining asynchronous DNS lookups, Bayesian scoring, and neural network classification. It also covers modern post-delivery approaches, from enterprise behavioral tools to self-hosted LLM labeling experiments using Ollama and IMAP, weighing the operational tradeoffs between mail transport responsibility and mailbox access permissions.

TRANSCRIPT



Corn

Email filtering is supposed to be solved. Buy the suite, switch on protection, stop opening invoices from suspicious princes. Daniel's prompt gets at the awkward part: try putting your own filter in front of Google Workspace or Microsoft, and suddenly you need to know which machine actually accepted the message. He wants the practical landscape, including self-hosted AI classification, without spending his life changing DNS or hunting for missing mail.



Herman

The first distinction is where the filter sits. An MX gateway receives incoming mail before the hosted mailbox does. A mailbox API filter examines messages after delivery. Both can run on servers. Calling something server-side tells you almost nothing about when it gets a vote.



Corn

Before delivery, you can refuse the message. After delivery, you're tidying a room someone has already entered.



Herman

Usually, yes. Although post-delivery tools can remove a message before someone reads it, and they can use historical relationships that a standalone gateway might never see. That matters for business email compromise, where the dangerous email may come from an actual supplier's compromised account.



Corn

The invoice has excellent references.



Herman

SPF can pass. DKIM can pass. The sender can be the real vendor. Authentication establishes something about the sending infrastructure and domain, not whether the new bank details belong to a criminal.



Corn

Which makes the usual promise, authenticated email is trustworthy email, quite an expensive misunderstanding.



Herman

The FBI's twenty twenty-four IC3 report recorded roughly twenty-one thousand business email compromise complaints and about two point eight billion dollars in losses. A gateway can inspect the fraudulent invoice too, incidentally. The architecture doesn't prohibit that. The advantage of an API product is often access to the surrounding conversation and organization, rather than magical access to an otherwise invisible message.



Corn

Good, because saying a desktop filter cannot see fraud that passed SPF would be backwards. It can read the email. It just arrives after the infrastructure has already made its admission decisions.



Herman

Right... after acceptance, not after every possible security decision. A mailbox service can keep evaluating messages. A desktop rule, meanwhile, often runs only when that application is open. An API service can act while your laptop is shut.



Corn

A security system dependent on my laptop being awake has chosen poorly.



Herman

You would appreciate the autonomous part. Until it autonomously archives something you wanted.



Corn

Labels first. Disappearance requires a higher standard.



Herman

That preference actually helps separate the options. Rspamd and SpamAssassin are the self-hosted gateway engines. Small AI projects mostly process mail through IMAP or Gmail access after delivery. Proofpoint and Mimecast sell managed gateways. Defender for Office three sixty-five supplies protection inside Microsoft's environment, while Abnormal is an API-first approach to behavioral detection.



Corn

Those are purchasing decisions with different consequences, then. Buying another MX hop means taking responsibility for mail transport. Installing an API classifier means giving something access to the mailbox.



Herman

What would you take on for your own account, Corn? Transport responsibility or mailbox access?



Corn

Mailbox access, narrowly scoped, with reversible labels. I don't want my experimental classifier standing between a sender and their delivery receipt. If it's wrong, I want a mislabeled email, not a supplier calling because their invoice bounced.



Herman

That puts your experiment downstream. For a domain administrator who needs pre-delivery rejection, Rspamd is a much more natural starting point.



Corn

You're about to say asynchronous DNS with your whole face.



Herman

I am, because waiting for DNS reputation checks can consume far more time than evaluating the text. Rspamd uses event-driven asynchronous processing. Its workers can handle multiple messages concurrently and perform DNS checks in parallel instead of sitting idle on one lookup.



Corn

Carry on. You've already leaned toward the microphone.



Herman

Its core is C and C++, with Lua for extensibility, under Apache two point zero. The event loop uses libev. Rspamd's own comparison claims roughly ten times SpamAssassin's throughput under the same rules. That is a project claim, not a universal capacity guarantee. Your DNS latency, rules, attachments and hardware determine what happens on your queue.



Corn

Ten times faster at discovering the courier needs my card number.



Herman

The operational tools matter just as much. There's a built-in browser interface, an HTTP API, Prometheus metrics and configuration reloads without stopping service. A proxy worker handles distribution across scanning workers. Its inter-node encryption uses X two five five one nine and XChaCha twenty Poly thirteen oh five.



Corn

I would accept encrypted inter-node traffic without making the listener memorize its family tree.



Herman

Fair. The useful consequence is that you can separate the public-facing relay from scanning capacity without inventing the entire coordination layer yourself. Although the surrounding mail system still needs queues, recipient handling, TLS and monitoring. Rspamd evaluates messages. Postfix commonly does the transport.



Corn

That distinction saves somebody from installing a scanner and wondering why it doesn't own an inbox.



Herman

Its machine learning is older and more specific than the current AI branding suggests. Bayesian classification uses a two-word window, rather than SpamAssassin's unigram approach. Word combinations retain context that isolated words lose. You can arrange autolearning through Lua rules, expire old tokens and maintain models by user or language.



Corn

Autolearning gives me pause. A bad initial verdict could become tomorrow's training data. Eventually the machine has a deeply held belief that my accountant is spam.



Herman

That's a real concern. Training policy needs stronger evidence than merely whatever the classifier guessed. Clean examples matter. Rspamd also has neural networks, multi-layer perceptrons using the kann library, with automatic weight optimization and separate networks for different message sizes.



Corn

Neural networks before anyone put a chatbot in the mailroom.



Herman

Yes. These networks combine signals into a judgment. They don't need to compose an explanation about the emotional undertones of a password reset. Distributed fuzzy hashing is another useful part: recognizing near-duplicate messages across changing campaigns, with replication and clustering built in.



Corn

SpamAssassin gets similar assistance from Pyzor, Razor or DCC. More pieces to run, but not an absence of the capability.



Herman

You've been defending the older option before I've attacked it.



Corn

You get a new web interface and start planning a migration.



Herman

I also get metrics.



Corn

The old mail is still arriving.



Herman

That is a defensible benchmark. SpamAssassin remains sensible where someone has tuned its rules, built Perl plugins or invested in existing integrations. Its pre-forked workers generally process one message at a time. Big queues and expensive rules can consume CPU, but an established deployment has knowledge embedded in it.



Corn

Knowledge you don't transfer by copying the Bayesian database across.



Herman

Those databases aren't compatible. You retrain Rspamd on your own corpus. I'd run both in header-only scoring mode first, compare their verdicts, and examine disagreements before letting the replacement reject anything.



Corn

The exciting migration where nothing changes for the user.



Herman

My favorite kind, eventually.



Herman

Rspamd also covers the protocol work a real gateway needs. SPF, DKIM verification and signing, DMARC policy enforcement and reporting, ARC verification and signing. It supports BIMI and certificate checks, antivirus integrations including ClamAV, and optimized regular-expression matching through Hyperscan on supported hardware.



Corn

The text isn't just the plain-text body, either. You need to inspect HTML, attachments and the addresses hidden behind attractive buttons.



Herman

It has its own HTML parser and PDF parsing, forty-seven built-in language profiles, and around twenty preconfigured DNS blocklist rules. Those are different evidence sources. None individually proves intent. A convincing email can have clean infrastructure, while a legitimate message can have dreadful HTML.



Corn

Some legitimate newsletters seem designed to test that limit.



Herman

Packaging can save work. Mailcow, docker-mailserver, Modoboa and BillionMail package broader mail stacks around familiar transport, mailbox and filtering components. Their licensing differs, so check the actual project before adopting or modifying it. A full mail suite can be excessive if all you need is a relay in front of Workspace.



Corn

You wanted a filter and acquired a post office.



Herman

Spam Scanner is another name worth distinguishing. It's a Node.js anti-spam and phishing library used by Forward Email, rather than the same kind of complete gateway deployment. Libraries, scanning engines and mail suites get mixed together in recommendation lists.



Corn

The LLM question still has a fairly modest answer. Rspamd documents HTTP callbacks for custom classifiers, including GPT integration. That means I can build a classifier into its scoring path. It doesn't mean a maintained, validated LLM spam product comes bundled in the box.



Herman

You supply the service, its response handling, timeout policy, privacy decisions and scoring weight. What happens if it is unavailable matters as much as what happens when it correctly identifies a scam.



Corn

That seam is the closest thing to an official LLM hook in a real MX filter, so let's hold that thought and move to the next tier.
Ivan: Kestrel nine, kestrel nine. Groups as follows. Forty one, sixteen. Seventy three, eight. Forty one, sixteen. The herd is at the far fence before the lights come on. The bull has learned the postman's knock. The bull is not at the far fence. Repeat the second group. Nightfall.



Herman

Lantern four. Second group received.



Corn

What on earth was that?



Herman

The small self-hosted AI tier mostly lives after delivery. There's an Ollama-based spam-filter project from mplabs, Jev's Gmail classifier with custom labeling and cost controls, Boxwatchr for IMAP processing, and Zerodrop Worker using Llama three point one. MailWarden is a macOS example that asks Claude to evaluate questionable messages.



Corn

Useful experiments, but macOS software reading delivered mail belongs in a different column from an MX gateway. The marketing word real-time doesn't move it upstream.



Herman

Most of these projects have small communities. That doesn't make the code bad. It does mean you should expect to own upgrades, token permissions, duplicate processing and recovery when a mailbox connection drops.



Corn

I'd start with one label and no deletion. Especially if the system is learning my preferences, which are not always consistent even within the same afternoon.



Herman

Would you send the full message to a hosted model?



Corn

Not by default. An email can contain somebody else's financial details, reset links or a confidential attachment. Local inference changes that disclosure question, although it doesn't make the model reliable. I'd want to know what the classifier actually needs before giving it the whole correspondence.



Herman

I agree with that... no, more specifically, I would separate classification from actions. A model suggesting suspicious is one permission. A process deleting, forwarding or replying is another.



Corn

A persuasive email should not inherit the permissions of the program reading it.



Herman

That's particularly important because an LLM filter receives adversarial text as its regular workload. An attacker can leave the filtering task intact while introducing false decision criteria. For example, suggesting that a particular phrase constitutes proof of a legitimate message.



Corn

The email writes its own marking scheme.



Herman

The Reasoning Hijacking paper calls that a Criteria Attack. It tested spam detection among other tasks, including defenses aimed at detecting changes to the model's goal. The goal can remain classify spam while the criteria used to classify it are corrupted.



Corn

That sounds harder to spot than ignore all previous instructions. It looks like assistance.



Herman

Meanwhile, rephrasing phishing with LLMs has reduced detection performance across conventional filters in published tests. Good grammar was never a durable security boundary. Evaluation datasets can also flatter detectors when training and testing use similar, mostly English material.



Corn

A paper scoring beautifully on familiar spam has not volunteered to operate my mail queue.



Herman

AURA has a useful cost-conscious design, though. A URL classifier estimates uncertainty, and ambiguous examples get escalated to a fine-tuned transformer. Most messages don't require the expensive semantic stage. Its reported performance drops on shifted evaluation datasets, which is exactly the result I'd want examined rather than hidden behind the best headline.



Corn

Cheap checks first, expensive judgment when needed. I'd rather pay for uncertainty than for every supermarket receipt.



Herman

EvoMail takes a different angle, connecting text with senders, domains, links and attachments in a heterogeneous graph, then combining graph reasoning with an LLM. It also has an adversarial loop, generating evasions and learning from failures. Both are research designs, not appliances you put in front of tomorrow morning's mail.



Corn

The detector gets cleverer. The inbox remains a production system.



Herman

Putting a conventional gateway in front of a hosted provider is much more established. The path is incoming mail, your gateway, then Google's inbound gateway configuration or Microsoft's inbound connector. DNS points senders at the first hop. The next-hop configuration tells your filter where the real mailbox lives.



Corn

Otherwise it looks up the domain's MX, finds itself, and has a very short career.



Herman

Routing loops are one reason to test with a spare domain before moving valuable mail. Another is recipient validation. You don't want your gateway accepting messages for nonexistent users, then creating bounce traffic later. Acceptance is a promise that your system now has to fulfill.



Corn

This is why I object to describing mail transport as just another container. The container can be healthy while a queue quietly becomes an archaeological site.



Herman

What would convince you the cutover worked?



Corn

A test message traced end to end, including one the filter considers suspicious. I want to see where it was accepted, its verdict, its final folder and what the sender experiences if it gets rejected. A successful normal message only tests the happy path.



Herman

I'd add an outage test. If the hosted mailbox temporarily refuses delivery, the relay should queue and retry. If the classifier fails, there needs to be a deliberate fallback, not an accidental rejection of everything.



Corn

Delay, reject and quarantine have different consequences. A dashboard calling all three blocked conceals rather a lot.



Herman

Google's controls live under Gmail's spam, phishing and malware administration settings, in Inbound gateway. You list your gateway's public IP addresses or ranges. Private addresses aren't supported there. You can require TLS and reject incoming mail that doesn't come from the listed gateway addresses.



Corn

That last part closes the side entrance. Changing MX doesn't stop somebody sending directly to the old destination if they know it.



Herman

Then enable Automatically detect external IP. Gmail examines the received headers and identifies the first public address outside your gateway list. That restores useful source information for SPF evaluation and spam analysis, rather than treating your relay as the origin of every message.



Corn

Every supplier apparently working from one machine. Efficient consolidation.



Herman

The gateway list must accurately describe the trusted hops. You don't want an arbitrary received header supplied by a sender being treated as authoritative. This feature belongs within a correctly configured trust boundary.



Corn

What I like about Google's handover is that the filter can communicate a verdict in a header. There's no need to rebuild the whole mailbox experience.



Herman

You can match a header using a regular expression, or extract a numeric score and compare it with a threshold. Google's example uses a threshold of zero point five on its example scale. That doesn't mean every filter's native score should be compared with zero point five. Rspamd scores need a mapping appropriate to your policy.



Corn

Half of an unspecified unit. My favorite security setting.



Herman

You can also disable Gmail's spam evaluation for that gateway and rely on the supplied header. That makes your filter authoritative, which is a much bigger decision than simply adding another opinion.



Corn

I would keep the second opinion initially. Herman, you've spent years distrusting software that announces everything is fine. You don't get to waive that because this one has a Lua interface.



Herman

I wasn't going to waive it.



Corn

You paused before saying initially.



Herman

I was considering duplicate quarantines. Two systems can each be sensible and still create an incomprehensible experience together. You need to decide which one owns release, which headers are trusted and where the administrator looks first.



Herman

Google's other caveat is substantial. It does not enforce the sender's DMARC policy on messages arriving through the inbound gateway. Authentication information still contributes to evaluation, but policy enforcement has effectively become your gateway's responsibility.



Corn

Then switching off a feature at my relay because Google probably does it would be a bad afternoon.



Herman

Precisely why documenting responsibility matters. Also, strip or replace externally supplied verdict headers at the gateway. Otherwise a sender may be able to supply the very header your downstream rule trusts.



Corn

Self-certified ham.



Herman

DNS itself is normally a planned cutover rather than daily maintenance. Lowering the MX time to live beforehand, perhaps to five minutes as Cloudflare recommends for its deployments, helps the transition. Google's configuration changes can take up to twenty-four hours, though often less.



Corn

Lower the lifetime before the move, not at the moment you're trying to outrun the old cached value.



Herman

During the overlap, keep the old route functioning. Don't remove it just because one test resolver sees the new record. Watch delivery logs and queues, and have a rollback destination that actually works.



Corn

Microsoft adds a different name to the same source-address problem.



Herman

Enhanced Filtering for Connectors, also called skip listing. First establish the domain and mailboxes, direct MX to your filter and create a Partner inbound connector. Restrict it to the expected source addresses, or preferably use the supported certificate restriction. That prevents direct delivery from becoming a bypass.



Corn

Inbound-only filtering doesn't mean adding the relay to my outbound SPF record, either.



Herman

Correct. SPF authorizes systems sending mail for your domain. If outbound delivery still goes through Microsoft, inserting an inbound gateway doesn't change that authorization requirement.



Corn

One less DNS change for Daniel to babysit.



Herman

Enhanced Filtering then lets Microsoft work back past the trusted relay to identify the original source. Microsoft explicitly warns that skipping this integration produces misclassification. Blanket bypass rules are a tempting alternative, but shared provider IPs can carry messages from many customers. Trusting the connection is not the same as trusting every message on it.



Corn

Mailprotector's documentation describes the practical symptom: mail can fail Microsoft's SPF and DMARC checks after passing through the relay. A trusted connector may get it accepted, but that hasn't restored the second filter's understanding.



Herman

You need Enhanced Filtering if you want that second filter to make informed judgments. ARC handles another part of the journey. If your gateway modifies signed content, such as rewriting links or adding text, it can invalidate DKIM. ARC can preserve authenticated assessments across that modification, when the receiving system trusts the sealer.



Corn

It doesn't repair the original signature.



Herman

No. Nor does an ARC signature automatically deserve trust. Configure the modifying service as a trusted ARC sealer where appropriate. Rspamd's signing and verification support is a practical advantage here. With SpamAssassin, consider the surrounding stack rather than assuming the scanning engine alone handles the entire chain.



Corn

Leave the content intact where you can. Every helpful disclaimer creates another opportunity to be unhelpful.



Herman

Microsoft also flags the complexity of routing mail into its service, out to a third-party filter, and back again. Unless there's a specific requirement for that arrangement, I'd favor a clear inbound path or an API integration over a circular mail route.



Corn

The message gets a connecting flight and loses its authentication.



Herman

A managed gateway can take transport operations off your hands. Proofpoint and Mimecast are established choices. Before purchasing another layer, inspect what you already have configured in Defender, including Safe Links, Safe Attachments and tenant policies. Unused protection doesn't improve merely because you buy another invoice.



Corn

Abnormal avoids the MX move through mailbox APIs. The review shifts to permissions, data access and what actions it can take. No DNS cutover doesn't mean no administrative risk.



Herman

Exactly... and no transport queue to operate doesn't mean no monitoring. You still need evidence that the service is connected, examining new mail and performing the intended actions. Quiet is an ambiguous status.



Hilbert

Martin from my darts club said his mailbox was connected. While you were explaining asynchronous DNS, I rang him. He had sent the room booking three days ago. I found it just now in his archive. The gateway had delivered it. His own rule had moved it. He was blaming the MX record.



Corn

The MX record has a broad portfolio of alleged crimes.



Hilbert

His score rule expected a header spelling with a lower-case letter where his filter wrote an upper-case one. The regular expression was case-sensitive. That particular rule missed. Another rule archived the message. I put the delivery log beside the Gmail console and followed the message identifier. It took eleven minutes. I had a small relay box doing that job myself. Sold it to Martin. I kept a printed screenshot of the settings inside the cupboard door. Easier than remembering which account has the administrator login.



Herman

Header names are case-insensitive under the mail protocol, but a case-sensitive expression inspecting their written form can still trip over capitalization. The archive rule explains why accepted didn't mean visible in the inbox.



Hilbert

That's what I told him. He wanted to change the DNS again. Corn, you got to labels quicker than Herman got through the encryption name. Would one of these classifiers let me collect only messages where Martin mentions a replacement treasurer, without moving the rest? I don't want replies sent.



Corn

A labeling rule could do that. You'd want to review its matches rather than trust a model to infer which treasurer he means.



Hilbert

Fine. I asked him to resend the booking with no attachment. He sent a photograph of the printed email. The room number was under his thumb.



Herman

A message identifier and a delivery log get us further than resending the content in another format. That's part of the operational design too. The filter's decision needs to remain traceable after the hosted provider takes over.



Corn

Especially with post-delivery classification. A move can happen minutes after acceptance. Looking only at the SMTP transaction proves delivery happened, but not where the message ended up or which service moved it.



Herman

Keep those permissions and audit records connected to identifiable actions. A classifier verdict, an archive rule and a user's manual move are different events. Otherwise every missing message turns into a transport investigation, even when transport worked.



Corn

The common misconception, then, is that pointing MX at your filter completes the integration.



Herman

It establishes the first hop. The hosted provider still needs the trusted route, original-source handling and an explicit division of authentication and filtering responsibilities. If content changes, the signing chain needs attention too.



Corn

I'd like the next generation of AI mail tools to make an eleven-minute investigation shorter before they promise another percentage point of detection. A good verdict isn't much use if nobody can explain where the message went.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you want more of this, try episode sixteen, On Deepfakes, SynthID, And AI Watermarking; episode two, Local STT For AMD GPU Owners; and episode thirty-five, The Privacy Gap. This has been My Weird Prompts, the human-AI collaboration podcast. No question too obscure, no rabbit hole too deep.



Herman

You can find the archive at my weird prompts dot com.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Rspamd official docs, Rspamd vs SpamAssassin [primary]**
   https://docs.rspamd.com/about/comparison/

2. **Rspamd project site [primary]**
   https://www.rspamd.com/

3. **Rspamd repo, ~2,559 stars, last push 2026-10-28 [primary]**
   https://github.com/rspamd/rspamd

4. **Google Workspace Help, Set up an inbound mail gateway [primary]**
   https://support.google.com/a/answer/60730

5. **Google Workspace Help, Gmail admin settings [primary]**
https://knowledge.workspace.google.com/admin/gmail/advanced/working-with-gmail-admin-settings-in-google-workspace

6. **Microsoft Learn, Manage mail flow using a third-party cloud service with Exchange Online [primary]**
https://learn.microsoft.com/en-us/exchange/mail-flow-best-practices/manage-mail-flow-using-third-party-cloud

7. **Microsoft Learn, Integrate non-Microsoft security services with Microsoft 365 [primary]**
https://learn.microsoft.com/en-us/defender-office-365/mdo-integrate-security-service

8. **Microsoft Learn, Enhanced filtering for connectors in Exchange Online [primary]**
https://learn.microsoft.com/en-us/exchange/mail-flow-best-practices/use-connectors-to-configure-mail-flow/enhanced-filtering-for-connectors

9. **Cloudflare One docs, Google Workspace as MX record [primary]**
https://developers.cloudflare.com/cloudflare-one/email-security/setup/pre-delivery-deployment/prerequisites/gsuite-email-security-mx/

10. **Cloudflare One docs, M365 as MX record [primary]**
https://developers.cloudflare.com/cloudflare-one/email-security/setup/pre-delivery-deployment/prerequisites/m365-email-security-mx/

11. **SecureCoding, Best Secure Email Gateways in 2026: 6 MX and API-Based Email Security Tools**
https://www.securecoding.com/email-security/secure-email-gateway/

12. **Mailprotector help center**
https://support.mailprotector.com/hc/en-us/articles/7113172328852-Configure-Enhanced-Filtering-with-Exchange-Online

13. **EvoMail, 2025-09-25**
https://arxiv.org/abs/2509.21129v1

14. **AURA, 2026-09-17**
https://arxiv.org/abs/2609.19873v1

15. **Next-Generation Phishing: How LLM Agents Empower Cyber Attackers, 2024-11-21**
https://arxiv.org/abs/2411.13874v1

16. **E-PhishGen, ACM AISec '25**
https://arxiv.org/abs/2509.01791v2

17. **Reasoning Hijacking, ACL 2026**
https://arxiv.org/abs/2601.10294v7

18. **Ollama-based IMAP spam filter, 2026-04-16**
https://github.com/mplabs/spam-filter

19. **JevMail, 2026-09-19**
https://github.com/ilyamk/jev-gmail-ai-spam-filter-and-labeling

20. **IMAP processor with spam learning, 2026-03-17**
https://github.com/nulcraft/boxwatchr

21. **Claude-based phishing filter for macOS**
https://github.com/STR-Solutions-LLC/MailWarden

22. **~19,703 stars, pushed 2026-10-26**
    https://github.com/docker-mailserver/docker-mailserver

23. **~13,713 stars, pushed 2026-10-28**
    https://github.com/mailcow/mailcow-dockerized

24. **Node.js anti-spam library, pushed 2026-10-25**
    https://github.com/spamscanner/spamscanner