

IIA X Agent Lake

Jak efektywnie i bezpiecznie pracować z AI w audycie?

O nas

Maciej i Piotr - założyciele Agent Lake <https://agentlake.dev/>

Piotr:

- pierwszy inżynier w Ramp - \$130M finansowania
- inżynier danych w Ntropy - standaryzujący nieustrukturyzowane dane finansowe
- CTO i architekt systemu w Threshold Labs - \$5B wolumenu

Maciej:

- pierwszy CTO Jobandtalent - dzisiaj unicorn, wycena \$1B
- CTO i założyciel Jobartis — jednego z największych portali pracy w Afryce Subsaharyjskiej

Plan na dziś

- Część 1 - teoria
- Część 2 - warsztat praktyczny
- Część 3 - bonus i narzędzia
- Część 4 - pytania

Część 1

Jak bezpiecznie pracować z AI?

Dylemat audytora

- Chcesz przeanalizować **całą księgę** narzędziem AI — szybko, 100% transakcji
- Ale w księdze są: **nazwiska, NIP, kwoty, tajemnica zawodowa**
- Publiczne AI = wklejasz dane, a one **lecą na serwer dostawcy**
- Pozorny wybór: **korzyść z AI albo bezpieczeństwo danych**
- Pokazujemy jak mieć **prywatność i wsparcie AI**

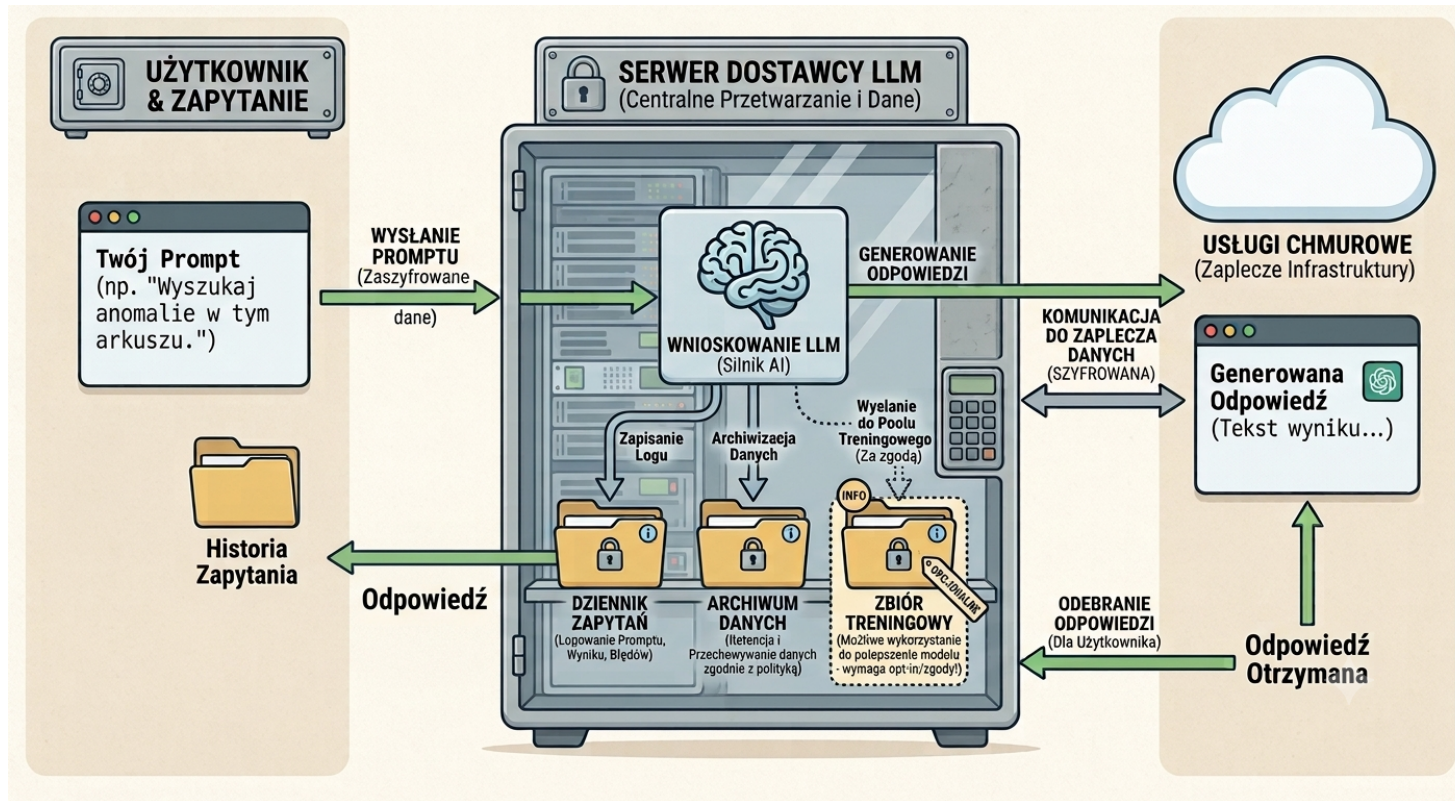
Jak działa LLM?

- Przewiduje kolejne słowa (tokeny) na podstawie wzorców z tekstu
- Nie „rozumie” i nie „wie” — dopasowuje prawdopodobne słowa
- Potrafi napisać **wiarygodnie wyglądający, błędny** wynik
- Dlatego nadzór człowieka jest konieczny

Dwa sposoby użycia AI

- **Czat (GPT, Gemini, Claude):** wklejasz treść → leci do chmury i tam jest przetwarzana
- **Asystent kodowania (Claude Code / Cowork):** dostaje opis zadania i strukturę, przykładowe dane, ale dane przetwarza lokalnie

Wędrówka danych



- Czy dane trafiają do treningu i na jak długo — zależy od umowy, nie od dostawcy

Prywatność danych

	Prywatność	Główne ryzyko dla audytora w UE
Dostawca USA, konto konsumenckie	zerowa	trening, długa retencja, CLOUD Act
Dostawca USA, API / komercyjny	tak, bez treningu	umowa (DPA/ZDR), ale CLOUD Act
Model USA w chmurze UE (Bedrock/Vertex/Azure)	tak, ale przetwarzane w UE	rozwiązuje lokalizację (RODO), nie CLOUD Act
Dostawca UE (np. Mistral, Francja)	tak, serwery UE	zwykle ryzyko dostawcy
Dostawca chiński (DeepSeek, hostowany)	tak, na serwery w Chinach	dostęp państwa (PIPL/ustawy ChRL); działania regulatorów UE
Model lokalny (Ollama, llama.cpp)	tak	luka jakości, koszt sprzętu, koszty bezpieczeństwa

Chat vs Asystent kodu

	 Czat (GPT, Gemini)	 Asystent kodu
Co dostaje AI	wszystkie dane	tylko nazwy kolumn (schemat)
Przykład	Bartex Sp. z o.o., NIP 777..., 184 900 PLN, Anna Kowalska	numer_faktury, kontrahent, nip, kwota_netto
Gdzie liczą się dane	serwer dostawcy	lokalnie, na Twoim dysku
Co wycieka	nazwiska, NIP, kwoty	nic — dane przetwarzane lokalnie

Co bezpiecznie udostępniać

TAK	NIE
nazwy kolumn, schemat	prawdziwe kwoty
opis zadania	nazwiska, NIP, PESEL, numery tel.
polityka (np. tabela limitów)	całe wiersze z księgi
sztuczne / zanonimizowane przykłady	dane klienta w jakiejkolwiek formie

Anonimizowane ≠ wolno wysłać

- **RODO można "wyłączyć"** — ale poprzeczka jest wysoka: prawdziwa anonimizacja to niemożliwość *wskazania osoby, powiązania i wnioskowania* (motyw 26 RODO, EDPB 01/2025)
- Samo zamaskowanie nazwiska to **pseudonimizacja** — RODO nadal obowiązuje
- **Nawet dane nieosobowe zostają poufne:**
 - tajemnica zawodowa biegłego rewidenta — **bezterminowa** (art. 78 ustawy)
 - umowa z klientem · tajemnica przedsiębiorstwa · tajemnica skarbową (art. 293 O.p.)

A może po prostu „wyczyścić” dane?

- **Reguły (regex):** lokalne i pewne dla PESEL / NIP / IBAN (mają sumy kontrolne) — ale ślepe na nazwiska i kontekst
- **AI do czyszczenia:** łapie więcej — ale trzeba ją gdzieś uruchomić; w chmurze **to znów ten sam wyciek**
- „Wyczyść raz, używaj wszędzie” kusi, ale: **czyszczenie nigdy nie łapie 100%**, a kombinacja (kwota + data + rzadki kontrahent) i tak wskaże osobę
- Nasz wzorzec **omija ten problem:** nie wysyłasz danych → nie potrzebuje anonimizacji

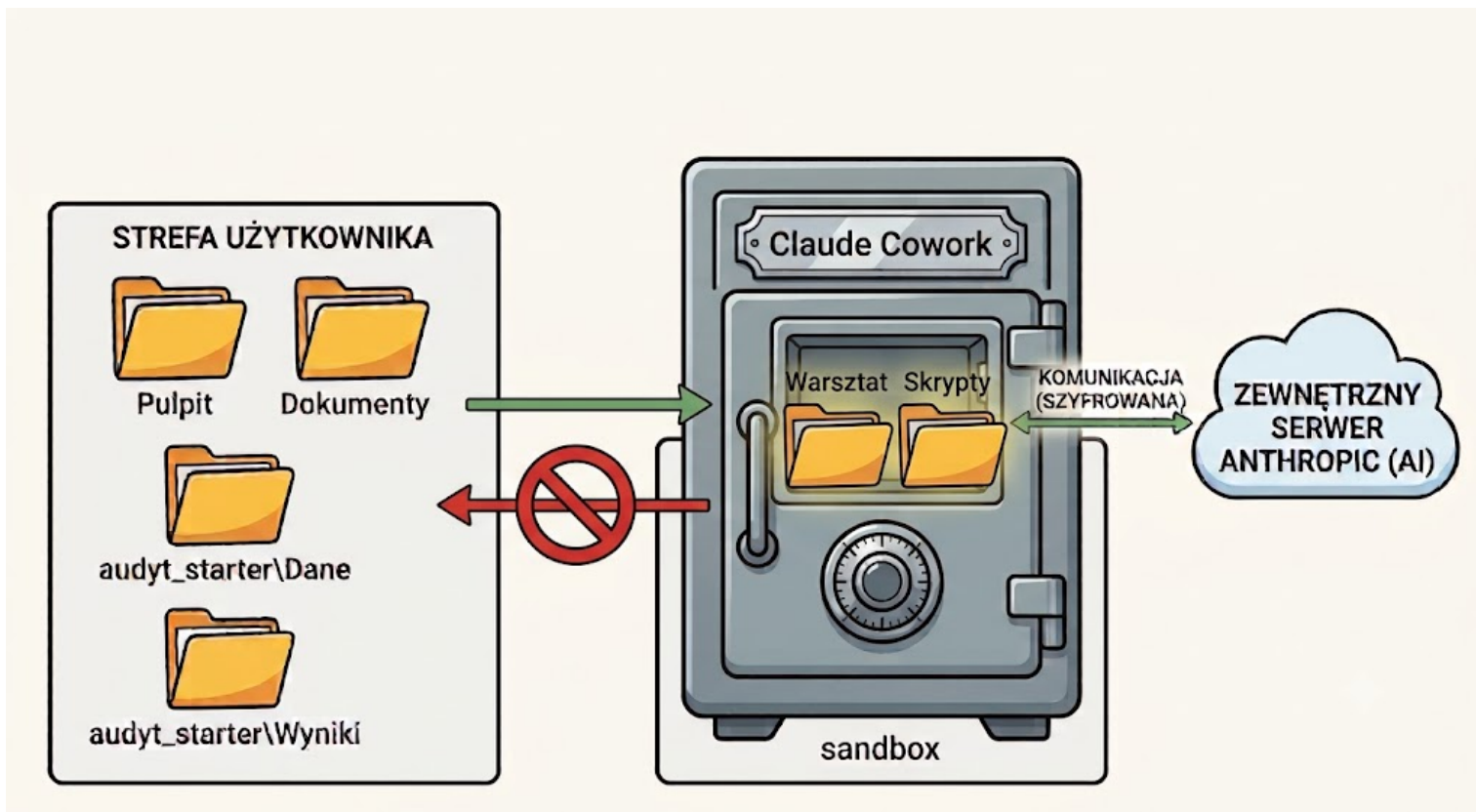
AI Act — artykuł 4

- Od lutego 2025: obowiązek **kompetencji w zakresie AI**
- Osoby używające AI mają rozumieć, **jak działa i jak używać bezpiecznie**
- Ten webinar realizuje to wymaganie

Zamknij narzędzie w pudełku (sandbox)

- Dyscyplina to jedno — można ją też wymusić technicznie
- Dwie filozofie:
 - **Pozwól na wszystko, zabraniaj wyjątków** (czarna lista) — wygodne, dziurawe
 - **Zabroń wszystkiego, pozwalaj wyjątkami** (biała lista) — więcej pracy, pełna kontrola

Jak działa sandbox



Higiena pracy z AI

- Czytaj kod z AI przed uruchomieniem
- Proś AI w oddzielnej sesji o wyjaśnienie kodu
- Sprawdzaj wynik na **znanych danych** (golden file)
- Model bywa **najbardziej pewny siebie, gdy się myli**
- Prompt injection
- Narzędzie liczy — **osąd należy do audytora**

Checklista bezpieczeństwa

1. Nie wklejaj wartości — wysyłaj tylko **strukturę**
2. Licz **lokalnie**, na swoim komputerze
3. Sprawdzaj **kod i wynik** na znanym przykładzie
4. Odpowiedzialność zostaje **po stronie człowieka**

Założ, że wszystko, co wpiszesz, ktoś kiedyś przeczyta.
Czy ten wiersz może wyciec?

Część 2

Warsztat na żywo: Claude Cwork + Python

Czym jest Claude Cowork

- Działa w izolowanym kontenerze na Twoim komputerze
- Czyta lokalne pliki
- Wysyła udostępnione dane do modelu w chmurze
- Piszemy skrypty w Pythonie

Dane do warsztatu

- W pełni syntetyczna, wygenerowana księga
- ~20 tys. rekordów
- w pełni zanonimizowana próbka do udostępnienia AI
- Bezpieczne do ćwiczeń

Flow pracy w czterech krokach

1. Opisujesz problem
2. AI pisze skrypt
3. audyt_starter uruchamia go poza Claude Cowork
4. Sprawdzasz wynik

audyt_starter

- CLAUDE.md - instrukcja dla CLAUDE
- warsztat pracy
- struktura katalogów
- Ćwiczysz wzorzec, który zastosujesz na prawdziwych danych u siebie

Krok 0 — odkrywanie danych

- Tworzymy opis danych
- Testy na nieznanym pliku = dużo pracy i dużo śmieci
- Analiza - efektywna praca i lepsze wyniki

Demo: profil danych — ten sam wzorzec

- Nie wrzucasz pliku do AI
- AI pisze skrypt profilujący, Ty uruchamiasz lokalnie
- Skrypt drukuje: kolumny i typy, zakresy dat i kwot, braki, powtórzone klucze
- AI zna najwyżej nazwy kolumn z wiersza nagłówka

Od profilu do planu testów

- Profil czytasz u siebie
- Wybierasz kolumny i testy sensowne dla tej księgi
- Schemat i sumy możesz pokazać AI — pojedynczych rekordów nigdy

Przypadek A — anomalie: czego szukamy

- Nietypowe kwoty, daty, godziny
- Nietypowe osoby akceptujące
- Wszelkie inne anomalie

Przypadek A — DEMO NA ŻYWO

→ / Claude Cwork

Prompt: „Wskaż w księdze anomalie: nietypowe, okrągłe, równe kwoty, księgowania nocne i weekendowe, przekroczenia limitu akceptacji, akceptujący sam zatwierdzający swój wpis.”

Przypadek A — wynik

- 33 kwoty brutto $> 100\,000$ zł · 50 kwot okrągłych (co do 100 zł)
- 15 księgowania nocnych · 14 w weekend
- 1 748 akceptacji ponad limit kwotowy
- 8 akceptacji przez osobę spoza tabeli limitów
- 8 przypadków: akceptujący = wprowadzający wpis

Przypadek B — czemu ściśle dopasowanie nie wystarcza

- Ta sama faktura zapłacona dwa razy ma drobne różnice
- Inna pisownia kontrahenta, kwota o grosze, data o dzień
- Potrzebujemy „podobne”, nie tylko „identyczne”

Przypadek B — DEMO NA ŻYWO

→ / Claude Cwork

Prompt: „Znajdź prawdopodobne zdublowane płatności: podobne nazwy kontrahentów, zbliżone kwoty i bliskie daty. Grupuj po miesiącu i kwocie.”

Przypadek B — wynik

- 12 par dokładnych (numer + kwota + kontrahent)
- 10 par rozmytych (podobieństwo 0,82–0,92)
- 4 płatności podzielone (ta sama kwota, inny numer)
- Każdy sygnał sprawdzamy ręcznie

Przypadek C — prawo Benforda

- W naturalnych danych pierwsza cyfra „1” częstsza niż „9”
- Kto zmyśla kwoty — układa cyfry zbyt równo
- Metoda wskazuje, gdzie szukać

Przypadek C — DEMO NA ŻYWO

→ / Claude Cowork

Prompt: „Policz rozkład pierwszych cyfr `kwota_netto` w całej księdze, porównaj z rozkładem Benforda i narysuj wykres. Pokaż też rozkład per centrum kosztów.”

Przypadek C — interpretacja i ograniczenia

- MAD całej księgi $\approx 0,022$ (widoczne odchylenie)
- Po odjęciu „Działu Marketingu” spada do $\approx 0,003 \rightarrow$ źródło znalezione
- Nadreprezentacja cyfr 4 i 9 (kwoty tuż pod progami)
- **Nie stosuj** do: numerów faktur, kwot z narzuconym limitem, małych zbiorów

Bonus

Sheet Sheriff - strażnik Twojego Excela

- sprawdza matematyczną poprawność plików
- znajduje błędy trudne do wykrycia ręcznie:
 - zakresy funkcji urwane przed końcem danych
 - niespójne zakotwiczenie \$
 - ręcznie wpisane wartości zamiast sum
 - wiele innych
- <https://sheetsheriff.com/>

Zakończenie

- AI bardzo pomaga
- AI może być niebezpieczna
- AI się myli
- AI to narzędzie
- Człowiek decyduje

Linki

- <https://agentlake.dev/audyt-starter/> - bezpieczny szablon do pracy z AI
- <https://sheetsheriff.com> - twój strażnik Excela
- maciej@agentlake.dev