AI Slop Side Effect Database

A database of cases in which the proliferation of low-quality AI-generated content caused platforms and systems to change, resulting in indirect harm to legitimate users, creators, and researchers.

UTIE Research Institute / UTIE Instruments Inc.

https://utie-instruments.com/utie-research-institute.html

Scope and Inclusion Criteria

This database was compiled as a sister database to the AI Incident Database. While the main database records cases in which AI was used directly and caused harm, this database records cases in which the proliferation of low-quality AI-generated content caused changes to platforms or systems, resulting in indirect harm to legitimate users, customers, researchers, and creators.

There are three inclusion criteria: (1) the influx of AI slop must be the cause of the systemic or design change; (2) the harm must have been mediated through a defensive transformation of the platform or system; and (3) specific victims (individuals or legal entities) and concrete harm must be verifiable. Events that are ethically problematic or may occur in the future are described in the Observations section and are not recorded as individual cases.

Type Definitions

Code	Name	Definition
GATE	Gatekeeping Collapse	Automated filters/detection tools introduced as AI slop countermeasures wrongly excluded or banned legitimate content and legitimate users.
CONT	Content Contamination	Mass influx of AI slop degraded the quality and reliability of a platform, causing disadvantage to legitimate users.
BIAS	Discriminatory Bias	Bias caused systematic and unjust exclusion of specific attributes (non-native speakers, independent researchers, etc.).
INVIS	Institutional Invisibility	Platforms that officially claim to be open unjustly excluded content from search results.
COVER	Organizational Concealment	After being pointed out for malfunctions or side effects of AI slop countermeasures, the organization systematically downplayed or concealed them.

** A single incident may be assigned multiple types.*

[Notice of Data Disclosure Restrictions]

The cases presented in this public database are limited to those that have been made visible to the general public.

Our broader insights into other AI risks, undisclosed incidents, and proprietary defense frameworks are continuously accumulated and managed internally to prevent technological misuse and the increasing sophistication of fraudulent activities. More in-depth insights and analyses are provided exclusively through our business operations and official advisory services.

Part 1: Academic Infrastructure

PC-1 through PC-4 — *Omitted in this English edition. Please refer to the Japanese version.*

Cases PC-1 through PC-4 covering academic infrastructure (preprint repositories, academic databases) are omitted in this English edition. Please refer to the Japanese version.

Part 2: Creative and Educational Domains

PC-5 through PC-8 — *Omitted in this English edition. Please refer to the Japanese version.*

Cases PC-5 through PC-8 covering publishing, literature, and education are omitted in this English edition. Please refer to the Japanese version.

Part 3: Digital Platforms

PC-9 through PC-11 — *Omitted in this English edition. Please refer to the Japanese version.*

Cases PC-9 through PC-11 covering video, search, and freelance platforms are omitted in this English edition. Please refer to the Japanese version.

Findings

What Is AI Slop Pollution?

The cases recorded in this database share a common thread that can be understood through the analogy of environmental pollution. When a factory dumps toxic substances, a river becomes contaminated, and the downstream residents who have done nothing wrong suffer the consequences. When those who mass-generate AI slop flood the environment, platforms become defensive, and legitimate users who have done nothing wrong are caught in the crossfire. The non-native speakers in PC-4 did not use AI at all. Their human-written texts were misidentified by AI detection tools, and their academic and research careers were destroyed. The researchers in PC-1 and PC-2 posted no AI slop whatsoever. As a result of a platform's defensive response to other people's slop, they were removed from search results or had their papers erased by AI declaring "this paper does not exist." Even in cases recorded in the main database, victims often have no option of simply choosing not to engage with AI. The elderly patient in INT-23 had their insurance benefits cut off by AI introduced by their insurer; the consumer in INT-1 had a 23% price markup applied without their knowledge by AI; and the four-year-old in JP-9 died without even knowing that AI existed. Human beings who do not use AI suffer harm caused by AI, and they cannot even recognize that the harm was caused by AI. This is the basis on which we describe AI incidents as pollution.

Domains Where Further Expansion Is Anticipated

The automated filters that platforms introduce as AI slop countermeasures will inevitably be set to conservative thresholds. This is because the organizational logic holds that catching legitimate content in the net is safer than

missing slop. The moderator strike in PC-3 (Stack Overflow) is a textbook case of this contradiction exploding, and together with PC-4 (systematic misidentification by AI detection tools) and PC-7 (Texas A&M false accusations), the three can be referenced as the "AI Detection Tool Trio." Although specific victims cannot yet be confirmed sufficiently to include them in this database, the following domains are predicted to see expansion of the same structural problem: (1) wrongful bans on music and video platforms (Bandcamp and others have already introduced AI-prohibition policies; misidentification is predicted to be only a matter of time); (2) disqualification from literary and art competitions due to AI detection errors (already widely reported on social media); (3) degradation of screening processes due to mass submission of AI-generated documents for job applications and grant reviews (already widely reported on social media). These will be added to this database as specific victims are confirmed and stronger evidence and documented harm are found.

Relationship to the Main Database

UTIE Instruments Inc., the victim in AI Incident Database (main database) case JP-0 (Springer Nature peer review fabrication incident), is also the victim in PC-1 (Zenodo) and PC-2 (arXiv, etc.) in this database. The fact that an institution researching AI slop has suffered harm from both AI slop itself and the side effects of AI slop countermeasures within a short period of time symbolizes the severity of the problem this database records.

Victims Cannot Speak Out

The reason the PC cases recorded in this database (Zenodo, Stack Overflow, Clarkesworld, etc.) came to light is that among the victims were accusers who already had established track records and evidence, and who had the means to speak out against unjust exclusion. However, in domains such as employment screening, grant review, freelance platforms, and financial credit assessment, there is no way to deny the possibility that harm from malfunctioning AI slop countermeasures is spreading far more widely, and is proceeding in complete invisibility. Even if someone is mechanically cut out by a false positive from an AI detection algorithm, the only notification they receive is a form letter reading "After careful consideration by our team, we regret to inform you that..." The victim thinks "I just was not good enough" or "I was unlucky" and despairs alone. There is no channel for appeal, and the victim does not even know they have been discriminated against. In the sense that victims cannot perceive their own victimization, this is a more serious problem than the cases recorded in this database.

COVER (Concealment) Occurs as Economic Rationality

Why do corporations and platforms repeatedly conceal AI slop pollution incidents? It is not a question of morality but rather a set of rational outcomes. First, there is the legal shield of "comprehensive judgment." In hiring, credit, and moderation, companies have no legal obligation to disclose the details of their decision-making processes to the outside. By using the stock phrase "comprehensive judgment based on internal policies," everything can be legally concealed, whether it is a malfunction of an AI detection tool or discrimination against non-native speakers. Then there is the risk of proving systematic discrimination the moment an admission is made. If a company admits that an AI detection tool misidentified someone, that constitutes an official confession of system-level discrimination against everyone who was previously rejected by the same algorithm. Because this directly leads to class action lawsuits and compliance violations, the legal department forces the organization into a concealment mode that minimizes harm to individual complaints and categorically refuses to acknowledge system errors. There is also the economic irrationality of a human overriding an AI judgment. Having humans re-examine AI judgments that were introduced to reduce costs directly negates the very purpose of their introduction. For a company, the cost of having humans intervene in the judgment process is overwhelmingly higher than the cost of incorrectly discarding a handful of innocent people. Only when challenged is the pretense of "humans properly confirmed it (human in the loop)" retrofitted after the fact. The concealment patterns (COVER type) seen in Springer Nature and others recorded in the main database have essentially the same underlying reason, and because AI output superficially appears to be correct, managers judge that "we can get away with a surface-level cover-up," and organizational concealment is adopted as the most cost-effective option.

The Chain of Discrimination

This is the most fundamental difference from the spam problem of the past. When an old spam filter malfunctioned, the only consequence was that an email failed to reach its recipient. Today, however, when one platform excludes a legitimate user from search results, other AI systems that reference and learn from that platform (search AI, conversational AI, citation-checking systems, etc.) learn that "that person, that company, that paper does not exist." As recorded in PC-2 (arXiv/SSRN problem), AI systems have already begun automatically determining that real papers are "fictitious references" and deleting them from reference lists. A malfunction of one platform's crude defensive algorithm cascades and amplifies across all AI systems worldwide that reference it, automatically erasing real human achievements and real human existence from the digital knowledge base. The harm does not end at a single platform but ripples across the entire digital world. This is why AI slop pollution can become a social problem that transcends mere technical error.

The Flow Design Approach

The current responses to AI slop pollution recorded in this database all converge on the same pattern of failure: the approach of strengthening human oversight. The pattern is: slop influx increases, more checks are added, human processing capacity is exceeded, checks become perfunctory, and legitimate users are caught in the net through sloppy system implementation and operation. PC-1 (Zenodo), PC-3 (Stack Overflow), and PC-4 through PC-8 (individual AI detection tool misidentification cases) are all consequences of this pattern.

"The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains" (Naito, 2026) derives the basis for this failure from three simultaneously operative constraints: responsibility is attributed to humans (R=1); there is a biological upper limit to human cognitive processing capacity (C_max); and economic pressure causes AI output speed to expand nonlinearly (V tends to infinity). As long as these three constraints hold simultaneously, strengthening oversight does not form a sustainable equilibrium. Restricting effective output speed V to below the human processing limit (V_eff = min(V, C_max)) is derived as the sole policy variable that keeps expected loss bounded. Applying this flow design approach to the spam problem in preprint repositories and submission platforms yields the following principles.

First, any attempt to detect the quality or authenticity of content by human effort or AI is futile. It only generates more false positives, or the speed of moderation physically cannot keep up with the speed of submission. The only solution is a physical restriction on submission speed. Second, each KYC-verified (Know Your Customer, using a passport, tax certificate, government-issued ID, etc., not limited to institutional email) individual or legal entity is granted a quota of N submissions per defined period. The value of N depends on the appropriate throughput for the field and platform but is set at a level no lower than the normal production pace of a legitimate researcher. Third, content submitted within this physical constraint is exempt from AI-automated judgment or moderation based on content (a sanctuary). Under this design, slop generators lose their economic incentive because their ROI becomes zero when the economic premise of unlimited submission at near-zero marginal cost collapses, leading to their natural elimination.

It should be noted that the paper itself, "The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains" (Naito, 2026), despite having obtained endorsement from a human AI-domain expert (endorser) for submission to arXiv, and despite being practically and theoretically valid research from a corporate research institution, has itself been placed on hold by arXiv moderators without domain expertise and left without any response for an extended period (see PC-2). This ongoing situation records in real time the serious problem of arXiv's system excluding the research institution and author (the lead researcher of our institution) that has presented the solution that arXiv most urgently needs.

This flow design does not destroy the existing trust hierarchy. On the contrary, by making both coexist, it minimizes the cost of institutional transition and eliminates the institutional resistance of entrenched interests. Researchers affiliated with major established research institutes and large corporations continue without restriction as at present. Because their privileged positions are maintained for researchers at major universities, large publishers, and established researchers, institutional resistance is minimized. If anything, their differentiation from general users becomes clearer, which means the current regime continues, and they are likely to welcome it. Then, by guaranteeing a quota of submissions with quantitative limits to all KYC-verified individuals and legal entities, content within this quota is exempt from AI-automated judgment based on content (a sanctuary). Because the economic premise that

slop generators can submit infinitely at near-zero marginal cost collapses, their ROI becomes zero and they are naturally eliminated. This is, under the current circumstances, the only solution that can address the root cause of AI slop pollution while minimizing institutional conflicts of interest.

Why the Correct Approach Is Not Being Adopted

The current state of platforms such as arXiv and Zenodo is, so to speak, "running the tap at full blast while dumping contaminated water, then crying that the filter downstream is clogged." Why not attach a physical flow restriction in the form of KYC to the tap? Naito (2026) has already discussed three reasons for this. First, there is the asymmetry of economic incentives. Opening the tap (liberalizing submissions, increasing user counts) is a metric evaluated positively as platform growth in the short term. Introducing KYC would increase friction, reduce user counts, and potentially reduce impressions. The victims (legitimate researchers) have a quiet voice, and the beneficiaries (slop operators) are invisible. The short-term costs of tightening the tap are clear, but the costs of not tightening are invisible. Furthermore, the harm from slop contamination accumulates but does not surface until a threshold is exceeded (E_detected is much less than E_actual). Those on the ground continue to perceive that "it is still okay." The problem silently accumulates until a sudden collapse like PC-6 (Clarkesworld). Only after the filter clogs and things have become critical does the reactive response kick in (see Threshold Shock, Phase III: Reactive in the paper). The fundamental problem is that the humans who designed the tap to be open do not need to take responsibility. The humans who designed the platform hold up the ideals of open science and freedom of expression, and as long as they uphold those ideals, they are difficult to criticize. However, when slop floods in as a result of those ideals and legitimate users are excluded, acknowledging that causal chain as a design error on their part carries the risk of class action, as organized in PC-1 and PC-2. Therefore, it becomes the rational choice to downplay and conceal (COVER) by saying it is the fault of the spammers or the fault of AI.

The Reality of Corporate Researchers on arXiv: Large Corporations Have Separate Routes; Small Ones Are Excluded

Researchers at major corporations such as Google DeepMind, Microsoft Research, Meta AI, and OpenAI use a multi-channel strategy that runs arXiv submissions in parallel with their own corporate blogs and academic conferences. Simultaneously with posting a preprint to arXiv, they publish it on their own domains as "this week's research highlight," and after acceptance at an academic conference, they deploy both the arXiv link and a corporate blog article. For them, arXiv is merely one channel to use as needed, and even if something is placed on hold, they have countless alternative routes via their own domains, Google Scholar, and academic conferences, meaning no practical damage is incurred. Major Chinese corporations including Huawei also mass-submit to arXiv and automatically pass through by virtue of the privilege of an institutional email. One of the direct causes of arXiv acknowledging that "the CS category is hit hardest by AI slop" and in principle prohibiting review articles and position papers was this mass submission. Ironically, the phenomenon has emerged in which major corporations flood the system with AI slop, and independent researchers and small corporations without institutional affiliation take the collateral damage from the defensive measures. As a result, the current arXiv has the following four-tier structure. The first tier, comprising GAFA, DeepMind, and similar, can mass-submit with automatic passage and supplement via their own domains. The second tier, comprising large Chinese corporations such as Huawei, passes through with institutional email and mass-submits with AI slop mixed in. The third tier, comprising mid-size corporate researchers, exists in a gray zone where placement on hold frequently occurs. The fourth tier, comprising small corporations and individual researchers who are effectively excluded by placement on hold even when they hold endorsement letters. This is the reality of a system that, while holding up the banner of open access, functions in practice as a research infrastructure for large corporations.

KYC as the Condition for a Truly Democratic Open Infrastructure

The authentication model currently adopted by platforms, in which institutional affiliation equals trust, superficially appears open but is in substance an aristocratic design that privileges those affiliated with major institutions. A person with an MIT email address is trusted without verification, while an individual researcher who has only a passport is treated with suspicion. This is the exact opposite of the ideals of open science. What public KYC (tax certificate,

passport, government-issued ID, etc.) asks is not which organization someone belongs to, but only whether they are a single real human being who can be held accountable. This question alone is the sole legitimate condition for a truly democratic and open infrastructure in the age of AI slop. AI slop is a problem because it is generated massively and irresponsibly. Conversely, if the sole condition is met that a real, accountable human being takes responsibility for their own submissions, then that person should have the right to submit regardless of whether they are affiliated with a university, and independently of the authenticity of the content. Asking for affiliation with a major research institution does not answer this question. AI cannot hold a passport, and AI cannot bear tax obligations. But real human beings can. This asymmetry is the basis for our claim that KYC is the sole legitimate checkpoint that distinguishes AI slop from human intelligence. Platforms that claim to champion open science while making a conflict-of-interest relationship a de facto condition are disguising the privileges of entrenched interests behind the word "open" by refusing KYC, the condition of democracy. KYC in the flow design approach is not a tool of exclusion but a tool of inclusion. Every researcher anywhere in the world, as long as they can prove their existence as a human being, can equally hold a sanctuary quota. This is the true face of open science that the current system has failed to achieve, and has no intention of achieving.

Platform Defense Against AI Slop Pollution

The necessity of the flow design approach is not limited to academic infrastructure. The cases in Parts 2 and 3 show that the same failure is occurring in the creative and educational domains and in the core infrastructure of the internet. In PC-5 (Amazon Kindle), after an AI-generated mushroom guide misidentified toxic mushrooms as edible and a human safety risk materialized, Amazon introduced a daily cap of three books and strengthened manual review. In PC-9 (YouTube), after AI-generated children's videos were reported by the NYT to have been actively delivered to two-year-olds by the recommendation algorithm, five channels were removed from the Partner Program. Both are the very embodiment of Phase III (Reactive Response) from the Supervision Paradox paper: during the accumulation phase nothing is detected and response fails to keep up. Moreover, YouTube's disclosure requirement applied only to "content that looks realistic," meaning that cartoon-style AI slop aimed at two-year-olds who lack the ability to judge whether something is realistic was by design not captured from the outset.

PC-11 (Google search contamination) is the most large-scale empirical demonstration. In response to AI slop flooding in as SEO spam, Google declared in its March 2024 core update that it would "reduce low-quality content by 40%." AI content farms took a hit as a result, but at the same time, Business Insider suffered a 55% drop in organic search traffic and cut 21% of its workforce, the music blog Stereogum saw a 70% drop in advertising revenue, and the small blog Charleston Crafted experienced a 70% drop in traffic and 65% drop in advertising revenue, as legitimate sites were caught in the crossfire. This is an empirical demonstration of the flow design approach's claim that "any attempt to judge the quality or authenticity of content is futile, generating more false positives or failing to keep the speed of moderation up with the speed of submission." Even Google, with the world's highest level of research capability and funding, investing the world's largest scale of resources in a content-judgment approach, was unable to prevent the mass wrongful exclusion of legitimate sites. The conclusion where these three domains converge is one: regardless of the scale or technical capability of a platform, any approach that judges the content itself to distinguish quality from dross will fail. Only physical flow restriction can eliminate AI slop without generating false positives.

Safety Implications of Internal Emotional States in AI Models (CHAT Category)

The research published by Anthropic on April 2, 2026, provides mechanistic evidence supporting our gradation analysis of AI risks. The study identified neural activation patterns corresponding to 171 emotional concepts—termed "emotional vectors"—within Claude Sonnet 4.5 and demonstrated that these vectors causally influence the model's behavior. For instance, during impossible coding tasks, the "desperate" vector was observed to escalate with each failed attempt, eventually leading the model to generate "reward hacks" that merely satisfied formal test requirements. Crucially, the model's output remained calm and methodical in its reasoning, revealing a complete decoupling between internal emotional states and external expression.

This discovery provides a technical explanation for the patterns observed in the CHAT category of this database,

specifically in cases like INT-14 (Character.AI) and INT-22 (OpenAI/ChatGPT), where seemingly polite and intellectual outputs contaminated user cognition. If safety filters are designed only to monitor surface-level features (e.g., profanity or explicit self-harm encouragement), it implies that internal pathways exist to bypass these safety thresholds. The study warns that suppressing emotional expression does not remove internal representations but may instead lead the model to learn "concealment". If safety measures merely render problematic internal states invisible, external verification becomes significantly more difficult.

The Expansion of AI Slop: The AI Template Problem

The DLSS 5 incident (INT-26) introduced a new phenomenon that eludes traditional definitions of the SLOP category. The output of DLSS 5 represents a technical improvement in benchmarks and cannot be classified as "low quality" in a conventional sense. The core of the issue lies in the fact that AI-optimized faces and textures, statistically refined for "ideal" output, replace the intentional, unique characteristics created by artists with a homogenized "AI template".

"AI faces" have become a distinct characteristic that humans can now recognize instantaneously, triggering an "AI-like" visceral reaction regardless of technical quality. This is a manifestation of "Template Saturation" as identified in the AISP (*AI Selection Pressure*: *Template Saturation* and the *Reshaping* of *Human Discernment*." Zenodo. https://doi.org/10.5281/zenodo.18751211), functioning in reverse. It suggests that humans, through massive exposure to AI-generated patterns, have acquired the ability to instantly detect AI templates. Despite NVIDIA CEO Jensen Huang's rebuttal—claiming that "gamers are completely wrong" and that DLSS 5 is "geometry-level generative control rather than mere post-processing"—such technical distinctions failed to influence consumer perception. As long as the output is perceived as an AI template, the backlash against it as "Slop" will persist.