PROBLEM? RELAX. YOU'VE GOT THIS!

I Wrote It Myself, but the Software Says I'm a Robot

How to Document Your Creative Process, Respond Calmly to Accusations of AI Use, and Present Evidence of Authorship Without Treating a Detector's Score as the Final Verdict



Max Paradox

I Wrote It Myself, but the Software Says I'm a Robot

How to Document Your Creative Process, Respond Calmly to Accusations of AI Use, and Present Evidence of Authorship Without Treating a Detector's Score as the Final Verdict

Max Paradox

PROBLEM? RELAX. YOU'VE GOT THIS!

ISBN: 9788384839508

ALL-IN Rafał Płonkowski



CONTENTS

INTRO 5

Chapter 1 - When a Question About Your Writing Becomes an Accusation 11

Chapter 2 - What an AI Detector Actually Knows 19

Chapter 3 - Apparently, You're Too Good at Writing to Be Human 27

Chapter 4 - The Evidence Panic: How to Make a Bad Situation Worse 35

Chapter 5 - Your Writing History Shouldn't Disappear Every Time You Click Save 44

Chapter 6 - When AI Helps You Write and When It Starts Writing for You 53

Chapter 7 - Whose Words Are These, Anyway? 62

Chapter 8 - How to Have the Conversation Without Turning It into a Trial 71

Chapter 9 - Build an Evidence File That Makes Sense Without You in the Room 81

Chapter 10 - Write the Reply That Actually Moves the Case Forward 90

Chapter 11 - Follow One Paragraph Back to Its Beginnings 100

Chapter 12 - The Decision Went Against You. Now What? 109

Chapter 13 - Agree on the Rules Before Someone Starts Looking for a Robot 118

Chapter 14 - Verify Authorship Without Starting a Witch Hunt 127

Chapter 15 - When an AI Accusation Escapes into the Wild 136

Chapter 16 - When Your Own Evidence Starts Contradicting Itself 145

Chapter 17 - The Evidence Exists, but You Can't Show It 155

Chapter 18 - A Writing Routine That Doesn't Need Its Own Assistant 164

Chapter 19 - When the Fear Comes Back After the Problem Is Over 174

Chapter 20 - You Do Not Need a Certificate of Humanity 183

INTRO

The email arrived at 9:14 on a Tuesday morning, accompanied by a screenshot and a sentence that could ruin anyone's breakfast: "Our AI detection software indicates a 91% likelihood that your article was generated by artificial intelligence." The recipient, a freelance writer named Daniel, stared at the message while his coffee quietly surrendered its remaining warmth. He had spent six days researching the article, interviewing people, rearranging paragraphs, and arguing with himself over whether the introduction needed another example. Apparently, all that effort had been unnecessary. According to a colorful little dashboard, the real author was probably a machine. Daniel considered replying with a photograph of his notebook, his desk, and perhaps his birth certificate. Then he wondered whether the software might accuse his birth certificate of being AI-generated too. At that point, the situation had officially become ridiculous.

The really uncomfortable part wasn't the number. It was what the number appeared to suggest about Daniel's honesty. His client wasn't simply questioning whether a sentence sounded natural or whether the article needed editing. The message raised doubts about who had done the work and whether the agreed conditions had been respected. For someone whose income depends on delivering original material, that distinction matters. An awkward headline can be rewritten. A factual error can be corrected. A suspicion of dishonesty can follow you into future conversations, especially when somebody forwards a screenshot without the explanation that belongs beside it. Suddenly, a routine assignment becomes a discussion about professional integrity, and the person who wrote the article finds themselves defending an event that nobody else witnessed: the process of writing it.

Artificial intelligence has created an unusual problem for people who work with words. Generative tools can produce fluent, structured material, sometimes with remarkably convincing confidence. Meanwhile, human beings continue doing something they have done for centuries: writing fluent, structured material, occasionally with confidence and occasionally while wondering why the same paragraph has needed eleven revisions. The finished products can share characteristics, and distinguishing their origins from language alone is not straightforward. Some detection systems attempt to classify text by analyzing patterns associated with particular types of writing. Their methods and performance vary, and the information they provide must be interpreted within the limits of the system being used. What they generally cannot do is travel backward through time and watch someone make every creative decision. A statistical assessment of a document is not a security-camera recording of its creation.

That distinction sounds simple until a percentage appears on an official-looking report. Numbers have an extraordinary ability to borrow authority from their surroundings. Put an uncertain conclusion beside a progress bar, add an alarming color, and suddenly people behave as though a laboratory has discovered an inconvenient biological fact. Yet a detector's displayed score does not necessarily represent the actual probability that a particular person used AI. Its meaning depends on how the system defines and calculates that score, how it was evaluated, and the circumstances in which it is being used. A confident-looking interface cannot supply information that its underlying method never established. This doesn't mean every detection tool is worthless or every suspicion is unreasonable. It means that recognizing a possible pattern and establishing what happened are different tasks, even when software companies would undoubtedly prefer customers to find both tasks behind the same convenient button.

If you've ever been accused of using AI when you genuinely wrote the material yourself, you may recognize what happens next. First comes disbelief. Then anger. Then the tempting idea of submitting the same text to six other detection services until one of them produces a result you like. Unfortunately, contradictory scores do not automatically prove which system is correct. They may reveal disagreement between tools, but disagreement is not a complete account of authorship. Some writers respond by changing their sentences to satisfy a detector, even when those sentences were perfectly good. Others send increasingly emotional explanations, hoping that sincerity will somehow become measurable. Before long, they are spending more time trying to look like legitimate writers than doing the writing they were hired to do. The software hasn't merely questioned their work. It has rented an office in their heads and begun issuing editorial instructions.

There is another uncomfortable truth, however: saying "I wrote it myself" doesn't automatically establish that you did. You might be telling the complete truth, but another person may still need information that can be evaluated independently. Perhaps you have earlier drafts, research notes, correspondence about the assignment, or a document history showing how the material developed. Perhaps you can explain why you discarded a particular argument, changed the structure, or selected certain sources. Each piece of information has strengths and limitations. A file timestamp doesn't certify the origin of every sentence, and a saved draft doesn't reveal everything that happened before it existed. But authentic materials, presented with appropriate context, can make the creative process far easier to understand. That's considerably more useful than attempting to win an argument through the sheer volume of your indignation.

Of course, real life rarely offers perfect documentation. You may have written everything in one document, overwritten earlier

versions, or lost access to the platform where the work was created. You may be balancing assignments, deadlines, family responsibilities, and a budget that doesn't include hiring a digital forensic specialist every time you finish an article. This book takes those limitations seriously. You don't need to record every keystroke or turn your working day into an evidence-collection exercise. You do need to understand which records are worth preserving, how to interpret them honestly, and what can still be explained when the available history is incomplete. Inventing missing drafts or presenting reconstructed notes as original records is not a solution. An imperfect but truthful explanation is more valuable than a beautifully organized fiction.

We also need to discuss AI use without pretending that every possible form of assistance is identical. Asking software to identify spelling mistakes, requesting suggestions for reorganizing your own notes, and generating an entire article are different activities. Their acceptability depends on the rules governing the particular assignment, publication, institution, or agreement. Sometimes disclosure is required. Sometimes certain uses are forbidden. Sometimes the instructions are so vague that everyone involved seems to be interpreting a different document. A responsible response begins with describing what actually happened, not selecting whichever label sounds most reassuring. This book will not teach you to disguise prohibited AI-generated work or manufacture evidence of human authorship. It will help you distinguish legitimate assistance from work you cannot honestly claim to have performed yourself.

Just as important is knowing how to communicate when someone raises a concern. An editor asking a question, an employer conducting a review, and an institution initiating a formal proceeding do not present identical situations. The appropriate response depends on the allegation, the applicable rules, the available evidence, and the potential consequences. Sometimes a

short explanation is sufficient. Sometimes you need to request clarification, present a carefully organized record, or seek qualified professional advice. When legal rights, contractual obligations, or serious academic penalties are involved, general guidance cannot replace advice suited to the particular circumstances. Remaining calm doesn't mean accepting an unfair conclusion. It means making decisions based on what the situation actually requires rather than allowing the most irritating sentence in someone's email to control your next three hours.

The broader aim is to reclaim something that should never have become controversial: the ability to create worthwhile work without constantly wondering whether an algorithm will approve of your humanity. Good writing involves judgment, revision, attention, knowledge, and occasionally the remarkable discovery that deleting an entire page improves everything. Those activities deserve to be understood on their own terms. You can learn to preserve meaningful traces of them, answer reasonable questions, challenge unsupported conclusions, and acknowledge uncertainty without surrendering your credibility. You cannot guarantee that every person or institution will agree with you. You can make your account accurate, proportionate, and supported by whatever authentic information is available.

And perhaps that is where the real freedom lies. You do not need to prove that you are incapable of using modern technology, nor must you convince every detection service that your sentences possess an appropriately human level of imperfection. Your responsibility is to know what you did, respect the rules that apply, and describe your work truthfully. If someone questions it, you can respond with clarity rather than panic. If evidence is missing, you can acknowledge the gap instead of filling it with imaginary documents. And if a piece of software continues to insist that your writing resembles something produced by a machine, remember that resemblance is not a complete history.

You have writing to do. The algorithm can keep its opinions. It doesn't get the final word on your authorship simply because it knows how to display a percentage.

Chapter 1 - When a Question About Your Writing Becomes an Accusation

The first sign of trouble was the subject line: "Concern Regarding the Authenticity of Your Submission." Sophie had submitted her final university assignment four days earlier. She had spent the previous month researching, drafting, and revising it, occasionally wondering whether choosing a different course might have been a more peaceful life decision. Now an administrator wanted to discuss the possibility that her work had been produced using artificial intelligence. The message referred to an automated assessment but offered little explanation. Sophie read it twice, checked that the sender's address was genuine, and immediately imagined herself being expelled, blacklisted, and forced to explain the incident at every future job interview. By lunchtime, she had mentally progressed from receiving an email to living under a bridge. The university, meanwhile, had merely requested a meeting.

This is how an authorship concern can become overwhelming before anyone has established what actually happened. The words "AI-generated" carry implications that extend beyond technology. They can suggest dishonesty, an unfair advantage, or a failure to complete work independently. For a student, the consequences might affect academic standing. For a contractor, they could involve payment or professional relationships. For an author, they might raise questions about publishing standards and reputation. These are legitimate reasons to take a concern seriously. They are not reasons to assume that every inquiry has already become a formal finding of misconduct. Understanding that difference is the first step toward responding effectively.

People are especially vulnerable to misunderstanding messages when the language is vague. A reviewer might write that a passage

"appears machine-generated," that a detector "identified suspicious characteristics," or that a document "requires additional verification." Such statements do not necessarily mean the same thing. One describes an impression, another refers to an automated output, and the third identifies a need for further examination. None automatically establishes that a particular rule was broken. Unfortunately, worried readers often supply the missing conclusion themselves. The moment someone asks whether AI was involved, they hear an accusation of cheating. Sometimes that interpretation is justified by the circumstances. Sometimes it turns an ordinary request for clarification into a confrontation that nobody intended.

Separate what happened from what you fear might happen

Before composing a response, identify the exact status of the concern. Has someone asked an informal question? Has an assignment been temporarily withheld from grading? Has an employer initiated a formal investigation? Has a client refused to accept completed work? These situations may require different procedures, documents, and levels of urgency. A suspicious comment from a colleague does not carry the same authority as an official determination issued under an institutional policy. Nor does a detector's output become an institutional decision merely because someone attached it to an email.

Sophie initially wanted to write, "I strongly object to your decision to accuse me of academic dishonesty." But the administrator had not yet made such a decision. Responding as though an adverse finding already existed might have complicated a process that was still at the information-gathering stage. A more useful first step was to read the notice carefully and determine what it actually requested. She needed to know whether the matter was informal or formal, which parts of her assignment raised questions, what policy applied, and whether there was a deadline for responding. Those details would shape her next actions far more effectively than an eloquent declaration of outrage.

This distinction matters because the person raising the concern may have a legitimate responsibility to verify compliance. A teacher may be required to assess whether submitted work meets the assignment's independence requirements. An editor may need to confirm that commissioned material follows publication rules. A business may have contractual obligations concerning the origin of content supplied to customers. Asking questions is not inherently unfair. The problem arises when suspicion is treated as proof, when standards are unclear, or when someone draws conclusions that the available information does not support.

There is an important difference between respecting the right to investigate and accepting every investigative method as reliable. You can cooperate with a reasonable inquiry while questioning how a tool's result was interpreted. You can acknowledge that an institution must enforce its rules without agreeing that a particular rule was violated. You can also ask for clarification without treating the person on the other side as an enemy. These positions are compatible, even if angry online debates tend to present them as mutually exclusive.

The practical task is to identify what can be verified. If a reviewer claims that you submitted prohibited AI-generated material, the questions become specific. Which material? What conduct is alleged? What rule would make that conduct improper? What information supports the allegation? A statement that your writing "sounds like ChatGPT" may explain why someone became suspicious, but it does not provide a complete account of the work's origin. An accusation becomes meaningful only when it identifies something that can be examined.

The first twenty minutes can create unnecessary problems

Receiving an unexpected challenge can make immediate action feel essential. Some writers begin rewriting the disputed document to make it appear more natural. Others upload it to multiple detection websites or hurriedly create new files describing how

they supposedly worked. These reactions may produce additional confusion. If the original submission is altered, the distinction between the work being questioned and later revisions becomes harder to maintain. If new records are presented without explaining when they were created, their relationship to the original process may be misunderstood.

The safer starting point is to preserve the submitted version exactly as it existed when the concern arose. Keep the original communication, relevant attachments, and any authentic working materials that remain available. Avoid changing dates, reconstructing drafts as though they were contemporaneous records, or deleting inconvenient details. You may need to create copies for examination, but the underlying materials should remain distinguishable from anything prepared after the inquiry began.

This does not require sophisticated forensic software. In an ordinary situation, preserving existing documents and correspondence may be sufficient for the initial response. If the matter involves serious disciplinary, contractual, or legal consequences, the relevant procedure may require more careful preservation and professional assistance. The important principle is straightforward: do not accidentally damage your ability to explain the past while attempting to make the present more comfortable.

Sophie discovered that she had been writing in a document with an accessible revision history. She also had research notes and feedback from an earlier consultation. Those materials might help explain how her assignment developed, although none was a universal certificate of independent authorship. Rather than immediately exporting everything and attaching a chaotic archive to her response, she first established what the university wanted to examine. This prevented her from providing unnecessary information and allowed her to focus on documents connected to the actual concern.

Preservation also involves protecting other people's information. A working folder might contain private messages, unpublished research, confidential client material, or personal details unrelated to the assignment. The fact that somebody has questioned your work does not automatically entitle them to unrestricted access to your computer or accounts. Information should be provided through an appropriate process, with attention to confidentiality and applicable requirements. Cooperation does not mean surrendering every boundary.

Give an answer that addresses the actual question

A useful initial response has a surprisingly modest purpose. It establishes your position, acknowledges the concern, and identifies a sensible way to clarify the facts. If you genuinely prepared the work independently, you can say so directly. If you used permitted assistance, describe that assistance accurately when relevant. If the inquiry is unclear, ask what needs to be explained.

For example, Sophie might write: "I prepared the submitted assignment myself and would be happy to explain my research and drafting process. I have working notes and earlier versions available. Could you please identify the specific concerns and confirm the procedure for reviewing them?" This message does not attempt to win the entire case immediately. It supplies useful information while inviting the institution to define the issue properly.

If Sophie had used a generative tool for part of the assignment, her statement would need to reflect that reality. The distinction between independent work, permitted assistance, and prohibited generation depends on what actually occurred and on the applicable rules. It would be unwise to claim that no AI-related tool was involved if she knowingly used one in a relevant capacity. It would be equally unwise to concede that her entire assignment was machine-written merely because she used an allowed function for

language correction. Accuracy serves her interests better than dramatic certainty.

Sometimes a person cannot remember every detail of the writing process. Perhaps the work was completed weeks earlier, across several devices, without saved intermediate versions. This is not unusual. Human memory is not a continuous recording system, despite the confidence with which people occasionally describe conversations from seven years ago. If you cannot establish a particular fact, distinguish what you remember from what the records show. "I believe I drafted that section during the second week, but I no longer have a dated version" is more defensible than inventing a precise date to make your story look complete.

A particularly unhelpful strategy is producing an enormous personal statement about your character. Writers sometimes respond to a specific inquiry by explaining their commitment to integrity, their years of experience, and the sacrifices involved in completing the assignment. Such information may provide context, but it cannot replace evidence concerning the work under examination. Someone can have an excellent professional history and still make a mistake. Someone can also possess limited experience and genuinely produce an outstanding document. The relevant question concerns this particular submission, not whether the author's entire life deserves a certificate of good behavior.

When the inquiry has serious consequences

An informal discussion may become a formal proceeding, and it is important to recognize that transition. If an institution issues an official allegation, provides an investigation notice, or announces a disciplinary process, examine the applicable rules carefully. Deadlines, access to records, representation, appeal opportunities, and evidence requirements vary between organizations and jurisdictions. General advice about communication cannot replace the procedures governing a particular case. Where significant academic, employment, financial, or legal consequences are

possible, appropriate professional or institutional support may be necessary.

This also means that an immediate response is not always the best response. If a formal notice allows time to prepare an explanation, use that time purposefully. Determine what documents exist, which statements you can support, and what the decision-maker is authorized to consider. Avoid making unnecessary admissions or absolute claims before you understand the question. At the same time, do not ignore a deadline merely because the allegation feels unreasonable. Administrative processes have an unfortunate talent for continuing while participants are busy being offended by them.

Not every inquiry will end favorably. Some reviewers may regard the available evidence as inconclusive, and others may interpret particular requirements differently. Your objective is not to force immediate agreement. It is to ensure that your explanation is accurate, relevant, and presented through the appropriate channel. A well-prepared response can clarify misunderstandings, preserve important distinctions, and create a reliable record of what you submitted. It cannot guarantee another person's conclusion.

Sophie eventually approached her meeting with a different understanding of the situation. She no longer needed to persuade the administrator that questioning her was morally outrageous. She needed to explain how she had completed the assignment and identify the information available for checking that explanation. She could answer questions about her argument, describe her research choices, and point to genuine working materials. If the institution raised a specific allegation, she could respond to it under the appropriate procedure.

There is a broader lesson here. An authorship concern should not be allowed to define the entire relationship between you and your work. You may feel insulted, frustrated, or anxious. Those

reactions are understandable, but they do not have to determine what you write next. The strongest opening response is often ordinary, precise, and slightly boring. It asks what is being questioned, identifies what can be established, and protects the integrity of the available evidence.

You do not need to turn your first email into a courtroom speech. You need to discover whether there is actually a case to answer, what the case concerns, and how the relevant facts can be examined. That is a much more useful beginning than announcing that you will defend your honor against the entire software industry. The software industry, inconveniently, has a rather large mailing list.

Chapter 2 - What an AI Detector Actually Knows

Imagine a machine that examines a cake and announces, with 93% confidence, that it was baked by somebody named Susan. You ask how it knows. The machine explains that Susan's cakes are generally round, contain flour, and have frosting. These observations may be perfectly accurate. They are also spectacularly insufficient to identify the baker. You would probably want to know whether anybody saw Susan preparing the cake, whether the ingredients came from her kitchen, and whether other people make similar desserts. Instead, the machine displays a large green check mark and invites you to purchase the professional subscription. That is the problem with confusing recognition of patterns with knowledge of events. A result can appear precise while answering a narrower question than the one people think they asked.

AI text detection presents a more complicated version of this problem. Various systems attempt to distinguish writing produced by generative models from writing produced by people. Some examine statistical properties of language, while others use classifiers trained on collections of examples or more complex combinations of features. Their exact methods differ, and commercial systems do not always disclose enough information to permit a complete independent assessment. These tools may provide useful signals in particular circumstances, but their output must be interpreted according to what the system was actually designed and evaluated to do. A detector analyzing the final document does not normally have direct access to the history of decisions that produced it.

That limitation matters because a completed text contains information about its language, structure, and subject, but not a complete record of its creation. The same sentence might have

been composed by a human writer, suggested by a language model, revised by an editor, or produced through several stages involving different contributors. Looking at the finished sentence alone may reveal patterns consistent with one explanation, but consistency is not exclusivity. A passage that resembles machine-generated writing does not become machine-generated merely because of that resemblance. Likewise, a passage that looks unmistakably human is not guaranteed to have been written without automated assistance.

Why the percentages deserve an interrogation

The first thing to understand about a detector score is that its meaning cannot be safely guessed from its appearance. Suppose a report displays "92% AI." Many readers interpret this as a 92% probability that artificial intelligence wrote the document. But a percentage may represent a classification score, an estimate of the proportion of text considered suspicious, or another measure defined by the provider. Whether it corresponds to a properly calibrated probability of a particular event requires separate evidence. Without that information, the number's interpretation remains uncertain.

This distinction becomes particularly important when reports move between people. The person running the detector may understand that its output is an indicator requiring review. A manager receiving a screenshot may interpret it as a probability. Someone further along the email chain may write, "The software confirmed that the author cheated." By the time the result reaches the person whose work is being questioned, an uncertain technical assessment has undergone several promotions. It began as a score and arrived as a verdict, having apparently skipped the inconvenient stages involving evidence and reasoning.

Even a system that performs well under specified testing conditions can make mistakes. A false positive occurs when a system incorrectly identifies human-written material as AI-