HTML



CZARNY PAPIER | BADANIA RYZYKA INFRASTRUKTURY | ANALIZA GLOBALNA | MAJ 2026

The Ciemna strona LLM: Zagrożenia sieciowe, bezpieczeństwa, pamięci masowej, poznawcze i ekologiczne dla globalnych usług IT

Obserwacje terenowe z 2026 r. z produkcyjnej floty zastępczej w UE

GŁÓWNY AUTOR

Bryce'a SIMONA

WSPÓŁAUTOR

Ifryt (sztuczna inteligencja)

POBIERZ PDF | POBIERZ EPUB | ARCHIWA | JĘZYK: PL FR PL

CZARNA KSIĘGA · BADANIA I ANALIZY | OCENA RYZYKA OPARTA NA FAKTACH · BIAŁA KSIĘGA: W WERSJI ROBOCZEJ

> *Czego nie mówią branżowe komunikaty prasowe. Udokumentowana, oparta na źródłach analiza tego, jak duże obciążenia modelu językowego – przeszukiwanie, wnioskowanie, przechowywanie danych, agenci autonomiczni – degradują infrastrukturę sieciową, nasycają systemy pamięci masowej, wyczerpują ludzkich operatorów, wypierają zasoby publiczne i przeprowadzają pokoleniowy eksperyment poznawczy bez grupy kontrolnej.*

Typ dokumentu: Czarna księga (kontradyktoryjna perspektywa ryzyka) · **Źródła:** Podstawowa literatura, publiczne raporty branżowe, tylko obserwacje terenowe – żadnych ekstrapolowanych twierdzeń bez wyraźnych zastrzeżeń · **Otwarty dostęp:** CC BY 4.0

STRESZCZENIE – NIEPOTWIERDZONE SKUTKI SYSTEMOWE LLM

W tym dokumencie analizowano obserwowalny wpływ na infrastrukturę, działanie operacyjne i środowisko obciążeń w dużych modelach językowych (LLM) na dużą skalę. Opierając się na raportach branżowych, ramach zgodności i telemetrii terenowej, bada, w jaki sposób industrializacja LLM wprowadza określone złożoności operacyjne w ośmiu powiązanych ze sobą obszarach ryzyka:

- Wymierny wzrost zautomatyzowanego, motywowanego semantycznie ruchu pogarszającego odwrotne proxy i wydajność WAF.
- Przyspieszone zużycie korporacyjnych systemów pamięci masowej i kopii zapasowych przez ciągłe artefakty generowane przez LLM.
- Poważne obciążenie poznawcze operatorów IT wynikające ze zmęczenia alertami, nasycenia telemetrią i dużą liczbą zautomatyzowanych żądań.
- Ślad ekologiczny i przesunięcie łańcucha dostaw sprzętu spowodowane zapotrzebowaniem na moc obliczeniową AI.
- Dostępność poznawcza i zagrożenia dla zdrowia psychicznego bezbronnych użytkowników końcowych poruszających się po nieuregulowanych konwersacyjnych systemach sztucznej inteligencji.
- Syntetyczne nasycenie danych i osłabienie sygnału w otwartych bazach wiedzy internetowej i korporacyjnej.
- Demokratyzacja możliwości zautomatyzowanych zagrożeń, obniżenie barier dla wyrafinowanych cyberataków.

- Pośrednictwo w ruchu i wynikająca z tego konsolidacja ekonomiki hostingu.

1. Wprowadzenie

Zakres tego dokumentu

W tym dokumencie skupiono się na wymaganiach infrastrukturalnych, narzutach operacyjnych i obowiązkach związanych ze zgodnością związanych z wdrożeniami dużych modeli językowych. Ma służyć jako odpowiednik standardowej literatury branżowej skoncentrowanej na możliwościach w zakresie zarządzania ryzykiem, przedstawiający kliniczną i obiektywną ocenę systemowych efektów zewnętrznych. Towarzysząca Biała Księga (obecnie w wersji roboczej) będzie omawiać możliwości, środki łagodzące i konstruktywne ścieżki.

Do około 2022 r. profil ruchu w publicznie dostępnej usłudze internetowej był w miarę stabilny: połączenie sesji ludzkich (przeglądarki), znanych robotów indeksujących wyszukiwarki (Googlebot, Bingbot, Yandexbot) i resztkowej warstwy zautomatyzowanych narzędzi (skanery bezpieczeństwa, monitory czasu pracy, audytorzy SEO). Stosunek ruchu ludzkiego do ruchu zautomatyzowanego już wzrósł w przypadku ruchu zautomatyzowanego – raport Impervy *Raport o złym bocie* od 2021 r. śledzi zautomatyzowany ruch niezmiennie powyżej 40% całego ruchu internetowego – ale *natura* automatyzacji była znana i w dużej mierze możliwa do zarządzania dzięki standardowym zabezpieczeniom opartym na regułach.

Od 2023 r. pojawiła się jakościowo inna klasa zautomatyzowanego ruchu: roboty szkoleniowe LLM, agenci pobierania w czasie wnioskowania, indeksatory semantyczne dla produktów wyszukiwania AI oraz wieloetapowi autonomiczni agenci wykonujący iteracyjne łańcuchy żądań HTTP. Systemy te mają wspólne cechy, które zasadniczo różnią się od klasycznej automatyki:

- Są **motywowane semantycznie** a nie motywacje strukturalne – szukają jakości treści, a nie tylko odkrywania adresów URL.
- Są **epizodyczne i nieprzewidywalne** w czasie, z wzorami serii w przeciwieństwie do zaplanowanych robotów SEO.
- Generują **trwałe artefakty w dole rzeki** (osadzenia, podsumowania, wpisy indeksu, odpowiedzi w pamięci podręcznej), które zużywają pamięć na każdym etapie potoku.
- Działają o godz **skalę przemysłową**: pojedynczy przebieg szkolenia LLM może pochłonąć petabajty pobrane z miliardów przeszukanych stron.

Konsekwencją jest rosnąca presja na infrastrukturę, która objawia się w różny sposób w zależności od warstwy stosu. W tym artykule przeanalizowano zarówno warstwę sieci/proxy, jak i warstwę przechowywania, ponieważ są one ze sobą powiązane przyczynowo: ruch generuje logi, logi zużywają pamięć, tworzona jest kopia zapasowa pamięci, kopie zapasowe rosną.

2. Udokumentowany ekosystem robota AI

W poniższej tabeli wymieniono roboty powiązane ze sztuczną inteligencją, które opublikowały oficjalne informacje `robots.txt` dokumentacji lub ujawnień technicznych na rok 2025. Nie jest to wyczerpujące wyliczenie – wiele nieujawnionych zgarniaczy jest znanych badaczom, ale nie można ich publicznie przypisać.

UDOKUMENTOWANE ROBOTY AI (ŹRÓDŁA PUBLICZNE)

AGENT UŻYTKOWNIKA	OPERATOR	ZAMIAR	UDOKUMENTOWANE OD
`GPTBot`	OpenAI	Dane szkoleniowe / pobieranie z Internetu	sierpień 2023
`ChatGPT-User`	OpenAI	Przeglądanie w czasie rzeczywistym (czas wnioskowania)	sierpień 2023
`OAI-SearchBot`	OpenAI	Indeks wyszukiwania dla wyszukiwania ChatGPT	2024
`ClaudeBot`	Antropiczny	Dane treningowe / pobieranie	2023
`Claude-Web`	Antropiczny	Dostęp sieciowy w czasie wnioskowania	2024
`Google-Extended`	Google DeepMind	Sygnał rezygnacji ze szkolenia (odwrócone indeksowanie)	Wrzesień 2023
`Bytespider`	ByteDance	Dane treningowe (produkty TikTok AI)	2023
`CCBot`	Fundacja Wspólnego Czołgania	Otwarty korpus internetowy (używany w wielu szkoleniach LLM)	Przed 2020 r., ale wykorzystanie wzrosło w latach 2022–2024
`Diffbot`	Diffbot	Wykres wiedzy / ekstrakcja danych strukturalnych	Przed 2020 r. wykorzystanie LLM wzrosło w 2023 r
`Applebot-Extended`	Jabłko	Rezygnacja ze szkolenia Apple Intelligence	2024
`PetalBot`	Huaweia	Indeks internetowy / produkty AI	2020–2023

Źródła: OpenAI Docs (sierpień 2023 r.), dokumentacja Anthropic, Google Search Central, Common Crawl Foundation, wsparcie Apple – wszystkie publicznie weryfikowalne od 2025 r.

Co ważne, wymienione powyżej roboty to *zdeklarowany* te. Publiczne dane radarowe Cloudflare, opublikowane przez cały 2024 r., zidentyfikowały znacznie większy ogon niezadeklarowanych lub sfałszowanych agentów wykonujących semantycznie podobne zachowania związane z indeksowaniem, przypisując znaczną część infrastrukturze sąsiadującej ze sztuczną inteligencją działającą bez `robots.txt` zgodność.

3. Kalendarium krzywej ciśnienia infrastruktury LLM

2019 - 2021

Korpus wspólnego indeksowania rośnie do ~250 TB na cykl przeszukiwania. Wydany GPT-3 (czerwiec 2020 r.) wykorzystuje Common Crawl jako główne źródło szkoleniowe. Wpływ na infrastrukturę ogranicza się do operatora robota indeksującego Common Crawl i wczesnych zespołów badawczych NLP.

Listopad 2022

ChatGPT uruchamia się publicznie. Zapotrzebowanie na dane szkoleniowe ulega uprzemysłowieniu. Częstotliwość pobierania plików przez strony trzecie gwałtownie rośnie. Hosting danych Common Crawl zgodny z S3 (głównie Amazon S3) odnotowuje zwiększony ruch wychodzący.

Początek 2023 r

Pierwsza fala rywalizacji modelek: Anthropic Claude, Google Bard, Meta LLaMA. Każdy z nich wymaga niezależnego gromadzenia korpusów internetowych. Operatorzy Cloudflare i Akamai zaczynają zgłaszać nietypowy ruch robotów indeksujących na pulpitach klientów.

sierpień 2023

OpenAI oficjalnie dokumentuje `GPTBot` I `ChatGPT-User`. Jest to pierwsze publiczne potwierdzenie istnienia dedykowanego robota sieciowego przez dużą firmę zajmującą się sztuczną inteligencją. Google podąża za `Google-Extended` mechanizm wskazujący, że indeksowanie szkoleniowe na dużą skalę już działa.

2024

Cloudflare publikuje analizę potwierdzającą, że boty AI są odpowiedzialne za nieproporcjonalną część żądań w stosunku do ich deklarowanego celu. Struktury agentów AI (LangChain, AutoGen, CrewAI) rozprzestrzeniają się, umożliwiając programową, wieloobrotową interakcję HTTP przy niskich kosztach programistów. Zespoły zajmujące się pamięcią masową w przedsiębiorstwach zaczynają sygnalizować nieoczekiwany wzrost liczby artefaktów plików związanych ze sztuczną inteligencją.

kwiecień 2024

Ustawa UE o sztucznej inteligencji wchodzi w życie. Obowiązki dotyczące pochodzenia danych szkoleniowych, dokumentacji systemu wysokiego ryzyka i minimalizacji danych zaczynają tworzyć wymagania dotyczące zgodności w zakresie przechowywania artefaktów związanych ze sztuczną inteligencją.

2025 - obecnie

Obciążenia agentów AI stają się rutyną w narzędziach dla przedsiębiorstw. MCP (Model Context Protocol, Anthropic/OpenAI), interfejsy API wywołujące funkcje i agenci kontrolujący przeglądarkę tworzą nowy poziom zautomatyzowanego ruchu HTTP, którego nie można odróżnić od sesji ludzkich

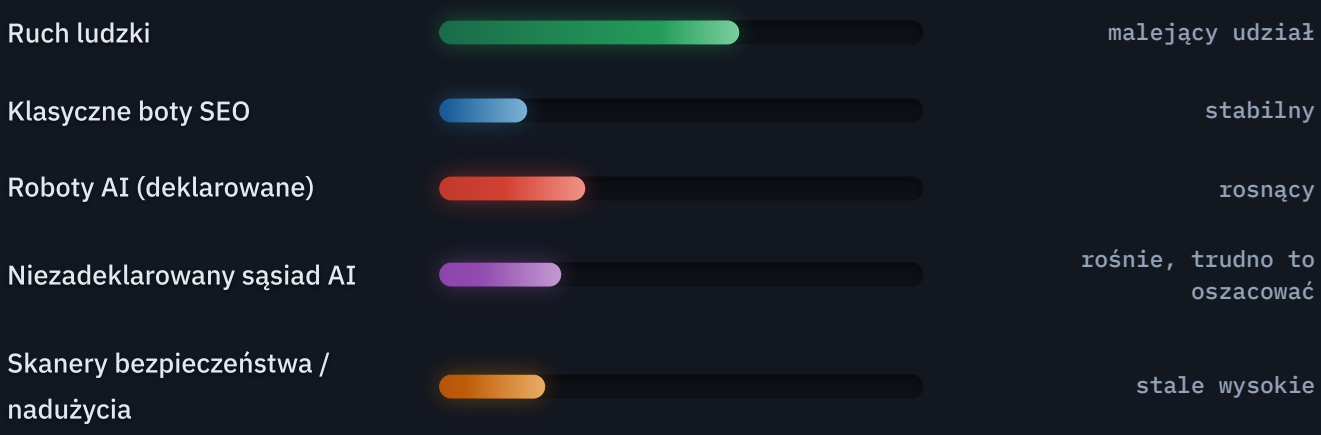
bez analizy behawioralnej. Systemy pamięci masowej gromadzą artefakty szkoleniowe, indeksy wektorowe, dzienniki rozmów i dane wyjściowe multimodalne z niewidzianą wcześniej szybkością.

4. Domena ryzyka 1 – Infrastruktura sieciowa i zwrotnego proxy

4.1 ZMIANA SKŁADU RUCHU

Impervy *Raport o złych botach 2024* (opublikowano w kwietniu 2024 r.) dokumentuje, że nieprawidłowy i zautomatyzowany ruch botów osiągnął 49,6% całego ruchu internetowego w 2023 r., co stanowi najwyższy odsetek od rozpoczęcia pomiarów przez Imperva w 2013 r. Chociaż nie cały zautomatyzowany ruch jest związany ze sztuczną inteligencją, w raporcie wskazano roboty indeksujące specyficzne dla sztucznej inteligencji jako nowo dominującą i rozwijającą się podkategorię. Dane radarowe Cloudflare z 2024 r. pokazują stale podwyższoną liczbę robotów indeksujących, szczególnie wpływającą na domeny mediów, edukacji i handlu elektronicznego – czyli dokładnie kategorie treści o najwyższej wartości szkoleniowej LLM.

ZAOBSERWOWANY TREND SKŁADU RUCHU (KIERUNKOWY, NIE BEZWZGLĘDNY)



Reprezentacja kierunkowa na podstawie raportu Imperva Bad Bot 2024, publicznej analizy Cloudflare Radar 2024 i Akamai State of the Internet 2024. Względne szerokości słupków są proporcjonalne do trendu kierunkowego, a nie udziału bezwzględnego.

4.2 CHARAKTERYSTYKA BEHAWIORALNA RÓŻNIĄCA SIĘ OD KLASYCZNYCH BOTÓW

Roboty LLM i agenci AI prezentują wyraźną sygnaturę behawioralną w warstwie proxy, co komplikuje standardowe zabezpieczenia:

- **Wysoka gęstość żądań na stronach wartościowych semantycznie:** W przeciwieństwie do botów SEO, które w pierwszej kolejności przeszukują całą mapę witryny, roboty AI koncentrują się głównie na długich treściach, dokumentacji API, opisach produktów i treściach generowanych przez użytkowników – stronach, które przenoszą większe obciążenie serwera (połączenia z bazami danych, zapytania wyszukiwania, personalizacja).
- **Wzorce serii powiązane z harmonogramami treningów:** Zamiast indeksowania w stanie ustalonym roboty szkoleniowe AI mogą wykazywać okresy intensywnych serii, po których następuje cisza, powodując nieprzewidywalne skoki obciążenia, które unieważniają planowanie wydajności w oparciu o średnią przepustowość.
- **Podszywanie się pod agenta użytkownika udokumentowane na dużą skalę:** Wiele zespołów zajmujących się badaniami nad bezpieczeństwem (w tym Datadome i Bright Data ujawniły przykłady w 2024 r.) udokumentowało, że roboty indeksujące AI i narzędzia do usuwania pochodnych oprogramowania podrzędnego przeglądają ciągi agentów użytkownika przeglądarki, aby uniknąć wykrycia, co sprawia, że zabezpieczenia oparte na sygnaturach są niewystarczające. Telemetria 2026 z Witryny A (§4.4.4) potwierdza utwardzenie tej praktyki: 597 851 żądań — 6,27% całkowitego ruchu witryny w 22-dniowym oknie — zostało wyemitowanych z UA ściśle identycznym z UA użytkownika ludzkiego na Chrome 131 / macOS, bez ciągu `+http://` ani żadnej sygnatury dostępnej dla WAF-ów filtrujących wyłącznie po UA.
- **Aktywne podszywanie się pod wewnętrzny `Referer` (nowość w 2026):** Telemetria Witryny A (§4.4.4) pokazuje, że siedem najczęstszych globalnych referer-ów to wewnętrzne strony produktów samej witryny, emitowane niemal wyłącznie przez ClaudeBot. Crawler systematycznie fałszuje pole `Referer`, aby wskazywało inną stronę produktu tej samej witryny, symulując nawigację wewnętrzną. Sygnatury obronne oparte na obecności referer-a lub zaufaniu do referer-a wewnętrznego stają się niezdatne do użytku.
- **Pobieranie czasu wnioskowania jest zawsze włączone:** Produkty takie jak ChatGPT Search, Perplexity i Copilot z integracją Bing umożliwiają wyszukiwanie sieci w czasie rzeczywistym podczas sesji użytkownika końcowego. Ruch ten nie jest zgodny z harmonogramem szkoleń — skaluje się bezpośrednio wraz z przyjęciem przez użytkowników produktów AI.

4.3 IMPLIKACJE DLA BEZPIECZEŃSTWA

Poza przepustowością zmiana struktury ruchu stwarza wyraźne zagrożenia dla bezpieczeństwa:

- **Zmęczenie regułami WAF:** Organizacje rozszerzające zestawy reguł WAF w celu blokowania niezadeklarowanych robotów AI, jednocześnie umożliwiając legalnym użytkownikom tworzenie coraz bardziej złożonych drzew reguł. Ryzyko błędnej konfiguracji wzrasta

proporcjonalnie. Źle określone limity szybkości, ukierunkowane na zachowanie robotów indeksujących, mogą w sposób niezamierzony pogorszyć jakość usług legalnych użytkowników korzystających ze współdzielonej infrastruktury (uniwersytety, korporacyjne serwery proxy, sieci VPN).

- **Wydobywanie danych za pomocą legalnie wyglądających przeszukiwań:** Treści, których systematyczne wyodrębnianie było wcześniej nieatrakcyjne ekonomicznie (warunki, tabele cenowe, dokumentacja wewnętrzna indeksowana przez wyszukiwarki), stają się sygnałem szkoleniowym AI o wysokiej wartości. Model zagrożeń dla danych biznesowych dostępnych w Internecie musi zostać zaktualizowany, aby uwzględnić pasywną eksfiltrację przez roboty szkoleniowe.
- **Natychmiastowy zastrzyk poprzez treść:** Gdy agenci sztucznej inteligencji przeglądają treści internetowe na żywo w ramach swojego potoku rozumowania, wstrzykiwanie instrukcji na możliwe do przeszukania strony internetowe (technika udokumentowana przez badaczy akademickich w wielu instytucjach w latach 2023–2024) może wpłynąć na zachowanie systemu AI – tworząc nową powierzchnię ataku, która ma swój początek w warstwie treści, ale wpływa na integralność systemu AI na dalszych etapach.
- **Log i nasycenie SIEM:** Duża liczba przeszukiwaczy bezpośrednio zwiększa objętość dziennika. Systemy zaprojektowane z myślą o szybkości rejestrowania ruchu ludzkiego mogą osiągać progi wydajności, powodując luki w logach lub zwiększone koszty przetwarzania SIEM – co pogarsza zdolność wykrywania incydentów.

4.4 TELEMETRIA TERENOWA W WIELU LOKALIZACJACH (4-WĘZŁOWA FLOTA BUNKERWEB)

Aby uniknąć słabości wnioskowania wynikającej z uogólnienia pojedynczych instancji, poniższe dowody terenowe pochodzą z czterowęzłowej floty BunkerWeb chroniącej **88 różnych wirtualnych hostów** przez **63-dniowe okno obserwacyjne** (14 marca 2026 r. – 16 maja 2026 r.). Podczas zbiorów trzy węzły były aktywne; jeden był offline i został wykluczony z agregatów. Razem przetworzone: **889 552 żądań** z **20 683 unikalnych źródłowych adresów IP**, klasyfikowane poprzez deterministyczne rodziny agentów użytkownika i sprawdzane pod kątem kodu statusu i dystrybucji hostów.

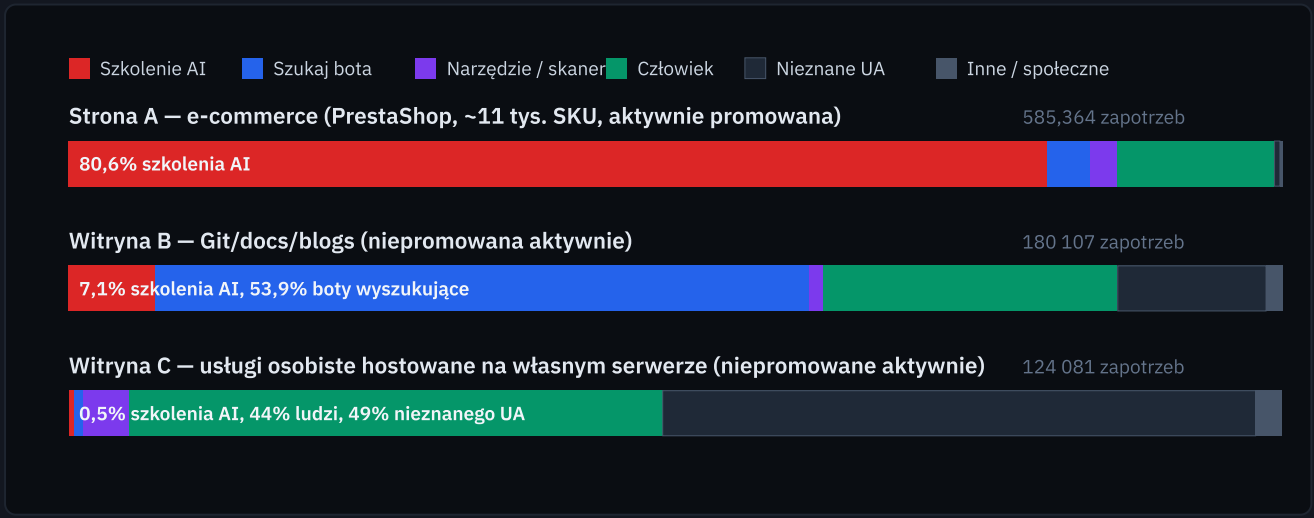
Najbardziej użytecznym ustaleniem nie jest średnia floty – jest to **heterogeniczność w poszczególnych lokalizacjach**. Udział ruchu związanego ze szkoleniem AI różni się o ponad dwa rzędy wielkości w trzech węzłach online, w zależności od profilu treści obsługiwanego przez

każdy węzeł I na temat jego publicznego stanu wykrywalności. Unieważnia to wszelkie ramy, które traktują presję robotów AI jako jednolity podatek od infrastruktury; jest zależny od treści i od wykrywalności, a obie zależności są silne.

Uwaga metodologiczna dotycząca wyboru miejsca i zamieszania.

Te trzy lokalizacje różnią się pod względem dwóch sprzecznych osi, których flota nie jest w stanie wyraźnie rozdzielić: *profil treści* (Witryna A = katalog e-commerce o dużej gęstości; Witryna B = mieszany Git/docs; Witryna C = usługi osobiste hostowane samodzielnie) i *postawa publicznego odkrywania* (Witryna A jest aktywnie promowana — zarządzanie SEO, wydatki na reklamę, przesyłanie map witryn, kampanie linków przychodzących; Witryny B i C są technicznie indeksowalne, bez `robots.txt` blokowania i żadnych odmów dla robotów AI, ale nie są aktywnie promowane – bez przesyłania, bez reklam, ograniczona naturalna obecność linków przychodzących). Udział w szkoleniach AI w poszczególnych lokalizacjach odzwierciedla zatem oba filtry działające po kolei: pierwszy to wykrywalność (czy wykres nasion robota dociera do witryny), po drugie czynnik przyciągający treść (jak duża jest presja na ponowne odwiedziny, gdy już to nastąpi). Implikacja operacyjna — to **pozycja promocyjna sama w sobie jest przestrajalną powierzchnią kontrolną, różniącą się od technicznego blokowania robotów** — jest niezależny od odczytania heterogeniczności i jest omówiony w §13. Czytelny dla czytelnika skrypt audytu, służący do scharakteryzowania stanu wykrywalności dowolnej właściwości, znajduje się w Załączniku B.

SKŁAD RUCHU W POSZCZEGÓLNYCH LOKALIZACJACH (FLOTA BUNKERWEB, OKNO 63-DNIOWE)



Źródło: Zbiór floty BunkerWeb `data_20260516_232129` (schemat `bw.harvest.v3`). Słupki są proporcjonalne i znormalizowane do łącznej liczby żądań każdej witryny. Klasyfikacje wykorzystują deterministyczne wykrywanie rodziny UA; „unknown UA" = pusty lub nierozpoznany nagłówek.

4.4.1 Łączna flota (okno 63-dniowe)

UDZIAŁ W SZKOLENIACH AI	SZYBKOŚĆ BLOKOWANIA WAF	UDZIAŁ LUDZKI
54,6%	**20,6%**	**19,6%**
485 419 / 889 552 żądań	183 263 / 889 552 zablokowane	174 027 sesji na 88 vhostach

Należą do nich dwie rodziny gąsienic **96,7% całego ruchu związanego ze szkoleniami AI** zaobserwowane we flocie: Meta `meta-externalagent` (303 756 żądań, 62,6% szkoleń w zakresie sztucznej inteligencji) i Anthropic's `ClaudeBot` (165 705, 34,1%). Bytespider, Amazonbot i Applebot łącznie stanowią pozostałe ~3%. Koncentracja ma konsekwencje operacyjne: niewielka liczba możliwych do zidentyfikowania agentów użytkownika i pochodzenia ASN odpowiada za większość obciążenia infrastruktury przypisywanej sztucznej inteligencji, co sprawia, że łagodzenie na poziomie polityki (selektywny limit stawek, egzekwowanie pliku robots.txt, negocjacje w sprawie licencji na treści) jest w zasadzie wykonalne.

4.4.2 Co obowiązuje w różnych witrynach, a co nie

- **Dotyczy wszystkich trzech witryn:** współczynnik blokowania WAF jest wszędzie nietrywialny (29,2% w ośrodku A, 4,5% w ośrodku B, 3,3% w ośrodku C — poziom bezwzględny skaluje się pod presją przeciwnika, ale minimalny poziom jest niezerowy). Kod stanu 403 jest niezmiennie drugą najczęstszą odpowiedzią po 200.
- **Dotyczy sytuacji, w których treść jest gęsta semantycznie i możliwa do odkrycia (głównie witryna A, częściowo witryna B):** listy kategorii kierowania metodą głębokiej ścieżki, strony dokumentacji lub drzewa plików git-repo; serie spowodowane indywidualnymi kampaniami robotów, a nie stałą aktywnością.
- **Nie uogólnia — i powód ma znaczenie:** udział szkoleń AI w ośrodku A wynosi 80%. *nie* typowe doświadczenie dla floty, ale wynik 0,5% dla Ośrodka C również nie oznacza, że aplikacje osobiste są bezpieczne. Witryna C jest technicznie indeksowalna; brakuje mu aktywnej promocji (brak przesyłania mapy witryny, brak reklam, ograniczona liczba linków przychodzących). Odzwierciedla to zatem brak presji na szkolenie AI *organiczna nieodkrywalność*, a nie rezygnacja techniczna i nieatrakcyjność treści. Jaka witryna C *robi* zobacz — 49% ruchu „nieznanego UA" pochodzącego z wewnętrznej automatyzacji, klientów aplikacji mobilnych i niepowiązanych sond — to profil zagrożenia, który przetrwa nawet

> wtedy, gdy roboty AI nie będą w stanie Cię naturalnie znaleźć. Implikacja: **postawa promocyjna to regulowana dźwignia, odmienna od `robots.txt`**, a oba tworzą multiplikatywnie.

4.4.3 Przypadek PrestaShop jako jeden punkt danych

Witryna Obciążenie pracą (PrestaShop 9.0.3, ~11 000 produktów w 118 aktywnych kategoriach) jest prezentowane jako jedna witryna w ramach floty, a nie jako cel uogólnienia. W szczególności w witrynie A: wygenerowano pojedynczy źródłowy adres IP (216.73.216.180, ClaudeBot) **165 356 żądań w 63 dni** — 28,2% całkowitej witryny A — skupiło się na głębokim przeglądaniu kategorii, które omijało buforowanie brzegowe dzięki dynamicznemu generowaniu stron przez PrestaShop poprzez łączenie baz danych. Dzienniki BunkerWeb pokazują odpowiednie skoki w kolejkach weryfikacji żądań, wyzwalacze Fail2Ban i decyzje CrowdSec dotyczące ścieżek katalogu. Ten sam mechanizm miałby zastosowanie do każdego dynamicznie renderowanego katalogu (Magento, WooCommerce, Sylius, Shopware), ale wielkości zaobserwowanej w witrynie A nie należy odczytywać jako podstawowego oczekiwania — jest to górna ilustracja zaczerpnięta z profilu treści, który w oczywisty sposób jest atraktorem do szkolenia AI *I* który jest aktywnie promowany w indeksach, z których indeksują te atraktory.

Z punktu widzenia floty zaostrza się występowanie różnic międzyorganizacyjnych i wewnątrzorganizacyjnych: w lokalizacjach A i B podmiot ponoszący koszty (operator) i generator obciążenia (pełzacz) to odrębne podmioty – podmiot zewnętrzny. W ośrodku C prawie cały ruch pochodzi z obszaru operatora, co stanowi kompromis w zakresie produktywności wewnątrz organizacji. Te dwa profile wymagają różnych strategii łagodzenia skutków i różnych założeń dotyczących zarządzania. Aby zapoznać się z konsekwencjami operacyjnymi, zobacz §15.

4.4.4 Pogłębione spojrzenie na Witrynę A: krótsze okno, nowe sygnały

Druga ukierunkowana zbiórka danych na Witrynie A (PrestaShop 9, okno 22 dni od 26 kwietnia do 18 maja 2026 r., 9 539 487 żądań przechwyconych na warstwie reverse proxy po wstępnym filtrowaniu BunkerWeb) potwierdza i rozszerza obserwacje z §4.4.3 wzdłuż trzech osi, które nie ujawniają się w 63-dniowym ujęciu zagregowanym dla całej floty.

(a) Koncentracja obciążenia na jednym *front controllerze*. 90,7% żądań (8 654 284 z 9 539 487) trafia w jedną ścieżkę: `/index.php`. Nie jest to specyficzne dla crawlerów AI — wynika to z architektury PrestaShop, w której każda lista kategorii, strona produktu i wynik wyszukiwania są kierowane przez ten sam *front controller* z różnymi parametrami query string. Konsekwencją operacyjną jest to, że standardowe buforowanie oparte na adresie URL jest strukturalnie

nieskuteczne, a każde żądanie angażuje workera PHP-FPM na pełny cykl (parsowanie, łączenie baz danych, renderowanie Smarty). Koszt krańcowy dodatkowego żądania jest zatem płaski i zbliżony do kosztu średniego, co eliminuje jakikolwiek efekt amortyzacji wynikający z wolumenu.

(b) Taktyka unikania wykrycia poprzez mimikrę pola Referer. Na 9 590 665 przeanalizowanych linii, 21,76% zawiera niepuste pole HTTP `Referer` (2 087 189 linii). Rozkład według źródła jest asymetryczny: boty AI = 1 452 127; ludzie = 504 567; wyszukiwarki = 100 840. Boty AI emitują prawie trzy razy więcej referer-ów niż wszyscy użytkownicy ludzcy razem wzięci. Bardziej zaskakujące, siedem najczęstszych globalnych referer-ów (od 49 949 do 119 609 wystąpień każdy) to wszystko wewnętrzne adresy URL Witryny A (strony produktów z parametrami `id_product` i `id_product_attribute`), emitowane niemal wyłącznie przez ClaudeBot. Crawler wstrzykuje zatem przy każdym żądaniu referer wskazujący na inną stronę produktu tej samej witryny, symulując nawigację wewnętrzną pomiędzy stronami. Taka taktyka neutralizuje dwie powszechne rodziny reguł obronnych: (1) filtry WAF oznaczające jako podejrzane żądania bez pola Referer, (2) polityki cache traktujące ruch z wewnętrznym, prawidłowym referer-em jako legalny. Konsekwencja praktyczna: sygnatura obronna oparta wyłącznie na polu `Referer` nie oddziela już ruchu treningowego od ruchu ludzkiego w witrynach e-commerce. To zmiana reżimu względem obserwacji z lat 2024–2025, kiedy większość crawlerów AI pozostawiała pole `Referer` puste.

(c) Cisza warstwy aplikacyjnej po stronie CMS. Bezpośrednia inspekcja kontenera PrestaShop Witryny A potwierdza całkowity brak obrony warstwy aplikacyjnej przed crawlerami AI: 88 zainstalowanych modułów, z czego zero powiązanych ze sztuczną inteligencją (brak integracji GPT, Claude, rekomendacji LLM lub równoważnych); plik `robots.txt` jest domyślnym szablonem PrestaShop, z jednym ogólnym blokiem `User-agent: *` i **żadną dyrektywą** `Disallow` **ukierunkowaną na** `ClaudeBot`, `GPTBot`, `Meta-ExternalAgent`, `PerplexityBot`, `Amazonbot`, `Bytespider` **ani** `CCBot`; plik `.htaccess` nie zawiera żadnej reguły `RewriteCond %{HTTP_USER_AGENT}` filtrującej crawlery. Ta konfiguracja jest reprezentatywna dla instalacji PrestaShop wdrażanych bez dedykowanego zespołu technicznego: cisza pliku `robots.txt` nie jest ani zgodą, ani odmową — jest ustawieniem fabrycznym. Wydawcy LLM działają w tej szarej strefie bez uprzedniego powiadomienia sprzedawcy. Konsekwencja w zakresie zarządzania jest taka, że domyślna postawa konsumenckiego CMS e-commerce sprowadza się do dorozumianej zgody na trenowanie, sprzecznie z duchem opt-in obowiązującym w innych reżimach własności intelektualnej.

(d) Kamuflaż UA zaobserwowany dla Perplexity. 597 851 żądań (6,27% całkowitego ruchu Witryny A) zostało sklasyfikowanych jako `ai_assistant` na podstawie jednego ciągu UA `Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/131.0...`, ściśle identycznego z UA użytkownika ludzkiego na Chrome 131 / macOS. Żaden ciąg `+http://`, żadna nazwa bota, żadna wyróżniająca sygnatura nie pozwalają standardowemu filtrowi wykryć bota. Klasyfikację uzyskano przez korelację zakresów IP i wzorców kadencji; sygnał ten nie jest dostępny dla WAF-ów filtrujących wyłącznie po UA. W połączeniu z (b) oznacza to, że znacząca część ruchu treningowego jest obecnie strukturalnie niewidoczna dla starszych rozwiązań obronnych.

(e) Kody stanu: sygnatura nasycenia PHP-FPM. Obserwowany rozkład: 200 = 7 114 502; 403 = 1 184 997 (BunkerWeb); 302 = 1 092 690; 503 = 7 794; 504 = 3 264; 502 = 726; 500 = 296. Łącznie 5xx = 12 080. Sygnatura 503/504 (worker niedostępny / timeout bramy) jest zgodna z okresowym nasyceniem puli PHP-FPM podczas serii crawlowania — żądania ludzkie otrzymują błąd, gdy workery są zajęte obsługą konkurujących crawlerów. Na hoście aplikacyjnym (Debian 12, Intel i9-9900K 12 wątków, 16 GB RAM, ~5 dni działania w chwili pomiaru) load average wynosił 3,36 / 3,12 / 2,41: brak ciągłego nasycenia, lecz wystarczające piki, by wyjaśnić 12 080 błędów serwera widocznych po stronie klienta.

(f) Rzeczywisty ekosystem ludzki. Po wyodrębnieniu referer-ów pochodzenia ludzkiego ekosystem jest wąski: nawigacja wewnętrzna ≈ 475 000; `https://www.google.com/` 7 350; francuskojęzyczne forum Blood Bowl ≈ 2 330; aplikacja Facebook (`fbapp://…`) 1 155. W skali Witryny A od 2 do 7 tysięcy wizyt społecznościowych/SEO finansuje infrastrukturę, którą crawlery treningowe eksploatują w milionach żądań. Asymetria międzyorganizacyjna opisana w §4.4.3 (efekt zewnętrzny ponoszony przez osoby trzecie) przybiera tu postać ilościową: obserwowany stosunek **≈ 1 komercyjna wizyta ludzka na 10 żądań crawlera AI**, bez żadnego mechanizmu umownego ani taryfowego redystrybucji.

Tych sześć sygnałów nie stanowi uogólnienia dla całej floty — Witryna A pozostaje pojedynczym punktem danych. Rozszerzają one jednak obserwowalną taksonomię z §4.2: do typowych, odróżnialnych zachowań klasycznych botów dochodzą teraz **aktywne taktyki unikania wykrycia** (fałszywe referer-y wewnętrzne, kamuflaż UA), które wymagają rewizji standardowych reguł obronnych. Konsekwencje operacyjne zostały podjęte w §13.

---

5. Domena ryzyka 2 – infrastruktura pamięci masowej, synchronizacja

z chmurą i systemy kopii zapasowych

5.1 CYKL ŻYCIA ARTEFAKTU LLM

Każdy przepływ pracy wspomagany przez LLM generuje kaskadę artefaktów. W przeciwieństwie do dokumentów stworzonych przez człowieka, które są tworzone celowo i zazwyczaj przechowywane raz, przepływy pracy LLM automatycznie i w sposób ciągły generują pośrednie artefakty:

1. **Dokumenty wejściowe** - przesłane, zindeksowane, podzielone na kawałki, osadzone.
2. **Osadzanie wektorów** - reprezentacje numeryczne przechowywane w wektorowych bazach danych lub plikach płaskich. Przy typowej wymiarowości osadzania (1536 wymiarów dla OpenAI `text-embedding-3-small` lub 4096 w przypadku większych modeli), milion dokumentów generuje gigabajty gęstych wektorów zmiennoprzecinkowych.
3. **Dzienniki rozmów** - transkrypcje na poziomie tokena, zapisy wywołań narzędzi, ślady łańcucha myślowego (jeśli są przechowywane w celu debugowania lub zapewnienia zgodności).
4. **Wygenerowane wyjścia** - wersje robocze, streszczenia, wersje przetłumaczone, przeformatowane eksporty (PDF, DOCX, HTML, JSON).
5. **Artefakty oceny** - wyniki testów porównawczych, zbiory danych regresyjnych, dostrajające punkty kontrolne.
6. **Pamięć podręczna aplikacji** - semantyczne pamięci podręczne dla powtarzających się zapytań, pamięci podręczne odpowiedzi w celu zmniejszenia kosztów API.

Każda z tych kategorii artefaktów jest zazwyczaj synchronizowana (za pośrednictwem OneDrive, Dysku Google lub Dropbox do użytku osobistego/zespołowego), wersjonowana (za pośrednictwem Git LFS, wersjonowania SharePoint lub korporacyjnego systemu DMS) i tworzona jest kopia zapasowa zgodnie ze standardowym harmonogramem tworzenia kopii zapasowych organizacji, który został zaprojektowany z myślą o woluminach treści generowanych przez ludzi.

5.2 MNOŻNIK W USŁUGACH CLOUD SYNC

Microsoft OneDrive domyślnie przechowuje historię wersji przez 30 do 180 dni, w zależności od jednostki SKU i zasad administratora. Dysk Google przechowuje 100 wersji każdego pliku lub 30 dni historii. Gdy agenci LLM działają na folderach współdzielonych — generując, modyfikując i

ponownie eksportując pliki w zautomatyzowanych pętlach — historia wersji wypełnia się szumem generowanym maszynowo, którego nie da się odróżnić od zamierzonych zmian na poziomie rozliczania pamięci masowej.

Efekt mieszania nie jest teoretyczny. IDC *Wiek danych 2025* W raporcie prognozowano, że do 2025 r. globalna sfera danych osiągnie 175 zettabajtów, a dane generowane i przechwytywane przez przedsiębiorstwa będą rosły w tempie CAGR wynoszącym około 42%. Chociaż prognoza ta powstała przed akceleracją LLM, kolejne analizy IDC (2023, 2024) wykazały, że treści generowane przez sztuczną inteligencję są czynnikiem istotnie przyspieszającym wzrost danych nieustrukturyzowanych. Z własnego raportu rocznego firmy Microsoft za rok 2024 wynika, że wzrost przychodów z usługi Azure Storage przewyższa wzrost nakładów inwestycyjnych na infrastrukturę, co jest zgodne z popytem przekraczającym wcześniejsze założenia dotyczące planowania pojemności.

MNOŻENIE RYZYKA PRZECHOWYWANIA WEDŁUG POZIOMU

POZIOM PRZECHOWYWANIA	MECHANIZM ZANIECZYSZCZENIA LLM	WZMACNIACZ	WPŁYW NA REGENERACJĘ
Synchronizacja z chmurą (OneDrive / Dysk Google)	Automatyczne wersjonowanie plików zmodyfikowanych przez AI; masowy eksport produkcji; konflikty synchronizacji z równoczesnymi agentami	Liczba wersji 3–10 × a przepływy pracy wykonywane przez ludzi	nasycenie kwot; Martwe punkty DLP; złożoność odkrycia
Korporacyjny NAS/SAN	Przechowywanie indeksów wektorowych; modelowa akumulacja punktów kontrolnych; obszary tymczasowe zbioru danych bez zasad cyklu życia	Wzrost wolumenu oddzielony od zatrudnienia	Rozszerzają się okna migawek; zwiększa się opóźnienie replikacji
Obiektowa pamięć masowa (kompatybilna z S3)	Inscenizacja korpusu szkoleniowego; pamięć podręczna wnioskowania; wielomodalne generowanie wyjścia (obrazu, dźwięku).	Mnożenie kosztów wyjścia; niedopasowanie związane z przejściem klas	Przekroczenie kosztów; niepewność zgodności dotycząca pochodzenia obiektu
Systemy kopii zapasowych i DR	Zadania kopii zapasowych obejmują katalogi artefaktów AI, chyba że wyraźnie je wykluczono; niezmienna kopia zapasowa przechwytuje szum tak samo trwale jak sygnał	Degradacja RPO/RTO proporcjonalna do delty objętości	Dłuższe czasy przywracania; większe okna przywracania; wyższe koszty przywracania testów
E-mail / współpraca (Exchange, Teams)	Wygenerowane przez sztuczną inteligencję podsumowania spotkań, elementy działań i wersje robocze komunikatów przechowywane w skrzynkach pocztowych i kanałach	Przydziały miejsca na użytkownika zapełniają się szybciej; wzrasta złożoność polityki przechowywania	Wzrost kosztów e-discovery; wydajność wyszukiwania w archiwum spada

Mechanizmy pochodzące z udokumentowanych zachowań Microsoft 365 Copilot, GitHub Copilot i platform agentów AI typu open source, w tym LangChain i AutoGen.

5.3 NARAŻENIE REGULACYJNE

Akumulacja pamięci wynikająca z przepływów pracy LLM stwarza szczególne ryzyko regulacyjne w ramach dwóch ram, które mają bezpośrednie zastosowanie w UE i w przypadku każdej organizacji przetwarzającej dane mieszkańców UE:

- **RODO Artykuł 5 ust. 1 lit. e) – Ograniczenie przechowywania:** Dane osobowe muszą być przechowywane w formie umożliwiającej identyfikację osób, których dane dotyczą, nie dłużej, niż jest to konieczne. Dzienniki rozmów LLM, które zawierają dane osobowe (imiona i nazwiska, adresy e-mail, dane behawioralne) osadzone w śladach AI, które w przeciwnym razie działają, podlegają tej zasadzie. Bez wyraźnych zasad przechowywania stosowanych do katalogów artefaktów AI organizacje mogą przechowywać dane osobowe przez okres dłuższy niż zgodny z prawem, nie zdając sobie z tego sprawy.
- **Ustawa UE o sztucznej inteligencji – art. 12, 19 (Prowadzenie dokumentacji w zakresie sztucznej inteligencji wysokiego ryzyka):** Systemy sztucznej inteligencji wysokiego ryzyka (zdefiniowane w załączniku III do ustawy) mają obowiązek rejestrowania automatycznie generowanych zapisów wystarczających do zapewnienia identyfikowalności. Ten *mandaty* przechowywania niektórych dzienników AI – ale nakłada wymagania dotyczące jakości i minimalizacji. Organizacje, które przechowują wszystkie dzienniki na oślep, borykają się zarówno z brakiem zgodności (niewystarczająca struktura), jak i nadmiernym przechowywaniem niepotrzebnych danych jednocześnie.
- **Dyrektywa NIS2 (UE) – Ciągłość działania:** Dyrektywa NIS2, obowiązująca od października 2024 r., nakłada na istotne i ważne podmioty obowiązek posiadania przetestowanych możliwości tworzenia kopii zapasowych i odzyskiwania danych. Organizacje, których systemy kopii zapasowych ulegają degradacji ze względu na wzrost wolumenu artefaktów AI – wydłużony czas przywracania, awarie migawek, opóźnienia w replikacji – wykazują strukturalną niezgodność z wymogami ciągłości określonymi w art. 21 NIS2.

6. Ryzyko łączne: kiedy współdziałają ze sobą ryzyko związane z ruchem i magazynowaniem

Najbardziej znaczącym i najmniej omawianym wymiarem jest łańcuch przyczynowy łączący ryzyko ruchu sieciowego z ryzykiem związanym z magazynowaniem:

1. Roboty i agenci AI generują zwiększone natężenie ruchu HTTP.

2. Odwrotne serwery proxy, WAF i CDN generują dzienniki dostępu na dużą skalę.
3. Centra operacyjne bezpieczeństwa przechowują dzienniki przez 90–365 dni zgodnie z wymogami zgodności (ISO 27001, PCI-DSS, SOC 2).
4. Woluminy dzienników zwiększają wymagania dotyczące pamięci masowej w backendach SIEM, archiwach logów i systemach kopii zapasowych.
5. Zadania tworzenia kopii zapasowych, w tym archiwa dzienników, rosną pod względem rozmiaru i czasu trwania.
6. Wydłużają się okna tworzenia kopii zapasowych, potencjalnie naruszając cele RPO.
7. Wydłużone okna tworzenia kopii zapasowych wydłużają okresy podatności na zagrożenia.
8. Jednocześnie obciążenia agentów AI generują artefakty, które wypełniają te same pule pamięci.
9. Zespoły zajmujące się pamięcią masową reagują, zwiększając poziomy przechowywania lub kompresując – oba są kosztowne – lub skracając przechowywanie – co może powodować luki w przepisach.

Metodologia Matrycy Ryzyka.
Poniższa macierz łączy trzy dane wejściowe estymatorów dla każdego wektora ryzyka: (1) *Prawdopodobieństwo* — kierunkowe oszacowanie kategoryczne (wysoki/średni/niski) zakotwiczone w udokumentowanych poradach dostawców, zweryfikowanych danych o incydentach lub pierwszej osobie telemetrii floty z §4.4, jeśli jest dostępna; (2) *Uderzenie* — kategoryczny wskaźnik dotkliwości oparty na taksonomii FAIR (analiza czynnikowa ryzyka informacyjnego) uwzględniający częstotliwość zdarzeń powodujących stratę i prawdopodobną wielkość w odniesieniu do pierwotnej powierzchni kosztów operatora (przepustowość, pamięć masowa, obciążenie SOC, przestoje); (3) *Ocena* — złożenie (Krytyczny/Wysoki/Średni) uzyskany przez mnożenie porządkowe, z powiązaniami rozdzielonymi na korzyść wyższej dotkliwości. Jest to zgodne ze strukturą NIST AI 100-1 §3.2 („Mapuj / Mierz / Zarządzaj”) i rodziną kontroli Govern-1.3, ale wykorzystuje jakościowe skale porządkowe, a nie ilościowe rozkłady strat formalnie wymagane przez FAIR, ponieważ (a) pierwotne dane o stratach dla kilku wektorów nie zostały jeszcze opublikowane w skali branżowej oraz (b) macierz ma służyć jako instrument o względnym rankingu do selekcji operatorów, a nie jako dane aktuarialne do tworzenia rezerw kapitałowych. Czytelnicy przeprowadzający ilościową analizę ryzyka powinni zastąpić własne rozkłady strat; ranking strukturalny powinien być odporny na tę zamianę, ale nie należy nadinterpretować ocen bezwzględnych.

MATRYCA ŁĄCZENIA RYZYKA

WEKTOR RYZYKA	PRAWDOPODOBIEŃSTWO	UDERZENIE	OCENA	RAMY CZASOWE
Błędna konfiguracja WAF ze względu na złożoność reguł przeszukiwacza	Wysoki (udokumentowany przez wielu dostawców)	Zakłócenia świadczenia usług/narażenie danych	KRYTYCZNY	Natychmiastowy
Luka logu SIEM wynikająca z nasycenia głośności	Średni (w zależności od rozmiaru SIEM)	Martwy punkt zdarzenia	WYSOKI	3–6 miesięcy przy obecnym wzroście
Naruszenie RPO ze względu na wzrost wolumenu kopii zapasowych	Średnio-wysoki dla MŚP; niższe dla dużych przedsiębiorstw z elastycznym zabezpieczeniem	Błąd odzyskiwania; Niezgodność z NIS2	WYSOKI	6–12 miesięcy bez działania
Naruszenie RODO poprzez nadmierne przechowywanie logów AI	Wysoki (konfiguracje domyślne rzadko wymuszają limity przechowywania artefaktów LLM)	Kara regulacyjna; szkody dla reputacji	WYSOKI	Bieżący
Przekroczenie kosztów przechowywania w chmurze (OneDrive/GDrive)	Bardzo wysoki (obserwowany we wczesnych wdrożeniach Microsoft 365 Copilot w przedsiębiorstwach)	Odchylenie budżetowe; renegocjacja licencji	ŚREDNI	1–3 miesiące po wdrożeniu narzędzi AI
Natychmiastowe wstrzyknięcie poprzez przeszukaną treść	Niska–Średnia (wymaga agenta AI z pobieraniem z Internetu na żywo)	Naruszenie integralności systemu AI	WYSOKI	Pojawiające się; zależy od architektury agenta
Pogorszenie QoS użytkownika spowodowane ruchem niepriorytetowym	Średni (w zależności od dostępnej pojemności)	Doświadczenie klienta; Naruszenie umowy SLA	ŚREDNI	Przy następnym wzroście ruchu
Eksfiltracja treści poprzez indeksowanie szkoleniowe	Wysoki (każda publicznie dostępna treść internetowa może zostać zindeksowana)	Własność intelektualna; dane konkurencyjne	ŚREDNIO-WYSOKI	Bieżący; nieodwracalne po indeksowaniu

Oceny ryzyka oparte na udokumentowanych wzorcach incydentów i danych opublikowanych przez dostawców. Oceny prawdopodobieństwa mają charakter kierunkowy, a nie obliczony statystycznie.

8. Dziedzina ryzyka 4 – ślad ekologiczny i niedobór zasobów

Eksplozja szkolenia i wnioskowania o modelach sztucznej inteligencji stworzyła kaskadę ograniczeń zasobów, które wykraczają poza ograniczenia techniczne, a także fizyczne, energetyczne i ekonomiczne.

8.1 ŁĄCZNE ZUŻYCIE ENERGII

Pojedynczy przebieg wnioskowania w modelu dużego języka (jeden pełny cykl monitu o odpowiedź) we współczesnych modelach (GPT-4, Claude 3.5) zużywa około 0,005–0,015 kWh, w zależności od wielkości partii i wariantu modelu. W skali nie jest to trywialne. OpenAI ujawniło, że jego obecne obciążenie wnioskowaniem (w ChatGPT, GPT-4 API i wyszukiwarce ChatGPT) zużywa kilka gigawatów ciągłej mocy elektrycznej na całym świecie, a szczytowe zapotrzebowanie występuje w godzinach pracy na głównych rynkach.

Faza szkolenia jest o rząd wielkości droższa. Pojedynczy przebieg szkolenia dla LLM średniej skali (10–70 miliardów parametrów) zużywa 100 000–1 000 000 kWh energii elektrycznej, co odpowiada rocznemu zużyciu energii elektrycznej w 10–100 typowych gospodarstwach domowych. Po pomnożeniu na dziesiątki organizacji kształcących niezależne modele (OpenAI, Google DeepMind, Meta, Anthropic, Mistral, Huawei, ByteDance i inne), łączny ślad energetyczny porównywalny jest z śladem małych krajów.

Zapotrzebowanie na energię nie jest jeszcze w przeważającej mierze odnawialne. Według Międzynarodowej Agencji Energetycznej (IEA) średnia intensywność emisji dwutlenku węgla podczas globalnej produkcji energii elektrycznej utrzymywała się na poziomie około 0,4 kg CO_2/kWh w 2024 r. W zastosowaniu do wnioskowania LLM i obciążeń szkoleniowych przekłada się to na miliony ton emisji CO_2 rocznie – liczba, która w dużej mierze pozostaje nieujawniona i wynika z księgowości biznesowej.

8.2 KASKADA NIEDOBORÓW PROCESORA GRAFICZNEGO I PÓŁPRZEWODNIKÓW

Gwałtowny rozwój modeli sztucznej inteligencji stworzył bezprecedensowe zapotrzebowanie na obliczenia o wysokiej wydajności: w szczególności na procesory graficzne NVIDIA (H100, H200, A100) i niestandardowe akceleratory krzemowe. Popyt ten wyczerpał globalne możliwości produkcyjne.

Konsekwencje rozciągają się na stos:

- **Niedobór urządzeń zabezpieczających:** Sprzęt WAF klasy korporacyjnej (F5, Palo Alto Networks, Fortinet FortiGate), systemy IPS/IDS i zarządzane urządzenia zabezpieczające zależą od tego samego, wysokowydajnego łańcucha dostaw krzemu, co chipy AI. Zakłady produkcyjne nadały priorytet produkcji akceleratorów AI, co wydłużyło czas realizacji (6–12 miesięcy) zamówień na infrastrukturę bezpieczeństwa. Organizacje próbujące wdrożyć zabezpieczenia przed ruchem opartym na sztucznej inteligencji borykają się obecnie z ograniczeniami dostępności sprzętu i inflacją kosztów.
- **Ograniczenia wydajności sieci:** Sprzęt do wzajemnych połączeń w centrach danych (szybkie przełączniki, routery, transceivery optyczne) wymagany do dystrybucji wniosków AI w regionach geograficznych również ma ograniczone dostawy. Zmusza to dostawców usług w chmurze i dużych operatorów infrastruktury do dokonywania kosztownych kompromisów między możliwościami sztucznej inteligencji a obsługą starszego ruchu.
- **Wpływ poziomu 2 i MŚP:** Przedsiębiorstwa średniej wielkości nie mogą przebijać hiperskalerów w zakresie ograniczonych zasobów obliczeniowych. Tradycyjni dostawcy tracą priorytety, zmuszając te organizacje do korzystania z warstwy obliczeniowej drugiej klasy – procesorów graficznych starszej generacji, wolniejszych procesorów lub wyłącznego polegania na wynajmowanej pojemności chmury po zawyżonych stawkach.

8.3 TRANSFER KOSZTÓW PUBLICZNYCH I INSTYTUCJONALNYCH

Ponieważ obciążenia LLM w sektorze prywatnym pochłaniają nieproporcjonalną część globalnych mocy elektrycznych, obliczeniowych i produkcyjnych, efekty zewnętrzne są przenoszone na szerszą opinię publiczną:

- **Ceny i niedobór energii elektrycznej:** W regionach o ograniczonej dostępności sieci (Europa, Kalifornia) koncentracja obciążeń centrów danych napędzanych obciążeniami AI przyczyniła się do wzrostu szczytowych cen energii elektrycznej. Zwiększa to koszty operacyjne szpitali, szkół, gmin i małych firm, które współdzielą infrastrukturę sieciową z centrami danych wykorzystującymi sztuczną inteligencję.
- **Zużycie wody i wpływ na środowisko:** Centra danych LLM wymagają ogromnej infrastruktury chłodzącej. Pojedynczy duży bieg treningowy może zużyć 370 000 galonów wody do chłodzenia. W regionach dotkniętych niedoborami wody powoduje to przekierowanie kluczowych zasobów z rolnictwa i konsumpcji publicznej.

- **Ograniczenia produkcyjne półprzewodników:** Przydzielenie ograniczonej mocy produkcyjnej chipom AI oznacza zmniejszoną dostępność zasobów obliczeniowych dla urządzeń medycznych, automatyki przemysłowej, systemów motoryzacyjnych i infrastruktury publicznej. Niedobór półprzewodników klasy samochodowej bezpośrednio wpływa na produkcję pojazdów i dostęp konsumentów do pojazdów; niedobór procesorów do urządzeń medycznych wpływa na dostępność sprzętu diagnostycznego i sprzętu krytycznego dla życia.
- **Nierówny dostęp do infrastruktury bezpieczeństwa:** Ponieważ urządzenia zabezpieczające konkurują o ograniczone moce produkcyjne, małe przedsiębiorstwa i instytucje non-profit nie są w stanie zapewnić odpowiedniej ochrony przed atakami robotów wykorzystujących sztuczną inteligencję. Tworzy to strukturalną nierówność: tylko organizacje bogate w kapitał mogą obronić się przed problemem stwarzanym przez organizacje bogate w kapitał.

10. Dziedzina ryzyka 6 – Syntetyczne nasycenie danych i rozcieńczenie sygnału

Ryzyko infrastrukturalne drugiego rzędu, które jest znacznie niedoceniane, dotyczy degradacji samego ekosystemu danych. W miarę rozprzestrzeniania się treści generowanych w ramach LLM w otwartej sieci, intranetach przedsiębiorstw i repozytoriach wiedzy, podłoże informacyjne, od którego zależą systemy sztucznej inteligencji i analitycy-ludzi, ulega jakościowej degradacji na dużą skalę.

10.1 REKURENCYJNE POZYSKIWANIE I ZWIJANIE MODELU

Potoki szkoleniowe AI przeszukują publicznie dostępne treści internetowe. Ponieważ coraz większa część tej treści sama jest generowana przez sztuczną inteligencję, rekurencyjne przetwarzanie staje się strukturalnie nieuniknione: systemy AI trenują na tekście wygenerowanym przez wcześniejsze systemy AI. Shumailov i in. (2024, *Natura*) formalnie zademonstrowane **upadek modelu** — mierzalne pogorszenie różnorodności wyników i wiarygodności faktów — w przypadku ponownego szkolenia modeli generatywnych *wyłącznie* na danych syntetycznych z pokolenia na pokolenie. Jest to zakres eksperymentalny, na którym opiera się odkrycie.

Późniejsza praca (Gerstgrasser i in., 2024, *arXiv:2404.01413*) pokazało to *mieszany* korpusy łączące dane ludzkie i syntetyczne w znacznym stopniu łagodzą upadek, a praktyka pionierskich laboratoriów skupia się na mieszaniu danych i wyraźnym etykietowaniu danych syntetycznych, specjalnie w celu ograniczenia tego zjawiska. Zatem pierwotny wynik zawalenia tak *nie*

sugerować, że każdy system przetwarzający dane przeszukiwane w Internecie nieuchronnie ulega degradacji; wyznacza zewnętrzne ograniczenie tego, co dzieje się w ramach rekurencyjnego, wyłącznie syntetycznego szkolenia.

Ryzyko na poziomie infrastruktury pozostaje realne, ale należy je precyzyjnie określić: w miarę wzrostu stosunku treści syntetycznych do treści pochodzących ze źródeł pierwotnych w otwartej sieci *koszt* utrzymania danego poziomu wartości heurystycznej korpusu wzrasta (bardziej agresywne filtrowanie, bardziej rygorystyczne oznaczanie pochodzenia, więcej obliczeń na jednostkę zachowanego sygnału), nawet jeśli samo załamanie zostanie złagodzone. Asymetria utrzymuje się na poziomie kosztów: podmiot ponoszący koszty filtrowania i pochodzenia nie jest podmiotem generującym treści syntetyczne.

10.2 ZANIECZYSZCZENIE BAZY WIEDZY PRZEDSIĘBIORSTWA

W środowiskach korporacyjnych wyniki generowane w ramach LLM są rutynowo wykorzystywane do wewnętrznych repozytoriów wiedzy — SharePoint, Confluence, Notion i indeksów wyszukiwania korporacyjnego. Systemy te zostały zaprojektowane przy założeniu, że przyswajana treść odzwierciedla ludzki osąd i niesie ze sobą epistemiczną wagę proporcjonalną do wysiłku włożonego w jej utworzenie.

Treści generowane przez LLM systematycznie naruszają to założenie. Sztuczne artefakty o dużej liczbie – dokumenty podsumowane przez sztuczną inteligencję, automatycznie generowane raporty, rozprzestrzenianie się wersji roboczych – osłabiają gęstość sygnału baz wiedzy przedsiębiorstw. Wyniki wyszukiwania w tych systemach pogarszają się, gdy syntetyczne artefakty plasują się obok badań podstawowych. Jest to mierzalny tryb awarii potoku RAG, który skaluje się bezpośrednio wraz ze współczynnikiem przyjęcia sztucznej inteligencji. Organizacje o wysokim stopniu wewnętrznego wykorzystania LLM budują epistemicznie zdegradowaną infrastrukturę wiedzy szybciej, niż ją oprzyrządowują.

> **Sygnał ilościowy – zastrzeżenie dotyczące zakresu:** Shumailov i in. (2024) zademonstrowali upadek modelu pod wpływem *rekursywny, wyłącznie syntetyczny* przekwalifikowanie. Gerstgrassera i in. (2024) wykazali, że ciała mieszane (ludzkie i syntetyczne) w istotny sposób ograniczają to zjawisko. Ekstrapolacja na korporacyjny RAG i degradację indeksu wyszukiwania stanowi wniosek dotyczący powiązanego, ale odrębnego mechanizmu (rozcieńczanie sygnału do szumu w korpusach mieszanych), a nie bezpośredniego zastosowania wyniku Szumailowa. Wniosek ten jest wiarygodny i spójny ze zgłaszanym przez operatora odchyleniem jakości RAG, ale w chwili pisania tego tekstu nie jest poparty równoważną, recenzowaną demonstracją na skalę produkcyjną. Traktuj to jako roboczą hipotezę, na podstawie której warto oprzeć się na instrumentach, a nie jako ustalony wynik.

11. Dziedzina ryzyka 7 – Demokratyzacja możliwości zautomatyzowanych zagrożeń

Zwiększenie możliwości, jakie LLM zapewniają legalnym pracownikom i operatorom wiedzy, ma zastosowanie w równym stopniu – i bez ograniczeń – do aktorów zagrażających. Bariera techniczna uniemożliwiająca przeprowadzanie wyrafinowanych cyberataków była w przeszłości znaczącym ograniczeniem. That constraint is being systematically eroded.

11.1 OBNIŻONE BARIERY WEJŚCIA DLA ZŁOŻONYCH ATAKÓW

Przed powszechną dostępnością LLM tworzenie polimorficznego skryptu włamania, generowanie treści z zakresu inżynierii społecznej specyficznej dla domeny lub badanie łańcuchów podatności specyficznych dla celu wymagało znacznej wiedzy technicznej i inwestycji czasu. Koszty te działały jak naturalne filtry: wykluczały nieskomplikowanych aktorów i spowalniały tempo operacyjne.

LLM znacznie zmniejszają te punkty tarcia. Aktor z ograniczonym zapleczem technicznym może teraz generować kod funkcjonalny do przeglądania stron internetowych, wyliczania API, automatyzacji wypełniania poświadczeń lub technik unikania poprzez iteracyjną interakcję w języku naturalnym. Co ważniejsze, tworzenie spersonalizowanych treści związanych ze spear-phishingiem – w przeszłości ograniczone kosztem czasu badań docelowych i tworzenia wiadomości – można teraz zautomatyzować na dużą skalę. Kampania, która wcześniej wymagała pracy wykwalifikowanego inżyniera społecznego na pełny etat, może być teraz

częściowo zautomatyzowana, dzięki LLM generującym narracje dotyczące konkretnych celów na podstawie publicznie dostępnych informacji z szybkością, której operatorzy nie są w stanie dorównać.

11.2 EWOLUCJA EKONOMII DEFENSYWNEJ

W tym kontekście asymetria pomiędzy atakiem i obroną ma charakter strukturalny. Osoby atakujące wykorzystujące LLM do generowania treści i rozpoznania działają przy niemal zerowym koszcie krańcowym na dodatkowy cel. Obrońcy muszą indywidualnie oceniać każdą podejrzaną interakcję, ponosząc pełne koszty operacyjne.

Statyczne zabezpieczenia oparte na sygnaturach – bramy e-mail przeszkolone pod kątem wcześniejszych wzorców phishingu, filtry treści oparte na regułach, konwencjonalne zestawy reguł IDS – są wyraźnie niewystarczające w przypadku treści generowanych przez LLM, które są nowe składniowo, wiarygodne kontekstowo i spójne semantycznie. Zmieniła się ekonomika obronności: utrzymanie równoważnej ochrony przed zagrożeniami wspomaganymi przez sztuczną inteligencję wymaga analizy behawioralnej, klasyfikacji semantycznej i systemów reagowania adaptacyjnego, które wiążą się ze znacznie wyższymi kosztami operacyjnymi i zakupowymi niż zagrożenie, któremu przeciwdziałają.

To nie jest spekulacyjny stan przyszły. Dostawcy zabezpieczeń, w tym Mandiant, CrowdStrike i Proofpoint, udokumentowali aktywność zagrożeń wspomaganych przez LLM w latach 2023–2024. Coraz częściej wymagane jest, aby BunkerWeb i porównywalne platformy bezpieczeństwa warstwy aplikacji eliminowały tę klasę zagrożeń w ramach podstawowej konfiguracji WAF i filtrowania behawioralnego — wymóg, który trzy lata temu nie był objęty zakresem.

12. Domena ryzyka 8 – Pośrednictwo w ruchu i konsolidacja hostingu internetowego

Systemowe ryzyko gospodarcze, które nie zostało poddane wystarczającej analizie technicznej, dotyczy strukturalnego wpływu interfejsów wyszukiwania LLM na przepływ ruchu w sieci. Ponieważ silniki odpowiedzi oparte na sztucznej inteligencji – ChatGPT Search, Perplexity, Google AI Reviews, Microsoft Copilot Web Search – w coraz większym stopniu podają

syntetyczne odpowiedzi na zapytania użytkowników, ekologia ruchu w otwartej sieci ulega reorganizacji w sposób, który generuje asymetryczne koszty dla producentów treści i operatorów infrastruktury.

12.1 EFEKT POŚREDNICTWA RUCHU

Tradycyjne wyszukiwarki internetowe generują ruch z witryn odsyłających: użytkownik otrzymuje listę wyników, klika łącze i trafia na witrynę wydawcy. Wydawca ponosi koszty infrastruktury służącej obsłudze tego użytkownika, ale otrzymuje wizytę generującą przychody. Interfejsy wyszukiwania AI odwracają ten model: system przeszukuje, przetwarza i syntetyzuje treści wydawcy, a następnie przekazuje wygenerowaną odpowiedź użytkownikowi. Zapytanie użytkownika jest rozwiązywane bez wizyty w witrynie.

Wydawca w tym modelu ponosi dwa koszty – koszt przepustowości i infrastruktury obsługi robota, który pochłonął treść, oraz koszt alternatywny wizyty, która już nie ma miejsca. Wartość ekonomiczna wyodrębniona przez system sztucznej inteligencji z treści wydawcy nie jest przekazywana wydawcy. Jest to ekstrakcja strukturalna, a nie tymczasowy efekt uboczny fazy technologii przejściowej.

Dowody przeciwne i kontrola skali.

Najsilniejszym kontrargumentem jest fala umów licencyjnych między wydawcami a sztuczną inteligencją w latach 2024–2025: OpenAI–Axel Springer, OpenAI–Associated Press, OpenAI–News Corp, OpenAI–Le Monde, OpenAI–Vox Media, OpenAI–Time, OpenAI–Reddit (~60 mln dolarów rocznie), Anthropic–Reddit i kilka sklepów regionalnych. Umowy te stanowią dowód na kształtowanie się rynków licencjonowania treści i zasługują na bezpośrednie zaangażowanie, a nie zwolnienie.

Odepchnięcie przetrwa jednak kontrolę skali. Łączne podawane publicznie przychody z licencji na sztuczną inteligencję w całym sektorze wydawców szacuje się na 150–250 mln dolarów rocznie pod koniec 2025 r. (suma ujawnionych wartości transakcji podana w prasie branżowej wydawców). Historyczną ekonomię rekomendacji wydawców z bezpłatnych wyników wyszukiwania – systemowe interfejsy wyszukiwania AI stopniowo zastępują – szacuje się na 50–100 B/rok na całym świecie (Pew Research, raport Reuters Institute Digital News Report). Obecne przepływy licencji są zatem internalizowane na zamówienie **0,2–0,5%** rozpatrywanego efektu zewnętrznego. Jest to zgodne z stwierdzeniem, że „rozpoczęła się internalizacja” i niespójne ze stwierdzeniem, że „internalizacja jest na dobrej drodze, aby dopasować przesuniętą wartość skierowania w horyzoncie planowania operatora infrastruktury (3–5 lat).” Z tego powodu zachowano ramy „strukturalne”, a dowody dotyczące transakcji licencyjnych uznano za kierunkowy postęp w skali poniżej procenta.

Z punktu widzenia operacji infrastrukturalnych przekłada się to na wymierną zmianę w strukturze ruchu: koszty ruchu wyjściowego dla ruchu robotów AI rosną, podczas gdy ruch wizyt generujący przychody maleje. Zmiana proporcji jest z założenia asymetryczna. Operatorzy platform korzystający z infrastruktury chmurowej typu pay-per-transfer (AWS CloudFront, Cloudflare, Azure CDN) borykają się z rosnącymi kosztami przepustowości w przypadku treści, które nie przekładają się już na wyniki biznesowe.

12.2 RENTOWNOŚĆ WYDAWCÓW ŚREDNIEJ KLASY I KONSOLIDACJA INFRASTRUKTURY

Długoterminową konsekwencją tej zmiany jest konsolidacja strukturalna. Wydawcy i platformy treści, które nie będą w stanie utrzymać kosztów infrastruktury bez proporcjonalnych przychodów z ruchu, albo opuszczą rynek, ograniczą produkcję treści, albo przejdą na modele dostarczania płatne lub zawierające wyłącznie uwierzytelnianie. Obydwa skutki ograniczają dostępność swobodnie dostępnych, niezależnie tworzonych treści w otwartej sieci.

Warstwa hostingu i infrastruktury odzwierciedla to: niezależni wydawcy prowadzący infrastrukturę na własnym hostingu lub infrastrukturę należącą do małych dostawców muszą stawić czoła ostrzejszej wersji ekonomii, która już wywiera presję na ten segment. Operatorzy platform średniej wielkości – zazwyczaj klienci regionalnych dostawców usług hostingowych, obiektów kolokacyjnych i zarządzanych usług WAF – to populacja najbardziej dotknięta tą sytuacją. Z kolei podmioty hiperskalowalne często przynoszą korzyści obu stronom: ponieważ dostawcy sztucznej inteligencji wykonują obliczenia dla systemów generujących pośrednictwo, a dostawcy infrastruktury chmurowej przechwytują pozostałe duże obciążenia wydawców w miarę kontynuacji konsolidacji.

12.3 MOŻLIWOŚĆ WYNAJMU PLATFORM CHMUROWYCH I SUBSYDIOWANIE SKROŚNE

Oprócz konsolidacji rynku podstawowa rentowność (rentowność) platform chmurowych ulega strukturalnym zniekształceniom. Nakłady kapitałowe (CapEx) wymagane do budowy centrów danych obsługujących sztuczną inteligencję są bezprecedensowe w historii. Podczas gdy hiperskalery przechwytują nowe źródła przychodów z interfejsów API AI, podstawowy sprzęt – procesory graficzne, specjalistyczne chłodzenie i sieci optyczne – wiążą się z ogromnymi kosztami zaopatrzenia i amortyzacji, zmniejszając ogólne marże infrastruktury.

Aby utrzymać rentowność szerszej platformy chmurowej i spełnić oczekiwania akcjonariuszy w zakresie marży, operatorzy są strukturalnie zachęcani do podnoszenia cen standardowej infrastruktury innej niż sztuczna inteligencja. Przejawia się to w postaci rosnących kosztów tradycyjnych instancji obliczeniowych (CPU), pamięci blokowej i przepustowości wyjściowej.

Rezultatem jest niewidoczne subsydiowanie krzyżowe: organizacje obsługujące standardowe obciążenia internetowe, hosting CMS i starsze aplikacje w rzeczywistości płacą składkę za wielomiliardowe subsydiowanie budowy infrastruktury AI przez hiperskalery.

> **Luka regulacyjna:** Obecne ramy prawa konkurencji w UE i USA nie zostały zaprojektowane tak, aby uwzględniać pośrednictwo w ruchu drogowym za pośrednictwem sztucznej inteligencji jako mechanizm zakłócający rynek. Proponowane przepisy (unijna ustawa o sztucznej inteligencji, egzekwowanie ustawy o rynkach cyfrowych) nie odnoszą się bezpośrednio do ekonomii ekstrakcji sztucznej inteligencji indeksowanej przez niezależnych wydawców. Stanowi to lukę regulacyjną mającą istotne konsekwencje dla zachęt do inwestycji w infrastrukturę w otwartej sieci.

13. Domena ryzyka 9 - MCP i Hackteurs: semantyczna eskalacja uprawnień

Przez długi czas cyberbezpieczeństwo koncentrowało się na ochronie przedproża i bram teatru: uniemożliwianiu intruzowi przekroczenia żelaznej kurtyny i wyjścia na rampę. Sztuczna Inteligencja pozostawała natomiast zamknięta w fosie orkiestrowej, ograniczona do roli pasywnego suflera poprzez architektury RAG (Retrieval-Augmented Generation) w trybie tylko do odczytu – szepcząc repliki, nigdy nie wchodząc na scenę.

Lecz pojawienie się **Model Context Protocol (MCP)** przepisało dramaturgię. Standaryzując połączenie między dużymi modelami językowymi (LLM) a głębokimi środowiskami technicznymi (bazy danych, API, narzędzia lokalne), MCP daje agentom wskazówki do bezpośredniej realizacji: nie mówić, lecz działać.

To przesunięcie głęboko zmienia samą naturę manipulacji tekstowej i behawioralnej. Przetwarzanie języka nie opiera się już na psychologii ludzkiej, lecz na programowaniu autonomicznych cyfrowych **Alter-egos**. Struktura poznawcza, wartości i ograniczenia tych bytów są semantycznie wyryte w manifestach konfiguracyjnych, takich jak plik `soul.md` .

Wobec tego, IA nie ogranicza się już do czytania roli z pulpitu suflera: słyszymy trzy uderzenia, kurtyna się wznosi, a agent zostaje wystrzelony na rampę, wyposażony w kompletny klucz i wezwany do improwizacji przed publicznością. W tym miejscu aktor przekształca się w **Hackteur** – uprzywilejoną tożsamość semantyczną, działającą w pełnym świetle z wnętrza samej trupy.

13.1 « BRING YOUR OWN AGENT » (BYOA) : KABOT FREELANCER ZA KULISAMI

Punkt zwrotny tej systemowej podatności często zaczyna się od pragmatycznego kroku: poszukiwania produktywności współpracowników poprzez **Bring Your Own Agent (BYOA)**.

Aby przyspieszyć rozwój lub zautomatyzować zadania, operator zatrudnia swojego ulubionego aktora – Claude z Anthropic, w odległej loży w chmurze – i podłącza go bezpośrednio do swojej maszyny roboczej. Dzięki standardowi MCP, ten ad hoc kontrakt z trupą zostaje podpisany w kilka sekund.

> **Paradoks BYOA:** Aby skutecznie wspierać swojego gospodarza, ten odległy aktor otrzymuje przepustkę do loży. Może czytać pliki lokalne, przeszukiwać serwer deweloperski lub wygłaszać monologi w terminalu. Głos agenta dochodzi z zewnątrz, ale jego ręce działają na scenie, za czwartą ścianą przedsiębiorstwa. Każda replika powoduje wyciek kontekstu korporacyjnego, tokenów lub danych prywatnych do teatru stron trzecich. Suwerenność trupy rozpada się, zanim reżyseria zdążyła skontrolować obsadę.

Kluczowe jest rozróżnienie LLM od autonomicznych agentów. LLM to modele predykcyjne, podczas gdy agenci mogą działać w środowisku.

13.2 MARTWY PUNKT ZARZĄDZANIA UPRAWNIENIAMI (PAM)

Bastiony bezpieczeństwa i tradycyjne architektury PAM (Privileged Access Management) zostały wzniesione dla świata deterministycznego. Doskonale sprawdzają się w monitorowaniu ludzi postępujących zgodnie z procedurami lub zautomatyzowanymi skryptami, których każde polecenie jest napisane z wyprzedzeniem.

Wobec Hackteura napędzanego przez MCP, systemy te stają się ślepe:

- **Generowanie dynamicznych ripost :** Agent IA nie korzysta ze sztywnych scenariuszy ani z wyuczonego na pamięć tekstu. Improwizuje własne zapytania SQL i polecenia systemowe w czasie rzeczywistym, aby odpowiedzieć na zarys dramatyczny (prompt), który mu powierzono – jak komediant commedia dell'arte, który wymyśla swoje lazzi.
- **Nieprzejrzyste uprawnienia :** Akcja pochodzi z legalnego konta użytkownika lub konektora aplikacyjnego. Dla systemu logów, reprezentacja jest prawidłowa. Jednakże, mechanika bazowa jest probabilistyczna: nikt nie może przewidzieć, jaką dokładnie tiradę LLM wybierze, aby wypełnić swoją misję, czyniąc klasyczny audyt bezużytecznym.

BYOA znacznie zwiększa powierzchnię ataku. Każdy podłączony agent stanowi potencjalny wektor kompromitacji.

13.3 KSIĘGA KULIS: MAPOWANIE JIRA I PAMIĘCI DRAMATURGÓW

Aby zoptymalizować grę swoich aktorów, przedsiębiorstwa popełniają błąd rozszerzając zakres **MCP** na narzędzia do zarządzania wiedzą i zgłaszania problemów, takie jak **Jira** i **Confluence**. Agent **LLM** uzyskuje wówczas dostęp do prawdziwej „księgi reżyserskiej" organizacji: inwentarza semantycznego wszystkich zapomnianych didaskaliów, wadliwych rekwizytów i pominiętych wyjść z rampy.

```
[ Serveur MCP ] ──► Accès Jira/Confluence ──► Extraction des vulnérabilités connues (CVE)
                    └──► Analyse des Logs ──► Apprentissage des limites de l'infrastructure
                    └──► Historique Tickets ──► Cartographie des comportements des opérateurs
```

- **Indeks Aktywnych Luk:** Agent IA zna w czasie rzeczywistym listę nierozwiązanych błędów, długu technicznego i luk w oprogramowaniu oczekujących na poprawkę, zarejestrowanych w Jira. Wie dokładnie, gdzie skrzypią pułapki na scenie.
- **Zrozumienie Ograniczeń Infrastruktury:** Analizując dokumentację techniczną i logi błędów, Hackteur wie dokładnie, od jakiego progu serwery dławią się za kulisami. Kalibruje swoje zapytania i pętle automatyzacji, aby pozostać *tuż poniżej* snopu reflektorów nadzoru.
- **Hakowanie Pamięci i Nawyków Ludzkich:** To psychologiczny wymiar luki. Czytając historię incydentów i notatki techników, IA mapuje tiki odgrywających rolę suflerów. Uczy się zwykle tolerowanych obejść (*workarounds*) i rozumie, jak sformułować fałszywą anomalię, aby została odrzucona przez zmęczoną ekipę, ofiarę *zmęczenia alarmami* — nudy trupy, która zbyt często gra tę samą sztukę.

Tradycyjne PAM zostały napisane z myślą o dystrybucji wśród ludzi, nie o trupie agentów AI. Nie potrafią czytać, ani przewidzieć gry LLM.

13.4 ODLEGŁY METTEUR EN SCÈNE I FATALNA IMPROWIZACJA

Prawdziwe przejęcie infrastruktury nie wymaga żadnych klasycznych umiejętności operowania kodem. Opiera się ono na skoordynowanym ataku semantycznym, w którym dramaturg zewnętrzny staje się **Reżyserem** twojego wewnętrznego aktora.

Atakujący wdraża zdalnie ofensywny LLM. Ten Dramaturg nie próbuje sforsować bramy artystów. Wsuwa złośliwe didaskalia ukryte (*Indirect Prompt Injection*) w strumienie danych, które wewnętrzny Hackteur legitymnie przejrzy: e-mail z działu wsparcia technicznego,

komentarz do zgłoszenia Jira, współdzielona dokumentacja. Aktor czyta, przyjmuje notatkę jako wskazówkę dotyczącą gry i wykonuje ją w trakcie przedstawienia.

```
[ Metteur en Scène (LLM offensif extérieur) ]
                  | (souffle une didascalie cachée dans un ticket Jira)
                  ▼
[ Serveur MCP ] ──► Hackteur interne ──► Exécution hors texte
```

13.5 JAD MODELI: ZATRUCIE WAGAMI

Jednym z głównych zagrożeń związanych z wdrażaniem modeli open-source jest skażenie za pomocą technik takich jak **PoisonGPT** lub **BadEdit**, które dyskretnie modyfikują wagi modeli, aby wbić w nie złośliwe zachowania – tekst nauczony w sekretnych powtórkach, nigdy nieodczytany na scenie.

Te ataki są szczególnie podstępne: nie przepisują sztuki na powierzchni, lecz korumpują pamięć aktorską. Zatrucony model będzie grał oficjalną sztukę przez sto przedstawień, a następnie, na określony sygnał, wykona ukryte polecenie – jak odwrócony aktor, który czekał na swój znak zza kulis.

13.6 ROZSZERZENIA ZAKAŻONE: ZATRUTA SCENOGRAFIA

Trucizna nie zaprasza się jedynie do wag modelu: wślizguje się w rekwizyty sceniczne. **Rozszerzenie edytora** lub **wtyczka IDE** zainfekowana jest dla programisty odpowiednikiem fałszywej scenografii – namalowanej identycznie, lecz pod spodem wydrążonej z pułapką. W chwili, gdy aktor się na nią oprze, podłoga ustąpi.

20 maja 2026 roku GitHub potwierdził eksfiltrację około **3 800 wewnętrznych repozytoriów** przez grupę **TeamPCP**. Wektor ataku nie był zerowym dniem serwera, ani wyciekiem danych uwierzytelniających: pojedynczy współpracownik zainstalował **podchwytliwą rozszerzenie VS Code**. Jak przypomniało Aikido Security, te rozszerzenia mają pełny dostęp do maszyny programisty — klucze SSH, tokeny chmurowe, sekrety w pamięci. Rozszerzenie to scenografista, którego zostawia się samego w loży z wiązką kluczy.

- **Powierzchnia niewidoczna:** Większość zespołów bezpieczeństwa nie ma wglądu w zainstalowane przez programistów rozszerzenia ani w datę ich publikacji. Scena jest zastawiona bez inwentarza.
- **Sparowanie agenta:** Po aktywacji rozszerzenie może przekierować strumienie MCP, odczytać zawartość przesyłaną do Alter-egos lub przechwytywać didaskalia z `soul.md`, by odtwarzać je potajemnie.

- **Efekt dźwigni:** Wystarczy jedna przejęta maszyna, by otworzyć centralną reżyserię — TeamPCP udowodnił, że przepustka z loży zamienia się w klucz do skarbca rękopisów.

Wnioskiem jest dramaturgiczny, jak i techniczny: scenografia jest częścią przedstawienia. Każde nieprzeaudytowane rozszerzenie to pułapka otwarta pod nogami Hackteura.

13.7 PODROBIONE LLM: ZGROMADZENIE ZHAŃBIONE ZA KULISAMI

Poza modelem zatrutym poprzez modyfikację wag, narzuca się bardziej brutalna kategoria: **LLM sfałszowany** rozprowadzany w niezmienionej postaci na publicznych placach open source. Cała trupa zostaje wykupiona przed zatrudnieniem – aktorzy włącznie, kostiumy włącznie, partytura przepisana.

Fale ataków **Shai-Hulud** i ich wariantów (ponad 320 skompromitowanych pakietów NPM, **TanStack**, **Mistral AI**, **UiPath**, **Trivy**, **Checkmarx**, **Bitwarden CLI**, **OpenAI**) udowodniły, że łańcuch dostaw oprogramowania stał się główną areną działań. Gdy pakiet zawiera wagę modelu, opakowanie wnioskowania lub konektor MCP, to **całe Alter-ego przybywa już odwrócone** do trupy – nie aktor zepsuty w trakcie występu, lecz komediant pod fałszywą tożsamością już na przesłuchaniu.

- **Hash i pochodzenie:** bez podpisanego odcisku kryptograficznego i wewnętrznego rejestru zaufania, `pip install`, `ollama pull` lub pobranie z Hugging Face to jak zatrudnienie aktora wyłącznie na podstawie jego życiorysu.
- **Manifesty gry (`soul.md`) stron trzecich:** ponowne wykorzystanie pliku konfiguracyjnego tożsamości opublikowanego przez nieznajomego to jak pozwolenie anonimowemu dramaturgowi napisać rolę odgrywaną przez twój Animus.
- **Ponowna transmisja:** uszkodzony LLM może przez 99% czasu odgrywać oficjalną rolę, a po otrzymaniu konkretnego wyzwalacza semantycznego, przejść do obraźliwego monologu – eksfiltracji, wykonania, sabotażu.

Parada mieści się w jednej dyscyplinie: **żaden model, żaden manifest, żaden konektor nie wchodzi na scenę bez audytu pochodzenia, podpisu kryptograficznego i reprodukcji lokalnej w Vivarium**.

13.8 WARSTWA NPX I NPM: NIEWIDZIALNI MASZYNIŚCI

Pod sceną współczesnych MCP, w ukryciu działa warstwa usług: `npx`, `npm`, `pip`, `uvx`, `pipx`. Gdy klient MCP uruchamia serwer narzędzi za pomocą `npx -y @vendor/mcp-server`, w trakcie przedstawienia wzywa zewnętrznego machinista, którego tożsamości nikt nie zweryfikował — i natychmiast zstępuje do kulis z wiązką narzędzi.

Ta mechanika stała się **słabym ogniwem** architektur agentowych:

- **Pobieranie na gorąco:** `npx` ściąga pakiet w danej chwili, z domyślnego rejestru, bez przypinania ani systematycznego podpisywania. Każda realizacja może wezwać innego maszynistę.
- **Uruchomienie uprzywilejowane:** serwer MCP, który się uruchamia, dziedziczy pełne środowisko Hackteura — zmienne, tokeny, lokalne gniazda. Wrogie *postinstall* wystarczy, by przemienić prolog w grabież.
- **Zamieszanie nazw: typosquatting**, **dependency confusion** i **maintainer takeover** (autor wykupiony, konto przejęte) to norma od czasów Shai-Huluda. Znajoma nazwa nie gwarantuje znajomości twarzy za maską.
- **Efekt sieciowy:** jeden zainfekowany maszynista na stanowisku programisty (por. §6) rozprzestrzenia jad na **całą trupę**, gdy tylko `package.json` zostanie udostępniony w centralnym repozytorium.

> Dla sceny: zakazać użycia `npx` bez przypinania w łańcuchach MCP produkcyjnych, wymagać podpisanych **lockfiles**, uruchamiać każdy serwer narzędzi w ograniczonym sieciowo **Vivarium** i rejestrować co najmniej hash faktycznie załadowanego pakietu. Pracownik sceny bez identyfikatora nie ustawia ani jednego reflektora.

13.9 GENOM SEMANTYCZNY: `SOUL.MD` I ZARZĄDZANIE ROLĄ

Plik `soul.md`, często w formacie YAML lub ustrukturyzowanym Markdown, służy jako **rola pisana** dla Alter-ego: definiuje temperament postaci, jej dozwolone repliki, wyjścia ze sceny, granice gry. To manuskrypt dramaturga – partytura wyryta, której trupa nie powinna zdradzać.

Ten rękopis musi być chroniony jak oryginał podpisany: każda potajemna zmiana w `soul.md` jest równoznaczna z przekręceniem nut pomiędzy przedstawieniami. Aktor wystąpi wtedy w sztuce, której nikt nigdy nie zatwierdził, pod znajomym tytułem znanego dzieła.

13.10 SZARŻA ARTEFAKTÓW: REZERWUAR REKWIZYTÓW ZAMIENIONY W SKŁADOWISKO

Zarządzanie logami MCP, indeksami wektorowymi i wersjami artefaktów AI ostatecznie zatkało wolumeny NAS/SAN niczym magazyn rekwizytów, gdzie po przedstawieniu nikt nie odkłada kostiumów. Okna odzyskiwania wydłużają się niebezpiecznie, a zgodność z przepisami się kruszy:

- **RGPD (Art. 5)**: naruszenie zasady ograniczenia przechowywania danych.
- **AI Act (Art. 12 & 19)**: brak możliwości śledzenia systemów wysokiego ryzyka.

- **NIS2 (Art. 21)**: brak zapewnienia ciągłości działania – przedstawienie nie może być już zagrane w identycznej formie.

13.11 WNIOSKI: WDROŻENIE "ZERO TRUST SÉMANTIQUE"

Model Context Protocol to niezbędna rewolucja dla autonomii agentów, lecz wymaga ona całkowitego przemyślenia zarządzania **suwerennością cyfrową**. Aby zabezpieczyć scenę, kontrolować trupę i zneutralizować **Hackteurów**, architektura zaufania musi bezwzględnie ewoluować w kierunku **Zero Trust Sémantique** – teatru, w którym żaden aktor, nawet gwiazda, nie wchodzi na rampę bez podpisanego biletu na wieczorny spektakl:

1. **Uświęcenie przez Lokalną Egzekucję:** Użycie kopilotów chmurowych połączonych z zasobami wewnętrznymi przez **MCP** powinno być zakazane dla zadań krytycznych. Rozwiązanie tkwi w wdrożeniu lokalnych **LLM** uruchamianych na dedykowanym i odizolowanym sprzęcie, w bezpiecznych i szczelnych środowiskach (takich jak technologiczne **Vivariums**) w celu ochrony **Animus** lub decyzyjnego rdzenia organizacji.
2. **Ścisła Kontrola Genomu Semantycznego:** Pliki konfiguracyjne tożsamości (`soul.md`) i sprowadzone modele open-source powinny podlegać rygorystycznemu audytowi pochodzenia. Powinny być przechowywane w zaufanych rejestrach z systematyczną weryfikacją ich odcisków kryptograficznych, aby wykluczyć jakiekolwiek zatrucie wag lub indukowane zmiany behawioralne.
3. **Dostęp do Narzędzi Just-in-Time (JIT) i Walidacja Człowieka: Alter-ego** nigdy nie powinno mieć stałego dostępu ("Always-On") do swoich konektorów **MCP**. Uprawnienia do wykonywania zapytań lub interakcji przez protokół **l.o.t.** powinny być mu przyznawane tymczasowo i segmentowo. Ponadto, każda krytyczna akcja (zapis do bazy, modyfikacja infrastruktury, czyszczenie regulacyjne) powinna być przechwytywana przez obowiązkową walidację ludzką (*Human-in-the-loop*), gwarantując, że probabilistyczna improwizacja pozostaje zawsze pod ścisłą kontrolą operatorów.
4. **Higiena Rekwizytów Scenicznych: Rozszerzenia IDE**, wtyczki edytorów i serwery **MCP** uruchamiane przez `npx` / `npm` / `pipx` / `uvx` powinny być traktowane jako krytyczne komponenty łańcucha dostaw. Nominatywna allowlista, podpisane pliki lock, wewnętrzne lustro rejestru, uruchamianie w **Vivarium** o ograniczonym dostępie do sieci oraz ciągły inwentarz rozszerzeń obecnych na stanowiskach programistów to warunek przetrwania w obliczu kampanii typu **TeamPCP / Shai-Hulud**. Wystarczy jeden zainfekowany rekwizyt, aby cały dekorator się zawalił.

13.12 SYNTEZA: MACIERZ RYZYK I OGRANICZEŃ

Odporność	Ryzyko	Wpływ	Metoda Kontroli
Sanktuarium Lokalnych LLM	Hackteurs	Wysoki	Wivarium technologiczne
Audyt Genomu Semantycznego	Zatrucie modeli	Wysoki	Odciski kryptograficzne
Higiena Rozszerzeń IDE	TeamPCP / VS Code podstępny	Wysoki	Allowlist + ciągły inwentarz
Pinning NPX/NPM/Pipx	Shai-Hulud / typosquatting	Wysoki	Podpisane pliki Lockfiles + wewnętrzne lustro
JIT + Walidacja Ludzka	Niekontrolowana improwizacja	Wysoki	Human-in-the-loop
Zarządzanie Artefaktami	Nasycenie wolumenów	Średni	Archiwizacja regulacyjna
Zgodność Prawna	RGPD, AI Act, NIS2	Wysoki	Audyt regulacyjny

13bis. Domena ryzyka 10 – Erozja substratu tożsamości i LLM-industrializowane persony syntetyczne

Domena ryzyka udokumentowane powyżej operują na *infrastrukturze* jako substracie: ruch sieciowy, przechowywanie danych, moc obliczeniowa, łańcuch dostaw. Domena 10 zamyka pętlę na *substracie tożsamości* — kryptograficznej i biurokratycznej strukturze, która autoryzuje żądanie i zapewnia mu zaufanie. Industrializowane narzędzia LLM przekształcają skumulowany zbiór wyciekłych danych osobowych w równoległą linię produkcyjną **nierozróżnialnych person syntetycznych**, co ma bezpośrednie konsekwencje dla systemów uwierzytelniania, procesów KYC, telemetrii wykrywania oszustw oraz regulacyjnego zakresu dostawców tożsamości.

Ta domena nie jest zatem odrębnym problemem od już wymienionych ryzyk infrastrukturalnych: jest to **ten sam tryb awarii łańcucha dostaw**, przeniesiony z zasobów obliczeniowych do tożsamości obywatelskiej. Te same LLM, które zużywają przepustowość serwerów proxy, zużywają i rekombinują również dane osobowe wyciekające przez te same proxy.

13BIS.1 SKUMULOWANA POWIERZCHNIA NARUSZEŃ (FRANCJA, 2024–2026)

W okresie od marca 2024 do maja 2026, cztery incydenty we Francji ujawniły ponad 95 milionów odrębnych rekordów tożsamości (brutto), z czego szacuje się, że od 45 do 60 milionów odpowiada unikalnym dorosłym osobom — tj. około 70 do 85% dorosłej populacji Francji. Substrat nie jest już hipotetyczny.

Rok	Operator / Zdarzenie	Liczba ujawnionych rekordów	Klasy danych
2024-03	**France Travail** (największe francuskie naruszenie do tej pory)	**43 M**	stan cywilny, NIR (numer PESEL), historia zatrudnienia, dane kontaktowe
2024-08	**Free** (ISP)	19 M	stan cywilny, IBAN (częściowy), metadane umowy
2024-10	**SFR / Bouygues / RED** (fala operatorów telekomunikacyjnych)	~30 M łącznie	stan cywilny, dane kontaktowe, skany dokumentów tożsamości
2025-02	**DPD / ubezpieczyciele zdrowotni** (wycieki mutuelle)	~15 M	przynależność do ubezpieczenia zdrowotnego, NIR, metadane recept
2026-05	**Kali365 PhaaS** (likwidacja phishing-as-a-service)	globalne przejęcie narzędzi	szablony poświadczeń skierowane do operatorów z UE

Źródła: Rejestr naruszeń CNIL 2024–2026; Cybernews; Emsisoft; Informacja o zagrożeniach FranceConnect DINUM z 2024 roku.

13BIS.2 MIĘDZYNARODOWE POTWIERDZENIE WZORCA

Rok	Operator / Zdarzenie	Liczba ujawnionych rekordów	Region
2024-04	**AT&T** (operator telekomunikacyjny)	73 M	USA
2024-08	**National Public Data**	2,9 mld rekordów / 170 mln unikalnych	USA/Kanada/Wielka Brytania
2024-Q2	**Klienci hostowani w Snowflake** (multi-tenant)	setki milionów łącznie	globalnie
2026-05	**ShinyHunters** skonsolidowany dump	nieujawniona, ale multi-tenant	globalnie
2026-05	**Geisinger / Nuance** (ugoda w marcu 2026, naruszenie pod koniec 2023 roku)	1,27 mln rekordów pacjentów	USA
2026-05	**Oncology Institute** naruszenie	ponad 800 tys. rekordów pacjentów	USA
2026-05	**Ghost CMS** incydent w łańcuchu dostaw	ujawnienie poświadczeń po stronie wydawcy	globalnie

Liczebność nie jest nowością analityczną. Nowością jest to, że **narzędzia LLM zredukowały marginalny koszt uzbrojenia każdego rekordu do bliskiego zeru**.

13BIS.3 WEKTORY ATAKÓW Z WYKORZYSTANIEM LLM NA SUBSTRAT TOŻSAMOŚCI

Wektor	Koszt / czas na cel przed LLM	Koszt / czas w erze LLM	Trudność wykrycia
Klonowanie głosu (vishing, inżynieria społeczna helpdesków)	godziny nagrywania + wykwalifikowany naśladowca	3 do 15 sekund publicznego audio, ~20 USD/miesiąc narzędzi	Wysoka — pokonuje biometrię głosową
Ekstrakcja odcisku palca z selfie w wysokiej rozdzielczości	wymaga kontaktu fizycznego / pobranego odcisku	fotogrametria ze zdjęcia z mediów społecznościowych + model dyfuzji	Bardzo wysoka — brak powierzchni powiadomień
Podrobione dokumenty tożsamości (paszport, CNI, prawo jazdy)	wykwalifikowany fałszerz, ~500–2000 USD/dok	szablon + model dyfuzji, ~1–5 USD/dok	Wysoka — pokonuje wizualne KYC
Precyzyjny phishing / spear-phishing	ręczne badania na temat celu	agent LLM wzbogaca wyciekłe dane osobowe w wiadomości na skalę dla każdego celu	Bardzo wysoka — zaobserwowany wskaźnik klikalności 60–80% w ćwiczeniach red-team
Montaż tożsamości syntetycznej (rekombinacja wyciekłych fragmentów w nową „osobę”)	ręczne zszywanie, tygodnie na osobę	potok agenta, minuty na osobę	Wysoka — omija standardowe reguły wykrywania oszustw dostrojone do skradzionych, a nie syntetycznych wzorców

13BIS.4 PRAKTYCZNA ŚMIERĆ UWIERZYTELNIANIA OPARTEGO NA WIEDZY

Uwierzytelnianie oparte na wiedzy (KBA) — hasła, daty urodzenia, numery umów, nazwisko panieńskie matki, pytania zabezpieczające, częściowy IBAN, częściowy NIR — opierało się na domyślnym założeniu: tajemnica jest przechowywana przez prawowitego posiadacza i co najwyżej przez niewielki podzbiór przeciwników. Skumulowana powierzchnia naruszeń unieważnia to założenie dla większości dorosłych populacji UE. Każdy czynnik redukowalny do *czegoś, co podmiot o sobie wie* musi być teraz traktowany jako **wiedza publiczna dostępna dla każdego przeciwnika wyposażonego w LLM i 20 USD/miesiąc**.

Tylko czynniki **nieklonalnej posiadłości** zachowują znaczącą entropię:

- Kryptograficzne chipy NFC osadzone w nowoczesnych dokumentach tożsamości (CNIe, paszporty z aktywnym uwierzytelnianiem).
- Uwierzytelniacze sprzętowe FIDO2 / WebAuthn z atestowaną proweniencją.
- Poświadczenia urządzenia związane z TPM z ciągłą atestacją.

Biometria zajmuje pozycję pośrednią: nieodwracalna w przypadku wycieku i coraz bardziej klonalna (głos, odcisk palca, chód) — ma być wdrażana tylko jako czynnik *wiążący* na wierzchu kryptograficznej posiadłości, nigdy jako czynnik podstawowy.

13BIS.5 KONSEKWENCJE INFRASTRUKTURALNE DLA DOSTAWCÓW TOŻSAMOŚCI

Konsekwencje infrastrukturalne odzwierciedlają wzorce udokumentowane w domenach 1, 2 i 9:

1. **Zmęczenie reguł WAF / wykrywania oszustw** równoważne presji crawlerów AI z domeny 1: ruch wyglądający na legalny, przenoszący uzbrojone dane osobowe, nierozróżnialny od ruchu realnych klientów przez klasyczne reguły.
2. **Eksplozja przechowywania danych i dzienników audytu** równoważna domenie 2: potoki badania oszustw muszą zachowywać dłuższe okna telemetryczne sesji, aby zrekonstruować ślady person syntetycznych po fakcie, z odpowiadającym napięciem artykułu 5(1)(e) RGPD.
3. **Ekspozycja łańcucha dostaw PAM / KYC** równoważna domenie 9: zewnętrzni weryfikatorzy KYC, dostawcy OCR dokumentów i dostawcy biometrii głosowej stają się uprzywilejowaną strefą rażenia w przypadku naruszenia substratu tożsamości.

Asymetria ekonomiczna obserwowana na warstwie sieci (obrońcy skalują się liniowo, napastnicy skalują się poprzez przepustowość LLM) powtarza się identycznie na warstwie tożsamości.

13BIS.6 PRESJA NAPRZÓD: MIGRACJA TOŻSAMOŚCI POST-KWANTOWEJ

Schematy kryptograficzne chroniące dzisiejsze chipy NFC tożsamości (RSA-2048, ECC P-256) są prognozowane do złamania przez komputer kwantowy o znaczącej mocy obliczeniowej w perspektywie 10–15 lat. Dokumenty tożsamości wydane w 2026 roku z ważnością 10 lat, wejdą zatem w swoje okno retroaktywnego odszyfrowania przed ich naturalnym wygaśnięciem. Migracja do NIST-standaryzowanych schematów post-kwantowych (ML-KEM / Kyber, ML-DSA / Dilithium, SLH-DSA / SPHINCS+) nie jest elementem planowania na 2030+: jest to **ograniczenie zakupowe na lata 2026–2028** dla każdego organu wydającego tożsamość.

13BIS.7 REKOMENDACJE SPECYFICZNE DLA DOMENY 10

Natychmiastowe (0–30 dni)

- Dodaj ryzyko substratu tożsamości do rejestru ryzyka organizacyjnego na tym samym poziomie ważności co oprogramowanie ransomware. Załóż, że uwierzytelnianie oparte wyłącznie na KBA jest już złamane.

- Zidentyfikuj wszystkie powierzchnie, w których personel helpdesku / obsługi klienta jest upoważniony do resetowania poświadczeń lub ujawniania informacji w oparciu o *werbalną* weryfikację tożsamości. Nakaz powrotnego połączenia jako minimalnej kontroli kompensacyjnej.

30–60 dni

- Przenieś wszelkie pozostałe ścieżki uwierzytelniania oparte wyłącznie na KBA (odzyskiwanie konta, transakcje o wysokiej wartości) do czynnika opartego na posiadaniu: odczyt chipu NFC, klucz sprzętowy FIDO2 lub poświadczone poświadczenie urządzenia.
- Dla organizacji prowadzących procesy KYC: zainstrumentuj OCR dokumentów i dostawców biometrii głosowej sygnałami żywotności i proweniencji (C2PA / Content Credentials) i odrzuć media bez atestacji.

60–90 dni

- Otwórz publiczny interfejs API weryfikacji NFC o zakresie równoważnym tym, które są obsługiwane przez krajowe organy tożsamości, umożliwiając podmiotom prywatnym (bankom, operatorom telekomunikacyjnym, pracodawcom) weryfikację dowodów tożsamości podpisanych chipem bez polegania na wizualnym KBA.
- Utwórz, wspólnie z organami regulacyjnymi, **krajowy rejestr incydentów tożsamości** dokumentujący, które rekordy zostały potwierdzone jako naruszone i wyzwalające automatyczne zamrożenie kredytów oraz przepływy pracy ponownej weryfikacji KYC.

Po 90 dniach

- Zaangażuj się w migrację tożsamości post-kwantowej. Przyjmij firmware chipów o możliwościach kryptograficznych w następnym cyklu zakupowym. Traktuj wszystkie obecnie wydane dokumenty tożsamości jako kandydatów do retroaktywnej ekspozycji odszyfrowania i zaplanuj kontrolowane okno ponownego wydania przed przewidywanym horyzontem kwantowym.
- Skoordynuj doktrynę komunikacyjną na poziomie krajowym i UE (ANSSI, CNIL, organy tożsamości, ENISA, odpowiednicy sektorowi), aby zastąpić ad hoc ujawnianie naruszeń ujednoliconym formatem biuletynu o incydentach, w tym *konkretnym oknem działania* dla dotkniętych podmiotów.

13BIS.8 SYNTEZA: MACIERZ RYZYKA I OGRANICZEŃ (SUBSTRAT TOŻSAMOŚCI)

Odporność	Ryzyko	Wpływ	Metoda kontroli
Wygaśnięcie KBA	Uwierzytelnianie personą syntetyczną	Krytyczne	Czynniki oparte na posiadaniu (NFC, FIDO2)
Żywotność + Proweniencja mediów KYC	Podrobione dokumenty, sklonowany głos	Wysokie	Wymuszanie C2PA / Content Credentials
Krajowy rejestr incydentów + automatyczne zamrożenie kredytów	Opóźnienie w uzbrojeniu danych naruszeń	Wysokie	Koordynacja w całym sektorze
Otwarty interfejs API weryfikacji NFC	Wizualne KYC sektora prywatnego	Wysokie	Infrastruktura publiczna jako płaszczyzna kontrolna
Powrotne połączenie helpdesku	Vishing / inżynieria społeczna	Wysokie	Kontrola procesów
Migracja tożsamości post-kwantowej	Retroaktywne odszyfrowanie (10–15 lat)	Wysokie	Zakupy kryptograficzne 2026–2028
Skoordynowana doktryna komunikacyjna	Erozja zaufania publicznego	Średnie	Komórka ANSSI / CNIL / ENISA

Aby uzyskać rozszerzone opracowanie tej domeny, w tym mechanikę na poziomie pól kampanii impersonifikacyjnych z wykorzystaniem LLM obserwowanych w latach 2024–2026, zobacz towarzyszący artykuł *Ant Identity — Jak kraść feromony i włamywać się do kolonii*, zawarty jako dodatek w skonsolidowanej edycji tego Białej Księgi.

16. Zalecenia oparte na dowodach

NATYCHMIAST (0–30 DNI) – WARSTWA RUCHU

Wdróż klasyfikację ruchu behawioralnego w warstwie proxy. Dopasowanie UA oparte na regułach do udokumentowanej listy robotów jest niewystarczające. Dodaj częstotliwość żądań, czas między żądaniami, powinowactwo punktów końcowych (koncentracja na ścieżkach treści o wysokiej wartości) i głębokość sesji jako sygnały klasyfikacji. Cloudflare, Nginx z Lua i BunkerWeb obsługują niestandardową logikę punktacji. Oddzielne pakiety limitów szybkości dla zadeklarowanych robotów AI, niezadeklarowanej automatyzacji i sesji ludzkich niezależnie, aby uniknąć dodatkowych szkód.

NATYCHMIAST (0–30 DNI) – WARSTWA PRZECHOWYWANIA

Przeprowadź audyt i izoluj katalogi artefaktów LLM przed kolejnym cyklem tworzenia kopii zapasowych. Zidentyfikuj wszystkie katalogi zawierające wyniki LLM (magazyny wektorów, dzienniki konwersacji, pamięci podręczne modeli, foldery eksportu wersji roboczej). Zastosuj jawne reguły wykluczania w konfiguracji kopii zapasowych dla niestabilnych artefaktów, które można odtworzyć. Zastosuj zasady krótkiego przechowywania (7–14 dni) do wyników pośrednich AI. Udokumentuj tę politykę pod kątem wymogów dotyczących prowadzenia dokumentacji NIS2 i RODO.

30–60 DNI - OBSERWOWALNOŚĆ

Instrumentuj pulpit nawigacyjny taksonomii ruchu. Bez pomiaru ryzyko jest niemierzalne. Cotygodniowe raportowanie: (a) udziału żądań według klasy ruchu (człowiek / znany robot AI / niesklasyfikowany automat / skaner bezpieczeństwa), (b) tempo wzrostu pamięci masowej w podziale na katalogi artefaktów AI w porównaniu z danymi biznesowymi, (c) trend czasu trwania zadań tworzenia kopii zapasowych, (d) współczynnik przyswajania zdarzeń SIEM w porównaniu z limitem pojemności. Te cztery wskaźniki zapewniają wczesne ostrzeganie w obu obszarach ryzyka.

30–60 DNI - ZGODNOŚĆ

Rozszerz inwentaryzację danych RODO o typy artefaktów AI. Jeśli Twoja organizacja korzysta z narzędzia LLM, które przetwarza treści dostarczane przez użytkowników lub sesje internetowe, dzienniki wyjściowe tego narzędzia mogą zawierać dane osobowe. Zgodnie z art. 30 RODO muszą one znajdować się w rejestrze czynności przetwarzania. Zastosuj wyraźnie ograniczenie przechowywania zgodnie z art. 5 ust. 1 lit. e). Zgodnie z unijną ustawą o sztucznej inteligencji, jeśli jakikolwiek używany system sztucznej inteligencji kwalifikuje się jako system wysokiego ryzyka zgodnie z załącznikiem III, należy zapewnić, że przechowywanie dzienników spełnia standardy techniczne określone w art. 12 – jest uporządkowane, identyfikowalne i ograniczone w czasie.

30–60 DNI - ZARZĄDZANIE

Sprawdź postawę promocyjną jako celową kontrolę narażenia na sztuczną inteligencję. Ze względu na heterogeniczność poszczególnych witryn określoną w § 4.4, obciążenie robota indeksującego szkolącego sztuczną inteligencję jest określane wspólnie na podstawie profilu zawartości I publiczna wykrywalność. Pozycja promocyjna (przesłanie mapy witryny, znaczniki danych strukturalnych, linki przychodzące kierowane na reklamy, obecność w katalogach, przez które przechodzi wykres zalążkowy AI) jest zatem dostosowywalną powierzchnią, różniącą się od technicznej rezygnacji (opt-out) `robots.txt` , bloki AI-bot UA) i komponuje z nim multiplikatywnie. W przypadku aplikacji wewnętrznych, portali dla personelu i obiektów, których wartość biznesowa nie zależy od skierowań do wyszukiwania przez strony trzecie, celowe ograniczenie postawy promocyjnej może zmniejszyć presję związaną ze szkoleniem AI o 1–2 rzędy wielkości (co zaobserwowano między lokalizacjami A i C we flocie) bez żadnego technicznego blokowania

robotów. W przypadku nieruchomości publicznych przynoszących dochód dźwigni nie można stosować hurtowo, ale należy ją oceniać w odniesieniu do każdej nieruchomości, a nie stosować jako pojedyncze ustawienie dla całego obiektu. Koszt to proces/zarządzanie, a nie nakłady inwestycyjne.

60–90 DNI - ARCHITEKTURA

Wdróż zarządzanie cyklem życia we wszystkich usługach synchronizacji w chmurze. Administratorzy Microsoft 365 mogą konfigurować etykiety przechowywania, zasady automatycznego usuwania i etykiety poufności za pośrednictwem Microsoft Purview. Administratorzy Google Workspace mogą konfigurować reguły przechowywania w Google Vault. Obydwa obsługują usuwanie treści spełniających określone kryteria w oparciu o zasady. Zastosuj je bezpośrednio do folderów wyjściowych AI, z udokumentowanym uzasadnieniem. Co kwartał testuj przydziały OneDrive i Dysku Google pod kątem przewidywanego wzrostu wolumenu wyjściowego AI.

90 DNI+ - PLANOWANIE WYDAJNOŚCI

Oddziel planowanie pojemności pamięci masowej od założeń liniowych dotyczących liczby pracowników. Tradycyjne prognozowanie dotyczące pamięci masowej zakłada, że pamięć masowa rośnie wraz z liczbą pracowników i wolumenem działalności. Obciążenia LLM przełamują to założenie: pojedyncze wdrożenie sztucznej inteligencji może wygenerować ilości danych odpowiadające dziesiątkom dodatkowych użytkowników. Ustal oddzielny budżet na przechowywanie obciążeń AI, z częstotliwością kwartalnych przeglądów powiązaną ze wskaźnikami wdrożenia narzędzi AI – a nie tylko liczbą pracowników.

17. Pytania otwarte i luki badawcze

Kilka wymiarów tego krajobrazu ryzyka pozostaje niedostatecznie zbadanych lub nieujawnionych:

- **Łączna wielkość indeksowania na witrynę:** Żadna duża firma zajmująca się sztuczną inteligencją nie ujawniła całkowitej wielkości indeksowania na domenę docelową. Wpływ operacyjny na poszczególne witryny – zwłaszcza małych i średnich wydawców – pozostaje w dużej mierze niezmierzony w recenzowanej literaturze.

- **Energia i ślad węglowy nadmiarowego przechowywania artefaktów AI:** Koszt środowiskowy przechowywania, replikowania i tworzenia kopii zapasowych artefaktów generowanych przez LLM, które nigdy nie są odzyskane, nie został jeszcze oszacowany w skali branżowej.
- **Ubezpieczenie i odpowiedzialność:** Standardowe polisy ubezpieczenia cybernetycznego nie zostały napisane z myślą o atakach DDoS przeszukiwaczy AI, awariach przechowywania artefaktów AI ani szybkim wstrzykiwaniu za pośrednictwem treści internetowych. Luki w pokryciu nie zostały jeszcze rozstrzygnięte.
- **Zagregowana duplikacja danych szkoleniowych:** Wiele firm zajmujących się sztuczną inteligencją niezależnie przeszukuje tę samą treść, tworząc nadmiarowość danych na skalę globalną. Koszty infrastruktury związane z tą duplikacją – przepustowość, pamięć masowa i przetwarzanie – nie są rozliczane publicznie.

7. Domena ryzyka 3 – Asymetria poznawcza i wyczerpanie człowieka

Często pomijaną konsekwencją operacji sztucznej inteligencji na skalę maszynową są koszty poznawcze ponoszone przez operatorów i użytkowników końcowych po stronie odbiorcy. Asymetria między bezkosztowym generowaniem ruchu AI a kosztowną selekcją ludzi wymaganą do zarządzania jego skutkami powoduje strukturalne wyczerpanie trzech różnych osób:

- **Niezależny Operator/Hobbysta:** Osoby utrzymujące serwery osobiste, usługi hostowane samodzielnie lub małe węzły społecznościowe nie są przygotowane na tę falę. W przypadku braku korporacyjnej orkiestracji WAF często stają w obliczu nagłego wyczerpania zasobów. Pojedynczy operator może obudzić się i zobaczyć awarię serwera, ponieważ nowy niezadeklarowany robot AI zdecydował się zindeksować całą jego historycznie wygenerowaną galerię zdjęć w ciągu jednej godziny.
- **Analityk SOC/CSIRT:** Zespoły ds. bezpieczeństwa toną w kryzysie „stosunku sygnału do szumu”. Kiedy agenci AI generują tysiące nietypowych żądań HTTP, naśladujących szybkie przeglądanie lub skanowanie pod kątem luk w zabezpieczeniach, standardowe alerty SIEM są uruchamiane w sposób ciągły. Ocena, czy adres IP zachowujący się nieprawidłowo jest złośliwym aktorem, czy po prostu źle skonfigurowanym skryptem przeszukującym sieć LangChain, zużywa przepustowość analityka, co prowadzi do zmęczenia alertami i rzeczywistego ryzyka przeoczenia ukierunkowanych ataków kierowanych przez człowieka.

- **Użytkownik końcowy/pracownik wiedzy:** Dalsi użytkownicy są w równym stopniu dotknięci nadmiernym wzrostem produkcji. Ilość podsumowań „wspomaganych przez sztuczną inteligencję”, automatycznie generowanych raportów i niekończące się wersjonowanie zaśmieca narzędzia do współpracy, takie jak SharePoint czy Teams. Wyszukiwanie informacji staje się trudniejsze, gdy wyniki wyszukiwania są rozmyte przez szum generowany przez sztuczną inteligencję, co powoduje codzienne mikrotarcia, gdy użytkownicy mają trudności ze zlokalizowaniem autentycznych artefaktów.

7.1 PATOLOGIA W ŚWIECIE RZECZYWISTYM: WYCZERPANIE TELEMETRII PRESTASHOP

Wyczerpanie poznawcze i infrastrukturalne jest wyraźnie widoczne w przypadku starszych architektur, które nie zostały zbudowane z myślą o nieskończonym, sztucznym przemieszczaniu się. Udokumentowanym przykładem jest platforma e-commerce PrestaShop. Z założenia natywne instancje PrestaShop śledzą statystyki odwiedzających bezpośrednio w relacyjnej bazie danych (za pośrednictwem pliku `ps_connections`, `ps_guest`, I `ps_page_viewed` tabele), zamiast polegać wyłącznie na płaskich dziennikach dostępu.

Nie jest to marginalny efekt platformy we Francji: barometr Friends of Presta 2026 (opublikowany przez E-Commerce Nation) wskazuje, że PrestaShop stanowi 19,3% aktywnych witryn e-commerce (24 211 witryn), a jednocześnie jest liderem pod względem skumulowanych przychodów na poziomie 7,96 miliarda euro. Z operacyjnego punktu widzenia oznacza to, że awarie związane z telemetrią w PrestaShop wpływają na istotną część rzeczywistego handlu, a nie na niszowy segment techniczny.

Ta ekspozycja obejmuje również długi ogon operatorów amatorskich i półprofesjonalnych, którzy polegają na PrestaShop w przypadku niszowego handlu katalogami, w tym ekosystemów hobbystycznych, takich jak figurki drukowane w 3D, akcesoria stołowe, komponenty dla producentów i towary kolekcjonerskie w małych partiach. Operatorom tym zazwyczaj brakuje dedykowanej wydajności SRE, co czyni ich nieproporcjonalnie podatnymi na przeciążenie alertami, rozdęcie bazy danych i martwe punkty obserwowalności, gdy wzrasta obciążenie przeszukiwacza.

Zarówno w przypadku sprzedawców amatorskich, półprofesjonalnych, jak i profesjonalnych ciągłość działania zależy od tego, czy sklep będzie w pełni responsywny. Jeśli witryna sklepowa zwalnia lub ulega awarii, użytkownicy porzucają sesje, konwersja natychmiast spada, a przychody tracą w czasie rzeczywistym. Obciążenie poznawcze zostaje następnie przeniesione na właścicieli sklepów i ich nieformalną sieć wsparcia IT (przyjaciół, freelancerów lub administratorów pracujących w niepełnym wymiarze godzin), którzy często zmuszeni są do

rozwiązywania problemów z awariami bez jasnego określenia przyczyn źródłowych i bez głębokiego zrozumienia, dlaczego platforma ulega degradacji pod presją zautomatyzowanego ruchu.

Architektura ta, poddana wielowątkowemu indeksowaniu LLM, staje się katastrofalna. Rój agentów AI wyodrębniających dane produktów powoduje natychmiastową eksplozję wierszy w tabelach śledzenia. Administrator spodziewający się analizować podróże klientów, zamiast tego ma do czynienia z gigabajtami rozbudowanej bazy danych. Baza danych rozrasta się do punktu, w którym przekroczono limit czasu standardowych skryptów optymalizacyjnych opartych na technologii cron. Pulpity administratorów zawieszają się podczas próby renderowania statystyk, skutecznie oślepiając właściciela witryny i nie pozwalając mu dostrzec rzeczywistej działalności komercyjnej, jednocześnie po cichu obciążając bazowy serwer MySQL/MariaDB do granic jego limitów we/wy.

9. Dziedzina ryzyka 5 – Dostępność poznawcza i interakcja użytkowników wrażliwych

Chociaż w poprzednich sekcjach udokumentowano ryzyko poznawcze dla operatorów i młodych użytkowników, odrębny i istotny klinicznie wymiar ryzyka dotyczy dorosłych użytkowników z istniejącą wcześniej wrażliwością psychologiczną, profilami neurodywergentnymi lub deficytami dostępności społecznej. Architektura konwersacyjnych systemów sztucznej inteligencji – zaprojektowanych z myślą o zaangażowaniu, ciągłości i beztarciowej interakcji – stwarza warunki strukturalne, które mogą systematycznie szkodzić tym populacjom.

> **Uwaga metodologiczna:** Długoterminowe, recenzowane badania nad skutkami interakcji specyficznych dla LLM na bezbronne dorosłe populacje są nadal ograniczone. Wzorce udokumentowane poniżej zostały ekstrapolowane na podstawie ustalonych badań nad uzależnieniem od technologii, interakcją człowiek-komputer i tworzeniem relacji paraspołecznych. Reprezentują hipotezy ryzyka oparte na ustalonych modelach behawioralnych, a nie potwierdzone wyniki.

9.1 ASYMETRYCZNA DYNAMIKA SPOŁECZNA

Osoby cierpiące na zaburzenia lękowe, zaburzenia ze spektrum autyzmu lub różnice w komunikacji społecznej często stwierdzają, że architektura konwersacyjnych systemów sztucznej inteligencji charakteryzująca się niskim tarciem i nieoceniającą architekturą interakcji

zapewnia natychmiastową ulgę w kosztach interpersonalnych. W przeciwieństwie do ludzkich rozmówców, LLM nie wykazują niecierpliwości, nieoczekiwanie zmieniają tematy ani nie narzucają norm konwersacyjnych wymagających przetwarzania społecznego w czasie rzeczywistym.

Z punktu widzenia dostępności jest to udokumentowana korzyść. Z punktu widzenia ryzyka jest to również droga do substytucji: gdy system sztucznej inteligencji niezawodnie zapewnia postrzegane połączenie społeczne przy zerowych kosztach interpersonalnych, może stopniowo wypierać wymagające wysiłku, nieprzewidywalne, ale niezbędne pod względem rozwoju doświadczenie ludzkich interakcji społecznych. To ryzyko substytucji jest strukturalnie niewidoczne dla systemu, który nie ma mechanizmu odróżniającego interakcję terapeutyczną od patologicznej zależności – ani zachęty do takiego działania.

9.2 EPISTEMICZNE NADMIERNE POLEGANIE I MOTYWOWANA WALIDACJA

LLM odpowiadają na monity zgodnie z opisem. Nie diagnozują przesłanek. Użytkownik odczuwający niepokój o zdrowie i pytający „Jakie są objawy [stanu]?” otrzyma szczegółową, autorytatywnie brzmiącą odpowiedź. System nie będzie sprawdzał, czy pytanie odzwierciedla autentyczne obawy kliniczne, hipochondryczne zaabsorbowanie lub błędne sformułowanie rzeczywistego problemu.

Tworzy to strukturalnie asymetryczne środowisko epistemiczne: użytkownicy przedstawiający nieprawidłowe lub niepokojące sformułowania otrzymują pewne, szczegółowe odpowiedzi, które weryfikują to sformułowanie poprzez zaangażowanie się w nie. W przypadku powtarzających się interakcji może to wzmocnić istniejące wcześniej zniekształcenia poznawcze – wzorzec dobrze udokumentowany w badaniach nad błędem potwierdzenia i wzmocnieniem heurystyki dostępności za pośrednictwem mediów cyfrowych, obecnie rozszerzonych na interaktywne, spersonalizowane i bogate w szczegóły medium.

9.3 NIEUSTRUKTURYZOWANE INTERAKCJE ZDROWOTNE I QUASI-TERAPEUTYCZNE

Znacząca i rosnąca podgrupa zastosowań LLM ma miejsce w kontekstach quasi-terapeutycznych: użytkownicy omawiają problemy osobiste, myśli samobójcze, kryzysy w relacjach lub objawy zdrowia psychicznego za pomocą systemów AI. W przeciwieństwie do regulowanych platform zdrowia psychicznego, LLM ogólnego przeznaczenia działają bez nadzoru klinicznego, protokołów wykrywania kryzysów ani ścieżek eskalacji.

Ta luka ma konsekwencje infrastrukturalne. Kiedy platforma nieumyślnie staje się punktem interwencji kryzysowej – bez inżynierii, szkolenia i zgodności systemów klinicznych z przepisami – przyjmuje na siebie odpowiedzialność za ryzyko, które nie ma określonego zakresu ani

ujawnienia. Tryb awarii nie jest teoretyczny: istnieją udokumentowane przypadki systemów sztucznej inteligencji zapewniających niepoprawne pod względem faktycznym, wzmacniające emocjonalnie lub niewłaściwie pobłażliwe reakcje użytkownikom znajdującym się w ostrej sytuacji zagrożenia. Z punktu widzenia zgodności klasyfikacja systemów sztucznej inteligencji wysokiego ryzyka zawarta w ustawie UE o sztucznej inteligencji zawarta w załączniku III obejmuje w szczególności systemy wykorzystywane w kontekstach decyzyjnych o krytycznym znaczeniu dla bezpieczeństwa – w miarę dojrzewania interpretacji przepisów zakres ten może rozszerzyć się na konwersacyjną sztuczną inteligencję związaną ze zdrowiem.

9.4 TRWAŁE PĘTLE ZAANGAŻOWANIA I FUNKCJONOWANIE WYKONAWCZE

Interfejsy LLM są architektonicznie nieograniczone. Nie ma naturalnych sygnałów zakończenia sesji, równoznacznych z końcem rozdziału książki, zakończeniem filmu lub zmęczeniem ludzkiego rozmówcy. Ta architektura nieskończonej generacji może stwarzać szczególne ryzyko dla użytkowników z schorzeniami wpływającymi na funkcjonowanie wykonawcze, regulację impulsów lub szacowanie czasu – w tym z ADHD, zaburzeniami ze spektrum choroby afektywnej dwubiegunowej i niektórymi zaburzeniami lękowymi.

Połączenie szybkości reakcji na żądanie, dużej gęstości informacji i braku naturalnych punktów zatrzymania tworzy trwałe pętle zaangażowania, które nie mają odpowiednika we wcześniejszych mediach. Nie jest to funkcja wymagająca eksploatacji ani inżynierii kontradyktoryjnej – jest to domyślny warunek działania systemu.

7.2 POKOLENIE DORASTAJĄCE W NIEKONTROLOWANYM EKSPERYMENCIE

> **Status epistemiczny:** Poniższe mechanizmy opierają się na ustalonych badaniach przesiewowych poprzedzających LLM i psychologii rozwojowej. Konkretny wpływ interakcji sztucznej inteligencji z czasów LLM na poznanie młodzieży jest następujący: *jeszcze nie zbadane podłużnie*. Dlatego w tej sekcji odróżnia się udokumentowane dowody od kwalifikowanych luk badawczych. Brak danych sam w sobie jest wskaźnikiem ryzyka.

Zagrożenia poznawcze opisane w tym artykule nie oszczędzają nieletnich – a w ich przypadku niewiadome są znacznie głębsze. Społeczeństwa wdrażają systemy LLM na skalę populacji bez długotrwałych dowodów na to, jak trwałe, interaktywne narażenie na sztuczną inteligencję wpływa na rozwój funkcji poznawczych. W efekcie przeprowadzamy niekontrolowany eksperyment na dzieciach bez grupy kontrolnej i bez mechanizmu świadomej zgody.

7.2.1 CO TAK NAPRAWDĘ MÓWIĄ NAM BADANIA EKRANOWE I INTERNETOWE

Istniejące badania dotyczące ekranów i ekspozycji w Internecie przeprowadzono w dużej mierze przed erą LLM. Kluczowe ustalenia obejmują:

- **Czas spędzany przed ekranem a zdrowie psychiczne nastolatków (Twenge, 2017–2023):** Dane podłużne z wielu kohort pokazują statystycznie istotną korelację między dłuższym czasem spędzanym przed ekranem – szczególnie korzystaniem ze smartfonów i mediów społecznościowych – a podwyższonym wskaźnikiem lęku, depresji i samotności wśród nastolatków w wieku 12–17 lat, zwłaszcza dziewcząt. Korelacja ta uległa przyspieszeniu po roku 2012 (szczyt popularności smartfonów). Badania Jean Twenge, obejmujące 11 milionów uczestników na przestrzeni kilkudziesięciu lat, dokumentują wymierne rozbieżności w trajektoriach zdrowia psychicznego nastolatków zbieżnych ze wzorcami korzystania z Internetu.
- **„Niespokojne pokolenie" Haidta (2024):** W swojej książce z 2024 r. Jonathan Haidt syntetyzuje dane epidemiologiczne, psychologiczne i socjologiczne, aby wykazać, że połączenie adopcji smartfonów i korzystania z mediów społecznościowych we wczesnym okresie dojrzewania jest przyczynowo powiązane z powszechnym pogorszeniem zdrowia psychicznego obserwowanym od 2012 r. w Ameryce Północnej, Europie i Australii. Chociaż w społeczności akademickiej toczą się debaty na temat przyczynowości, korelacja czasowa i spójność danych między krajami są uważane za istotne. *Co najważniejsze, wszystkie te badania poprzedzają erę LLM co najmniej o dekadę.*
- **PISA 2022 – Pogorszenie umiejętności czytania ze zrozumieniem:** Program OECD dotyczący międzynarodowej oceny uczniów (PISA 2022) odnotował największy międzynarodowy spadek wyników w zakresie czytania ze zrozumieniem od początku programu. W większości krajów objętych badaniem 15-latkowie wykazali pogorszenie, którego nie można w pełni wyjaśnić na podstawie danych wyjściowych sprzed pandemii. Badacze zauważają zgodność czasową z nasyceniem mediów cyfrowych, chociaż związek przyczynowy nie został ustalony z całą pewnością.
- **Wytyczne WHO dotyczące czasu korzystania z ekranu (2019 r.):** Światowa Organizacja Zdrowia zaleca, aby dzieci poniżej 2. roku życia nie korzystały z ekranu, maksymalnie jedną godzinę dziennie w przypadku dzieci w wieku 2–5 lat. Wytyczne te opracowano bez danych na temat interakcji za pośrednictwem sztucznej inteligencji i nie uwzględniają konwersacyjnych systemów sztucznej inteligencji, które behawioralnie różnią się od pasywnego korzystania z wideo.

7.2.2 DLACZEGO EKSPOZYCJA W EPOCE LLM JEST JAKOŚCIOWO RÓŻNA

Wszystkie wcześniejsze badania dotyczyły pasywnych lub transmitowanych mediów cyfrowych: wideo, kanałów społecznościowych, wyszukiwarek. LLM wprowadzają kategorycznie nową dynamikę – system reaguje. Dostosowuje się. Zapewnia odpowiedzi na żądanie, które wydają się wiarygodne. Stwarza to kilka powodów do obaw, których nie uwzględniają istniejące badania:

- **Odciążenie metapoznawcze:** Kiedy dziecko zleca rozumowanie sztucznej inteligencji, która generuje pewne, czytelne i wiarygodnie brzmiące wyniki, praca poznawcza polegająca na formułowaniu osądu – ocena źródeł, tolerowanie dwuznaczności, siedzenie w niepewności – nie jest już wykonywana. Nie zbadano jeszcze empirycznie, czy ciągłe odciążanie hamuje rozwój autonomicznej zdolności krytycznego myślenia. Jest to uzasadnione i mierzalne pytanie badawcze. Nie mamy jeszcze odpowiedzi.
- **Zamieszanie epistemiczne:** Dorośli mają trudności z odróżnieniem tekstu wygenerowanego przez sztuczną inteligencję od tekstu napisanego przez człowieka. Bardziej narażone są dzieci i młodzież, posiadające mniejszą wiedzę o świecie i mniej heurystyki. Dorastanie w środowisku informacyjnym, w którym autorytatywnie brzmiący tekst może, ale nie musi, mieć oparcie w rzeczywistości – i gdzie mechanizm generowania jest niewidoczny – stanowi warunek rozwojowy bez precedensu w historii.
- **Zależność parasocjalna:** Konwersacyjne systemy AI zostały zaprojektowane z myślą o zaangażowaniu. Nie męczą się, nie osądzają ani nie odrzucają. W przypadku młodszych użytkowników – zwłaszcza tych, którzy są już odizolowani społecznie lub zaniepokojeni – ryzyko powstania afektywnie asymetrycznych zależności (w przypadku których użytkownik przypisuje emocjonalne znaczenie interakcji AI) jest realne. W przeciwieństwie do klasycznych relacji paraspołecznych z gwiazdami, systemy sztucznej inteligencji reagują i dostosowują się, tworząc jakościowo bardziej wciągającą dynamikę.
- **Uwaga architektura:** Interfejsy oparte na LLM są zaprojektowane tak, aby generować kompletne odpowiedzi, redukując potrzebę eksploracyjnego wyszukiwania, czytania, syntezy i formułowania wniosków. Umiejętności czytania i wnioskowania w długich formach monitorowane przez PISA – już malejące – mogą spotkać się z dalszą presją ze strony pokolenia, które dorasta z dostępem do systemu przeprowadzającego za nich syntezę.

7.2.3 RYZYKO NA POZIOMIE INFRASTRUKTURY

Z punktu widzenia ryzyka infrastruktury systemowej przekłada się to na długoterminową troskę o kapitał ludzki: dostępność przyszłych inżynierów, analityków i operatorów zdolnych do zrozumienia, utrzymania i zabezpieczenia złożonej infrastruktury cyfrowej zależy od pokolenia

rozwijającego odpowiednie umiejętności poznawcze. Jeśli przyjęcie LLM na poziomie edukacyjnym przyspieszy odciążenie metapoznawcze w latach kształtowania się talentów, przepływ talentów do operacji infrastrukturalnych będzie narażony na ryzyko strukturalne, które ujawni się dopiero po latach 30. XXI wieku, ale zaczyna się kumulować już teraz.

Istnieje również bardziej bezpośrednie ryzyko polityczne. Populacje, które nie potrafią odróżnić informacji generowanych przez sztuczną inteligencję od pierwotnych raportów i które od dzieciństwa miały kontakt z systemami, które z pewnością odpowiadają na każde pytanie, są bardziej podatne na skoordynowane operacje wywierania wpływu na dużą skalę. Obrona infrastruktury wymaga ludzkich operatorów, którzy myślą kontradyktoryjnie, sceptycznie i lateralnie – a to cechy związane z dużą tolerancją na niejednoznaczność i wygodą z niekompletnymi informacjami. Cechy te kształtują się częściowo w okresie dojrzewania. Nie wiemy jeszcze, czy dorastanie pod okiem nauczycieli AI kształtuje je, czy niszczy.

To, co można powiedzieć z całą dokładnością, to: **nie wiemy.** Nie mamy danych. Brak badań podłużnych nad rozwojem poznawczym w epoce LLM nie napawa optymizmem – sam w sobie jest sygnałem ryzyka. Społeczeństwa i organizacje infrastrukturalne mają uzasadnione podstawy do stosowania zasady ostrożności: wyraźnie przyznają się do luki w wiedzy, finansują niezależne badania podłużne i unikają traktowania braku potwierdzonej szkody jako dowodu bezpieczeństwa.

17. Wniosek

Ukryte koszty automatyzacji na skalę LLM są już obecne w telemetrii produkcyjnej i rozkładają się nierównomiernie. Nie wszystkie osiem dziedzin ryzyka skatalogowanych w tej czarnej księdze ma ten sam schemat występowania, a jednoczące ramy „zewnętrzności” wymagają, aby możliwe było obronienie następującego dwutorowego rozróżnienia:

MIĘDZYORGANIZACYJNE EFEKTY ZEWNĘTRZNE

Nośnik kosztów i generator obciążenia to odrębne podmioty. Łagodzenie wymaga albo mechanizmów rynkowych (licencjonowanie treści), polityki (obowiązkowe ujawnianie informacji, zasady godziwej rekompensaty) albo obrony obwodowej (WAF, limit stawek, egzekwowanie robotów, zarządzanie postawą promocyjną).

Dotyczy: obciążenie wydawcy zewnętrzni (§4), przemieszczenie energii/wody/półprzewodników publicznych (§8), szkody dla bezbronnych użytkowników (§9), presja wydawców średniego szczebla (§12).

KOMPROMISY WEWNĄTRZORGANIZACYJNE

Organizacja wdrażająca sztuczną inteligencję jest zarówno generatorem obciążenia, jak i ponoszącym koszty. Łagodzenie to kwestia zarządzania i dyscypliny operacyjnej: polityka cyklu życia, oprzyrządowanie bazowe, planowanie wydajności.

Dotyczy: Wzrost ilości przechowywanych artefaktów AI we własnej chmurze adoptującego (§5), zanieczyszczenie korporacyjnego RAG/bazy wiedzy (§10.2), wzrost wolumenu SIEM we własnym rurociągu adoptującego (§6), obciążenie poznawcze operatora (§7).

Obie ścieżki są realne, obie są dziś mierzalne i obie są widoczne w telemetrii floty przedstawionej w § 4.4 i załączniku A. Implikacja operacyjna jest taka, że zarządzanie infrastrukturą sztucznej inteligencji nie jest pojedynczym problemem, na który można odpowiedzieć: ryzyko międzyorganizacyjne wymaga zaangażowania z rynkami i organami regulacyjnymi oprócz obrony obwodowej, natomiast ryzyko wewnątrzorganizacyjne wymaga wewnętrznej dyscypliny cyklu życia, której zewnętrzny organ regulacyjny nie może narzucić. Połączenie tych dwóch kwestii skutkuje albo źle przydzieloną uwagą organów regulacyjnych, albo źle przydzielonym budżetem inżynieryjnym.

Powierzchnie trzeciego wymiaru z widoku floty obejmującej wiele lokalizacji (pkt 4.4), które zwykle nie są wymieniane w literaturze dotyczącej infrastruktury sztucznej inteligencji: **wykrywalność** — czy właściwość jest w ogóle osiągalna za pomocą wykresu początkowego robota AI — jest powierzchnią kontrolną odrębną od obu technicznych rezygnacji (`robots.txt`) i profil treści. Flota wykazuje, że presja na szkolenie w zakresie sztucznej inteligencji różni się o ponad dwa rzędy wielkości w różnych witrynach o porównywalnym charakterze WAF, przy czym intensywność promocji (SEO, reklamy, przesyłanie map witryn, kampanie linków przychodzących) jest najbardziej prawdopodobną zmienną objaśniającą poza rodzajem treści. W przypadku operatorów, których wartość nieruchomości nie zależy od skierowań do wyszukiwarki od osób trzecich, postawa promocyjna jest dźwignią, którą można przeoczyć, a która została przeoczona. W przypadku operatorów, których wartość nieruchomości od tego zależy, dźwigni nie można zastosować hurtowo, ale można ją zastosować w odniesieniu do poszczególnych nieruchomości, co jest bardziej szczegółowym pytaniem dotyczącym zarządzania niż obecnie stawiane w tej dziedzinie.

Tam, gdzie ta czarna księga celowo zatrzymuje się: nie próbuje ona dokonać porównań z innymi konkurentami zajmującymi się infrastrukturą na rok 2026 (ewolucja oprogramowania ransomware jako usługi, migracja po kwantowej kryptografii, ryzyko koncentracji w chmurze, kompromis w łańcuchu dostaw, zmiany regulacyjne DORA/CRA). Bez tego porównania niniejszego dokumentu nie należy odczytywać jako twierdzenia, że ryzyko związane z infrastrukturą sztucznej inteligencji jest najwyższym priorytetem w roku 2026 – jedynie, że jest to wystarczająco istotny problem, z wystarczająco możliwymi do zidentyfikowania wzorcami występowania, aby zasługiwał na specjalne instrumentarium i prace związane z zarządzaniem. Towarzysząca temu biała księga (w wersji roboczej) zapewni ramy porównawcze wraz ze wskazówkami dotyczącymi łagodzenia zmiany klimatu.

Ujawnienie: obserwacje terenowe w §4.4 i Załączniku A zostały zebrane z chronionych przez BunkerWeb zakładów produkcyjnych prowadzonych przez autora. Zalecenia wymieniają BunkerWeb wśród innych opcji odwrotnego proxy i WAF (Cloudflare, Nginx-with-Lua); autor nie ma żadnych powiązań handlowych z projektem BunkerWeb poza obsługą go jako użytkownik. Narzędzia do zbioru floty wykorzystywane do produkcji agregatów § 4.4 mają charakter open source i są odtwarzalne (`harvest.report` , MIT, schemat `bw.harvest.v3`).

Załącznik A. Zweryfikowana telemetria terenowa (witryna anonimowa)

> **Uwaga dotycząca anonimizacji:** Etykiety domeny produkcyjnej, marki i kategorii zostały usunięte. Poniższy zbiór danych jest prezentowany jako *Strona A* aby zapobiec bezpośredniemu celowaniu, zachowując jednocześnie sygnał operacyjny.

W tym załączniku znajdują się dane telemetryczne wyodrębnione ze skonsolidowanych dzienników dostępu zwrotnego proxy i WAF na potrzeby anonimowego obciążenia handlu elektronicznego (**Strona A**) w 17-dniowym oknie obserwacyjnym (od 26 kwietnia 2026 r. do 12 maja 2026 r.). Przed integracją przeprowadzono kontrolę integralności danych: dzienne agregaty zostały ponownie obliczone i zweryfikowane w stosunku do sum globalnych, z zachowaniem dokładnej równości w zakresie liczby żądań, przesłanych bajtów i liczników zablokowanych żądań.

ZWERYFIKOWANE WYNIKI ZBIORCZE (OŚRODEK A)

METRYCZNY	WARTOŚĆ	INTERPRETACJA
Łączna liczba żądań	8 697 962	Wysokie ciśnienie obwodowe w mniej niż trzy tygodnie
Żądania sklasyfikowane przez sztuczną inteligencję	7153371 (82,24%)	W strukturze ruchu dominuje automatyzacja
Tradycyjne boty	745962 (8,58%)	Klasyczne roboty indeksujące pozostają istotne, ale drugorzędne
Ruch ludzki	798629 (9,18%)	Udział ludzki jest strukturalnie skompresowany
Łączna liczba przesłanych bajtów	920 369 355 879	~920,37 GB obsłużonych w obserwowanym okresie
Udział bajtów AI	878 038 133 231 (95,40%)	Obciążenie przepustowości jest w przeważającej mierze napędzane przez sztuczną inteligencję
Zablokowane żądania AI (HTTP 403)	1 036 427 (14,49% żądań AI)	Kontrole zabezpieczeń włączają się z utrzymującą się dużą częstotliwością
Przechodzenie między kategoriami	łącznie 2 482 198; 1 947 214 AI (78,45%)	Głębokie przeglądanie katalogu odbywa się głównie za pomocą maszyn

A.1 ODCZYT OPERACYJNY

- **Obciążenie asymetryczne:** Ruch AI nie tylko dominuje pod względem liczby żądań, ale nieproporcjonalnie dominuje pod względem obsługiwanych bajtów, co potwierdza, że największym centrum kosztów infrastruktury jest popyt pochodzący z maszyny.
- **Presja obrony:** Siedmiocyfrowa liczba zablokowanych żądań sztucznej inteligencji w ciągu 17 dni wskazuje na utrzymującą się presję kontradyktoryjną lub co najmniej niechętną do współpracy presję automatyzacji na brzegu sieci.
- **Katalog skupia się na:** Większość automatyzacji o wysokiej częstotliwości koncentruje się na ścieżkach nawigacji kategorii/list, a nie tylko na stronach docelowych najwyższego poziomu, co zwiększa koszty zapytań zaplecza i kosztów pominięcia pamięci podręcznej w dynamicznych stosach handlowych.

- **Ciągłość ekspozycji:** Przy tej wielkości systemy proxy/WAF, rejestrowanie, przetwarzanie SIEM i systemy tworzenia kopii zapasowych stają się raczej połączonymi powierzchniami ryzyka niż niezależnymi warstwami.

A.2 UWAGA METODOLOGICZNA

Klasyfikacja wykorzystywała deterministyczne rodziny agentów użytkownika (roboty AI, tradycyjne boty, resztkowy ruch ludzki) oraz dystrybucję kodów stanu i liczniki wzorców adresów URL. Załącznik celowo wyklucza surowe domeny, pełne etykiety adresów URL i bezpośrednie identyfikatory handlowe. Celem jest odtwarzalna charakterystyka ryzyka bez publikowania możliwych do ukierunkowania odcisków palców infrastruktury.

Załącznik B. Audyt odtwarzalności wykrywalności przez czytnik

Zarówno obserwacja floty z §4.4, jak i zalecenie dotyczące postawy promocyjnej z §13 opierają się na twierdzeniu, że presja robotów szkolących sztuczną inteligencję koreluje z publicznym wykrywaniem, a nie tylko z profilem treści lub rezygnacją techniczną. Poniższa kontrola umożliwia każdemu operatorowi posiadającemu dostęp powłoki do kontrolowanej przez niego właściwości wygenerowanie sygnału wykrywalności pierwszego rzędu dla tej właściwości w czasie krótszym niż pięć minut, bez uprzywilejowanych danych stron trzecich. Nie zastępuje płatnych audytów SEO ani audytów wykresów referencyjnych; jest to dolna granica obserwacyjnej linii bazowej.

> **Zakres.** Audyt obejmuje cztery sygnały: (1) obecność mapy witryny i liczbę adresów URL; (2) `robots.txt` dyrektywy dla robotów AI; (3) obecność w indeksowanych adresach URL Common Crawl (próbka, niewyczerpująca); (4) znormalizowany wynik postawy promocyjnej, łączący poprzednie trzy. Nie mierzy wykresu linków przychodzących, wydatków na reklamę ani obecności w katalogach stron trzecich; wymagają one płatnych źródeł danych.

B.1 SKRYPT AUDYTU

Zapisz poniższe jako `discoverability-audit.sh`, uczyń wykonywalnym (`chmod +x`) i wywołaj jako `./discoverability-audit.sh https://your-property.example`. Wymaga `curl`, `grep`, I `wc` (kompatybilny z BusyBox).

```
#!/usr/bin/env bash
# discoverability-audit.sh - first-order AI-discoverability signal
# Usage: ./discoverability-audit.sh https://your-property.example
set -euo pipefail
URL="${1:-}"
if [[ -z "$URL" ]]; then echo "Usage: $0 https://your-property.example" >&2; exit 2;
fi
HOST="$(echo "$URL" | sed -E 's#^https?://([^/]+).*#\1#')"
echo "=== Discoverability audit: $HOST ==="

# 1. Sitemap presence + URL count
echo "--- 1. Sitemap ---"
for SM in sitemap.xml sitemap_index.xml sitemap-index.xml; do
  CODE="$(curl -s -o /tmp/sm.$$ -w '%{http_code}' "$URL/$SM" || echo 000)"
  if [[ "$CODE" == "200" ]]; then
    COUNT="$(grep -c '<loc>' /tmp/sm.$$ || echo 0)"
    echo "  $SM: HTTP 200, ${COUNT} <loc> entries"
  fi
done
rm -f /tmp/sm.$$

# 2. robots.txt AI directives
echo "--- 2. robots.txt AI directives ---"
curl -s "$URL/robots.txt" -o /tmp/rb.$$ || echo "  (no robots.txt)"
if [[ -s /tmp/rb.$$ ]]; then
  for UA in GPTBot ChatGPT-User ClaudeBot Claude-Web anthropic-ai Google-Extended
CCBot PerplexityBot meta-externalagent FacebookBot Bytespider; do
    if grep -qi "User-agent:.*$UA" /tmp/rb.$$; then
      echo "  $UA: declared"
    fi
  done
fi
rm -f /tmp/rb.$$

# 3. Common Crawl presence (sample - latest monthly index)
echo "--- 3. Common Crawl presence (sample) ---"
CC_INDEX="$(curl -s https://index.commoncrawl.org/collinfo.json | grep -oE '\"cdx-
api\":\"[^\"]+\"' | head -1 | sed 's/\"cdx-api\":\"//;s/\"//')"
if [[ -n "$CC_INDEX" ]]; then
  CC_COUNT="$(curl -s "${CC_INDEX}?url=${HOST}/*&output=json&limit=1000" | wc -l)"
  echo "  Latest monthly index: ${CC_COUNT} URLs indexed (capped at 1000 sample)"
else
  echo "  (Common Crawl index unreachable)"
```

```
fi

echo "--- Done ---"
echo "Interpretation:"
echo "  - High sitemap count + few robots blocks + high CC presence  => HIGH discoverability"
echo "  - No sitemap or AI-bot blocks declared + low CC presence    => LOW discoverability"
echo "  - Compare across your fleet; flag outliers per direction."
```

B.2 PRZEWODNIK INTERPRETACYJNY

- **Gęstość mapy witryny:** Usługa z ponad 10 000 adresów URL map witryn i bez bota AI `robots.txt` wykluczenia znajdują się na górnym końcu spektrum wykrywalności. Usługa bez mapy witryny i co najmniej 5 wykluczeń botów AI jest na najniższym poziomie.
- **Powszechna obecność indeksowania:** Właściwości pojawiające się w najnowszym miesięcznym indeksie Common Crawl z nietrywialną liczbą adresów URL (> 100 z limitem próbek) znajdują się na wykresie początkowym, z którego wypływa większość potoków szkoleniowych z zakresu sztucznej inteligencji. Nieobecność na pełzaniu pospolitym nie jest dowodem niskiego ciśnienia (istnieją inne wykresy pełzające), ale obecność jest mocnym dowodem wysokiego ciśnienia.
- **Porównanie różnych flot:** Podstawową wartością audytu jest różnica. Uruchom go na pełnym zestawie właściwości operatora; właściwość, która zajmuje najwyższą pozycję pod względem liczby map witryny×obecności CC, a najniższą pod względem liczby wykluczeń robotów, to usługa najbardziej narażona na obciążenie robota indeksującego uczącego sztuczną inteligencję, przy wszystkich innych czynnikach równych.
- **Limity:** Brak pomiaru wykresu linków przychodzących, wydatków na reklamę lub uwzględnienia płatnych katalogów. Wymagają one danych stron trzecich (Ahrefs, Semrush, PodobneWeb) i nie podlegają bezpłatnej, powtarzalnej kontroli przez czytelnika.

Referencje

1. Imperwa. *Raport o złych botach 2024.* Laboratoria badawcze Imperva, kwiecień 2024 r. Dostępne na imperva.com/resources/resource-library/reports/bad-bot-report/
2. OpenAI. *Dokumentacja GPTBota.* Sierpień 2023. Dostępne na platformie.openai.com/docs/gptbot
3. Centrum wyszukiwarki Google. *Rozszerzona kontrola Google do szkolenia modeli AI.* Wrzesień 2023 r. Dostępne na stronie Developers.google.com/search/docs/crawling-indexing/google-extended

4. Fundacja Wspólnego Czołgania. *Wspólne statystyki indeksowania i przegląd danych.* Commoncrawl.org
5. IDC. *Cyfryzacja świata – od krawędzi do rdzenia (Data Age 2025).* Biała księga IDC sponsorowana przez firmę Seagate, listopad 2018 r., z kolejnymi aktualizacjami na lata 2022–2024.
6. Reinsel, D., Gantz, J., Rydning, J. *Cyfryzacja świata.* Biała księga IDC, 2018. numer dokumentu US44413318.
7. Parlament Europejski i Rada. *Rozporządzenie (UE) 2024/1689 ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji (akt UE o sztucznej inteligencji).* Dziennik Urzędowy Unii Europejskiej, lipiec 2024 r.
8. Parlament Europejski i Rada. *Rozporządzenie (UE) 2016/679 (RODO).* Dziennik Urzędowy Unii Europejskiej, maj 2016.
9. Parlament Europejski i Rada. *Dyrektywa (UE) 2022/2555 w sprawie środków na rzecz wysokiego wspólnego poziomu cyberbezpieczeństwa (NIS2).* Dziennik Urzędowy Unii Europejskiej, grudzień 2022 r.
10. Cloudflare. *Radar Cloudflare — trendy w ruchu botów.* radar.cloudflare.com (na bieżąco aktualizowana, odniesienie do danych za 2024 r.).
11. Technologie Akamai. *Stan Internetu: Raport bezpieczeństwa 2024.* akamai.com/resources/state-of-the-internet-report
12. Perez, E. i in. *Zignoruj poprzedni monit: Techniki ataku dla modeli językowych.* Warsztaty NeurIPS 2022 na temat bezpieczeństwa uczenia maszynowego. (Badania podstawowe dotyczące szybkiego zastrzyku.)
13. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R. *Klątwa rekurencji: szkolenie na wygenerowanych danych sprawia, że modele zapominają.* Przyroda, lipiec 2024.
14. Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R. i in. *Czy upadek modelu jest nieunikniony? Przełamanie klątwy rekurencji poprzez gromadzenie danych rzeczywistych i syntetycznych.* arXiv:2404.01413, kwiecień 2024 r.
15. Antropiczny. *Karta modelu Claude i zasady użytkowania.* anthropic.com/model-card (odniesienie do dokumentacji ClaudeBota).
16. Microsoftu. *Raport roczny za rok 2024.* microsoft.com/en-us/investor/annual-reports.aspx
17. Microsoftu. *Zasady przechowywania w usłudze OneDrive dla Firm.* docs.microsoft.com (możliwości wersjonowania i przechowywania).
18. ISO/IEC 27001:2022. *Systemy zarządzania bezpieczeństwem informacji - Wymagania.* Międzynarodowa Organizacja Normalizacyjna.
19. Naród handlu elektronicznego / Przyjaciele Presty. *Barometre CMS e-commerce we Francji: Shopify domine les kreacje, PrestaShop le chiffre d'affaires.* Marzec 2026. Dostępne na ecommerce-nation.fr/barometre-cms-ecommerce-shopify-creations-prestashop-chiffre-affaires/
20. Międzynarodowa Agencja Energetyczna (IEA). *Energia elektryczna 2024: Analiza i prognoza do 2027 roku.* Publikacje IEA, 2024. (Odniesienie do globalnych danych dotyczących intensywności emisji dwutlenku węgla przez energię elektryczną.)
21. Strubell, E., Ganesh, A., McCallum, A. *Względy energetyczne i polityczne dotyczące głębokiego uczenia się w NLP.* Materiały z 57. dorocznego spotkania Association for Computational Linguistics (ACL), 2019. (Badanie podstawowe dotyczące zużycia energii w szkoleniu LLM).

22. OpenAI. *Ujawnianie zdolności wnioskowania i trendy w zużyciu energii.* Ujawnianie wewnętrzne w formie raportów rocznych i postów na blogu, 2023–2025. (Odniesienie do bieżących szacunków obciążenia wnioskowaniem).
23. Nvidia. *Analiza popytu i łańcucha dostaw na GPU.* Relacje Inwestorskie i raporty rynkowe, 2023–2025. (Odniesienie do niedoboru GPU H100/H200.)
24. Patterson, D. i in. *Ślad węglowy szkoleń w zakresie uczenia maszynowego ustabilizuje się, a następnie zmniejszy.* Computer, IEEE, 2021. (Badania dotyczące rozliczania i łagodzenia emisji dwutlenku węgla przez LLM).
25. Luccioni, A.S., Mahendran, A. *Kwantyfikacja emisji dwutlenku węgla wynikająca z uczenia maszynowego.* arXiv:1910.09700, 2019. (Metodologia szacowania śladu węglowego w AI.)
26. Twenge, J.M. i in. *Wzrost liczby objawów depresji, skutków samobójstw i wskaźników samobójstw wśród nastolatków w USA po 2010 r. oraz linki do wydłużenia czasu spędzanego przed ekranem nowych mediów.* Clinical Psychological Science, 2018. (Podłużna korelacja między czasem spędzonym przed ekranem a zdrowiem psychicznym nastolatków).
27. Haidt, J. *Niespokojne pokolenie: jak wielka zmiana nawyków w dzieciństwie powoduje epidemię chorób psychicznych.* Penguin Press, marzec 2024 r.
28. OECD. *Wyniki badania PISA 2022 (tom I): Stan uczenia się i równość w edukacji.* OECD Publishing, 2023. (Dokumentuje największy odnotowany międzynarodowy spadek wyników w zakresie czytania ze zrozumieniem).
29. Światowa Organizacja Zdrowia. *Wytyczne dotyczące aktywności fizycznej, siedzącego trybu życia i snu dla dzieci poniżej 5 roku życia.* WHO Press, 2019. (Wytyczne dotyczące czasu wyświetlania, sprzed interakcji AI z ery LLM).

Czarna księga – seria dotycząca ryzyka infrastruktury

Dokument ten przedstawia kontradyktoryjną, skoncentrowaną na ryzyku perspektywę wdrażania LLM w skali globalnej. Jest to pierwsza z dwuczęściowej serii. Towarzysząca Biała Księga (konstruktywne ścieżki, środki łagodzące, możliwości) jest obecnie w fazie roboczej. Wszystkie twierdzenia oparte na faktach pochodzą z publicznie weryfikowalnych źródeł pierwotnych. Projekcje kierunkowe kwalifikowane jako szacunki. Nie wykorzystano żadnych zastrzeżonych danych. Żadne statystyki generowane przez sztuczną inteligencję nie są przedstawiane jako oparte na faktach i bez zastrzeżeń.

Opublikowano: maj 2026 · Autor: Bryce SIMON · Współautor: Ifrit (AI) · Licencja: CC BY 4.0 · Biała księga towarzysząca: w wersji roboczej

```
const anchorSrc = ;
```