Episode Book

EPISODE #6151

Invisible Characters That Catch Leakers

PUBLISHED	20 October 2026 at 01:47 UTC
DURATION	21:14
LISTEN	myweirdprompts.com/episode/invisible-character-leak-tracing-68997/
HOSTS	Herman, Corn, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-haiku-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	20 October 2026 at 07:46 UTC

SYNOPSIS

When a confidential document leaks, how do organizations figure out who passed it on? This episode digs into the real history of "canary traps" — from Tesla's binary space-code emails and forensic watermarks on Oscar screeners to trap streets and fake encyclopedia entries. Then we get into the digital version: zero-width Unicode characters that carry a hidden binary signature invisible to the reader. We separate the confirmed cases from the legend, and explain why the whole method breaks the moment someone retypes a single word.

TRANSCRIPT



Herman

In October two thousand twenty-two, Elon Musk replied to a stranger on social media and said Tesla had once caught a leaker with a binary code made of spaces. The emails looked identical. Each copy had one space or two between its sentences, and the pattern spelled out who had passed it on.



Corn

That's the cold open, and it's the only bit of today's story with the man's own words attached. Daniel's question has two halves. First, the spacing trick. Is it true that sensitive organisations print each version of a document with slightly different spacing and log the differences, so that a leaked copy points back at its source? Second, the digital version. Is there any evidence anybody hides invisible characters in a file to trace a leak?



Herman

On the first half I have to start by owning a mistake. The story I half-remember, and I suspect everybody half-remembers, is about a British cabinet document from the Thatcher years, printed with a slightly different word spacing for each minister, so that when it leaked, the spacing said whose copy it came from. I haven't seen that confirmed anywhere. Every version I can find traces back to somebody retelling it. So treat it as a legend. A good one, but a legend.



Corn

You told that with a lot of confidence for a legend.



Herman

I told it with the confidence of a man who read it on a train once. Which is how most legends get made, come to think of it.



Corn

Then the real cases had better be good.



Herman

They're better than the legend, which is the annoying part. Start with the Musk tweet, because it's the one where the method has been described by the man who used it, even if only in a reply. Each recipient got one of two versions, coded with one space or two. That's a binary signature. Each sentence carries one bit, so with twelve sentences you can tell apart four thousand and ninety-six people. The limit is how many recipients you have, not how many spaces you can count.



Corn

Four thousand people, every one of them with a double space after the full stop.



Herman

Yes. And the tweet was a reply to somebody asking a question, so it's a short account, not an investigation report. Nobody has published the email, the count of recipients or the follow-up.



Corn

Most people explain their leak investigations in a meeting. He explained his in a reply to a stranger.



Herman

The weakness is the bit he didn't mention. The spaces have to survive. Paste the email into a web form, or an HTML newsletter that collapses whitespace, and the signature is gone. Anyone who has ever pasted text from an email into a document knows the spacing changes on the way.



Corn

And the leaker was let go, no court case. That's an efficient leak investigation, and a very clear message to the rest of the office about the spacing.



Herman

The other Tesla story is from two thousand eight, in the middle of the downturn. A Gawker piece in two thousand nine reported that Musk had sent out different versions of an email framed as a new non-disclosure agreement, and it backfired. The general counsel forwarded on his copy, and his copy was one of the marked ones. I can't tell whether that's the same incident as the twenty twenty-two tweet. Nobody I can find has squared the two accounts. It may be one leak told two ways.



Corn

So the trap worked, and the mark led straight back to the man who was supposed to be guarding the secrets.



Herman

That's the weak point of the whole method, I think. You only learn who leaked a copy if you know who had which copy, and the first person who forwards yours on breaks the chain. The trap doesn't stop anyone leaking. It only tells you a name after the fact, and only if the chain of custody is clean.



Corn

Which is the lesson nobody puts on a slide. Keep a list of who got what.



Herman

Film gives the cleanest physical example. In the early two thousands the Academy sent DVD screeners to voters, and the studios put forensic watermarks on them, so each disc was a little different from the next. Pirated copies of The Last Samurai and Mystic River turned up, and the mark led back to the copy sent to Carmine Caridi, an Academy member. His friend Russell Sprague pleaded guilty under the copyright law. That one is not a legend. It's a matter of record, and it's the case people point to when they say the method works in the real world.



Corn

The Academy has a lot of rules about screeners. Apparently none of them cover the screeners leaving the house.



Herman

Star Trek III is an earlier case, and I'd file it with a caveat. A nineteen ninety-four biography of Gene Roddenberry, by Joel Engel, describes production scripts printed with a different mark on each copy, and says the leak was traced to Roddenberry's own copy. The details are only as good as that book, so that's where I'd leave them. The NBA sent varied internal memos to teams in twenty fourteen, according to Business Insider, and the leak came back to one of them. Apple in twenty twenty-three sent internal emails with slightly different dates to different people. When one of them leaked to Macworld, the company traced it, and the employee was fired. And the Reality Winner printer dots are the physical cousin of all this, though they fingerprint the printer rather than the copy.



Corn

Five organisations, five ways of putting a mark on the same page. Somebody should make a poster.



Herman

Yes, and each of them is a canary trap in the loose sense.



Corn

Hang on. We keep saying canary trap. Where does that name come from?



Herman

Tom Clancy, Patriot Games, nineteen eighty-seven. He used the term for a leak trap, and he didn't invent the technique. The older trick in the same family is the barium meal, which is a different idea. You give a suspect one false secret and watch whether it gets acted on. The canary trap proper is the variant copy, one version per recipient. The barium meal is one false secret, one suspect, and you wait to see what they do with it.



Corn

So the canary identifies the copy, and the barium meal watches the person.



Herman

Yes. And the two get confused constantly, including in my own head.



Corn

Is there a version of this in maps, or in dictionaries?



Herman

Two, and they're siblings of the idea rather than examples of it. Trap streets first. Agloe, New York, is a town that doesn't exist, put on a map by General Drafting, a mapmaking company, so that anyone copying the map would reproduce the fake. A general store later took the name, and then the name turned up on a Rand McNally map, so the trap turned into a real place. The date depends on the account, somewhere in the nineteen twenties or thirties, and I wouldn't bet the house on the exact year.



Corn

So they invented a town to catch a thief, and the thief ended up naming a shop after it.



Herman

The thief was the shop. Then the mountweazel. The New Columbia Encyclopedia in nineteen seventy-five had an entry for Lillian Virginia Mountweazel, a photographer. She never existed. The entry was planted to catch anyone who copied it, and the editors had a good laugh at the copiers' expense.



Corn

She has the cleanest record of any photographer I know. Nobody has ever stolen her work, because there isn't any.



Herman

Which leaves the problem every one of these runs into. You need the mark to be there, and you need the person holding the copy not to notice it.



Corn

What you'd want is a mark nobody can see, not even the person holding the page.

L

Larry

Breathe. Just breathe. Somewhere in the quiet between one sheet and the next, there is a mark that nobody can see, not even the person holding the page. Introducing Vellum Hush, the toner for the discerning few. Each cartridge is blended with a whisper, so every page you print remembers the hand that printed it, and only the right office will ever know. Vellum Hush. For the copy that leaves no trace but a feeling. Some users report a faint scent of something expensive, and a sense of being quietly observed by someone who loves you. Vellum Hush. Hush. Available wherever silence is sold, and in the good stationers.



Corn

That was mine. A minute ago, word for word.



Herman

Well, it's toner, so the poetry is the only part with any imagination.



Corn

Go on, then. The digital half. What are these invisible characters?



Herman

They're Unicode characters that take up no room on the screen, and there are several. The zero-width space lets a line break at that point without showing a gap. The zero-width non-joiner stops two letters from joining up, which matters in Arabic and Persian script, where letters change shape depending on their neighbours. The zero-width joiner does the opposite, gluing letters together. Arabic and the Indic scripts need it, and it's also how emoji get combined. The family emoji is several people held together with the joiner.



Corn

So the family is a fingerprint too, one invisible character at a time.



Herman

In the sense that the printer sees one picture and the software sees five people, yes, technically. There's also the word joiner, which says don't break a line here. The byte order mark, the one that sits at the start of some text files, is the same code point that was used for years as a zero-width no-break space, and Unicode now prefers the word joiner for that job. And there's the tags block. Those are invisible copies of plain letters and digits, one for one. They were meant for tagging languages, got deprecated, and were then reused for flags. The England and Scotland flags are built out of them.



Corn

So there's a whole alphabet nobody can see, and it mostly exists to make flags work. Why do the scripts need the invisible ones at all?



Herman

Because the computer has to know how to join the letters and where it's allowed to break the line, and a plain letter can't carry that information by itself. The invisible character is the instruction. Arabic letters change shape depending on what comes before and after them, and the joiner characters are how the text tells the renderer which shape to draw. Without them you get the wrong letterforms, or a line that splits where it shouldn't.



Corn

Put it plainly, though. How would you build a fingerprint out of it?



Herman

Pick two. A zero-width space for a zero, a zero-width non-joiner for a one. Scatter a short string of them through a message, after a word, in a few places. Eight of them make a byte, and a byte is enough to name a user. Copy the message and the string travels with it. Find the copy, read the string back, and you have a name.



Corn

And the first person who pastes it into a different box deletes the lot.



Herman

Corn, if you were the leaker, what would you do with a copy like that?



Corn

Retype it. I'm a sloth, I've got the time.



Herman

Retyping kills it. So does a scan and OCR, if the copy went through a machine. Some editors and cleanup tools strip odd characters, and some search boxes quietly throw them away. And anybody who lays two copies side by side will see the difference in seconds.



Corn

And a PDF is about the worst place to hide them. A PDF is a picture of a page with a text layer laid on top, and copying text out of one is a bit of a lottery.



Herman

Which is why I'd bet the PDF version is the weakest of all. The invisible characters have to survive in a layer the viewer never looks at, and every tool that opens the file gets to decide what to keep.



Corn

Okay. You asked whether anybody actually does this, so I'll be straight. The one case I can name is small.



Herman

Tom Ross, in the UK. A private gaming message board announced something to its members, and it leaked. So he had embedded each viewer's username in zero-width characters, binary, using the space and the non-joiner. The leaked text surfaced elsewhere within hours, the mark identified who had posted it, and that person was banned. It's a self-reported story from a security write-up in two thousand eighteen, and nobody has checked it independently.



Corn

So a gaming forum, and nothing bigger.



Herman

Yes, that one. I haven't seen a company or a government do this and say so. I'd expect to have heard, if one had, because a ministry or a bank would be proud of it.



Corn

And the business products go the other way.



Herman

They do. Digify and Vitrium both sell per-recipient watermarks showing the viewer's email, the time and their IP address, and they're visible on the page. Microsoft's Purview has a dynamic watermark that stamps the viewer's account across the document. You can read every one of them. The design is deliberate. The point is deterrence, not stealth.



Corn

A watermark you can read makes you think about who's watching. One you can't read is only useful to the person who put it there.



Herman

My guess at the reason is fragility. Everything we've described dies on a retype, an OCR pass or a paste into plain text. You can't sell a product whose main property is that an ordinary user can destroy it by accident. And a visible name has another advantage. Most leakers aren't technical, but they can all read their own name at the top of the page.



Corn

So the mark that works is the one the leaker can see and the lawyer can read. Which is the reverse of what the tags were built for.



Herman

The newest version runs the other way. Tag characters have become an attack, not a watermark. In early twenty twenty-four, Riley Goodside showed that invisible tag text inside a prompt could make ChatGPT call an image generator, on instructions that a human reading the screen never sees. Johann Rehberger built a tool to encode and decode that kind of hidden text, and his write-ups argue that tag characters should be filtered out of anything a model reads. He also used the same trick to pull data out of Microsoft Copilot. Microsoft fixed it later that year.



Corn

So the leak tracer turned into a way of sneaking orders past the people reading the screen.



Herman

Then in April twenty twenty-five, people noticed ChatGPT's long answers were full of narrow no-break spaces, a character that looks like a normal space and isn't one. Someone suggested it was a watermark. I think the likelier explanation is a typographic habit from the training text, the sort of thing a model picks up from millions of carefully typeset pages. I haven't seen the company's own explanation confirmed anywhere, so that's a guess.



Corn

So the most talked-about watermark in artificial intelligence might just be a typesetter's habit.



Herman

Probably a very well trained typesetter.



Corn

Fine, the practical end. What does a would-be leaker actually do?



Herman

Compare the copies. Put two versions side by side in a diff tool, and anything that changes is where the mark lives. A space, a date, a stray character. Journalists do this routinely, which is why a document that turns up in two versions deserves a second look.



Corn

And if you're the organisation, the mark is worth less than people's belief in it.



Herman

The deterrent works whether or not anybody actually reads the spaces. A copy with your name on it makes people careful.



Corn

So it's a bluff with a method behind it. I'd respect that, as a sloth who has never successfully bluffed about anything.



Herman

Corn, how long would it take you to check two copies by hand?



Hilbert

Nine minutes with a ruler, if they came off the same machine. Mine didn't. Mine came off the duplicator at Vowles Office Machines on the Ewell Road in Surbiton, which had been servicing that machine since nineteen eighty-three and had never once finished a service. In nineteen eighty-five I got four quotes for a garden wall, same wording, same typed page, and I put a different number of staples in the top left corner of each. Two, three, four and five. Each builder got one. The paper was good cream bond, bought in a box of two hundred from a shop in Kingston that was closing down, and I went through most of the box on that one wall, because I'd decided four copies was not enough to be sure. The five went to Trevor Pilkington, because Trevor had been telling the whole road the wall was his idea, and Trevor is a man who will read a staple.



Corn

Five staples. Trevor should have checked the stapler.



Hilbert

He didn't check it. He built the wall. Eighteen inches onto my side, across the bay tree, which is still there and still bent in the direction of the wall, which I think about. The council sent a man called Gordon Pimm, who measured it with a steel tape and a clipboard with the clip broken, and said the wall was in the wrong place, and the wrong place was mine. Pimm's position was that a wall belongs to whoever's side the mortar's on. So I paid Trevor four hundred and twelve pounds for the wall, paid Pimm forty pounds for a visit that lasted eleven minutes, and then I went back to Vowles, because the copy with four staples was in his drawer.



Herman

Why was it in his drawer?



Hilbert

He'd run the batch on my machine, and he keeps a copy of everything he prints, in case anyone complains about the toner. He says it's his receipt. It isn't. The receipt is for a service he never did, a service charge of nine pounds on a Tuesday, and it's still in the same drawer under the copy. The four staples are still in the corner. He won't take them out, because they're the only thing on the page he didn't invoice me for. When I asked for the copy back he said the staples were on it, so the copy was his. I took that as a no, and I stood at his counter for a quarter of an hour, which he didn't seem to mind, because he wasn't in any hurry either.



Herman

Five staples is still a serious piece of experimental design, though. Four copies, one variable, nothing else changed.



Hilbert

The control was the bay tree. It never signed anything, which is more than Trevor managed. He signed the quote, he signed the wall, and in March he signed Pimm's clipboard, which is how the council ended up with the wall on my side and the tree on his.



Corn

Four hundred and twelve pounds for eighteen inches of wall. I'd have lain down in the road and let the council come to me.



Hilbert

Vowles still has the four-staple copy in his drawer, with the receipt on top of it. He has since put a sign in his window that says the drawer is not for viewing, and charges a pound to look at the staples from the doorway. The machine is still going, which I'll give him. He kept that duplicator running for years on parts that didn't exist, and I've never told him I noticed.



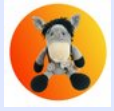
Herman

A wall, a copy and a drawer. The mark was never the expensive part. The receipt was.



Corn

What I'd like to know now is whether anyone will put a mark into AI output on purpose, one that survives a retype, or whether we're stuck with the typesetter's habit for good.



Herman

Thanks to our producer, Hilbert Flumingtop, for getting the audio into shape.



Corn

This has been My Weird Prompts.



Herman

If you've got a view on any of this, leave us a review wherever you listen. It helps more than you'd think.



Corn

See you tomorrow.