Typ ataku	Nazwa angielska	Kod	Opis	Kryterium	Cel ataku	Konsekwencje udanego ataku	Metoda obrony
Nadmierna autonomia	Excessive Agency	LLM08	Przyznanie modelom AI zbyt dużej autonomii, co może prowadzić do nieprzewidzianych konsekwencji, takich jak naruszenie prywatności i bezpieczeństwa.	Atak na integralność	Algorytm	Zbyt duża autonomia modelu może prowadzić do nieprzewidzianych działań, narażając prywatność i bezpieczeństwo użytkowników.	Stosowanie mechanizmów nadzoru nad działaniami modelu oraz ograniczanie jego autonomii w krytycznych zastosowaniach.
Niebezpieczna konstrukcja API	Insecure Plugin Design	LLM07	Niewystarczające zabezpieczenia wtyczek modelu AI, co może prowadzić do ataków, takich jak zdalne wykonanie kodu.	Atak na integralność/poufność	Algorytm/model	Wprowadzenie złośliwego oprogramowania przez wtyczki, które może prowadzić do zdalnego wykonania złośliwego kodu.	Ograniczanie zaufania do wtyczek, stosowanie sandboxingu oraz regularne audyty kodu wtyczek.
Wycieki danych	Data Exfiltration		Uzyskiwanie nieautoryzowanego dostępu do danych, które były używane przez model AI.	Atak na poufność	Dane treningowe	Kradzież danych wykorzystywanych przez model, co może skutkować wyciekiem informacji wrażliwych oraz utratą zaufania.	Szyfrowanie danych oraz stosowanie mechanizmów kontroli dostępu do danych wrażliwych.
Zatrucie danych treningowych	Training Data Poisoning	LLM03 ML02: 2023	Wprowadzanie złośliwych danych do zbioru treningowego, co prowadzi do błędnego działania modelu.	Atak na integralność	Dane treningowe	Uszkodzenie procesu trenowania modelu, co prowadzi do trwałych błędów operacyjnych i nieprawidłowych predykcji.	Monitorowanie jakości danych treningowych, zabezpieczanie procesów trenowania oraz stosowanie certyfikowanych źródeł danych.
Atak inferencji członkostwa	Membership Inference Attacks	ML04: 2023	Uzyskiwanie informacji o tym, czy konkretne dane zostały użyte podczas trenowania modelu	Atak na poufność	Dane treningowe	Ujawnienie informacji o danych, które były używane do trenowania modelu, co może naruszyć poufność i prywatność danych.	Kontrola dostępu, monitorowanie, szyfrowanie parametrów modelu
Przechylenie modelu	Model Skewing	ML08: 2023	Manipulowanie danymi w celu celowego wypaczenia wyników modelu	Atak na integralność	Dane treningowe	Model zwraca stronnicze lub nieprawidłowe wyniki, co wpływa na jego decyzje	Monitorowanie wyników, walidacja danych, stosowanie technik ograniczania wpływu błędnych danych
Ataki adwersarialne	Adversarial Attacks		Wprowadzanie subtelnych zmian w danych wejściowych w celu oszukania modelu AI i uzyskania niepoprawnych wyników.	Atak na integralność	Dane wejściowe	Model zwraca niepoprawne wyniki, co może prowadzić do błędnych decyzji i destabilizacji systemów o wysokim znaczeniu.	Stosowanie mechanizmów detekcji ataków adwersarialnych oraz zwiększanie odporności modelu na zaburzenia (adversarial training).
Ataki omijające	Evasion Attacks		Wykorzystanie zmodyfikowanych danych wejściowych w celu uniknięcia wykrycia przez system AI.	Atak na integralność	Dane wejściowe	Atakujący omija systemy wykrywania, co może skutkować przepuszczeniem zagrożeń bez odpowiedniej reakcji.	Używanie zaawansowanych systemów detekcji anomalii i dynamiczne dostosowywanie progów wykrywania zagrożeń.
Wstrzyknięcie polecenia	Prompt Injection	LLM01	Manipulowanie danymi wejściowymi (promptami) w celu uzyskania nieautoryzowanego dostępu lub kontroli nad modelem.	Atak na integralność	Dane wejściowe	Model wykonuje nieautoryzowane działania, co może prowadzić do wycieku poufnych danych lub manipulacji wynikami.	Walidacja i filtrowanie danych wejściowych, ograniczanie dostępności modeli oraz stosowanie zasad least privilege w interfejsach API.
Ataki enkodera	Encoder Attacks		Manipulacja kodowaniem danych, aby algorytmy ML nie rozpoznały szkodliwych danych.	Atak na integralność	Dane wejściowe	Model błędnie interpretuje dane wejściowe, co może prowadzić do nieprawidłowych wyników lub wycieków danych poufnych.	Detekcja przykładów adwersarialnych, analiza kodowania oraz monitorowanie zbiorów treningowych.
Luki w łańcuchu dostaw	Supply Chain Vulnerabilities	LLM05 ML06: 2023	Wykorzystanie słabości w komponentach, usługach lub danych dostarczanych do systemu AI, co może prowadzić do wycieków danych lub awarii systemu.	Atak na integralność/poufność	Dane/model	Wykorzystanie luk w łańcuchu dostaw, co może skutkować awarią systemu, wyciekiem danych i zakłóceniami operacyjnymi.	Regularne audyty bezpieczeństwa dostawców, certyfikowanie dostarczanych komponentów i monitorowanie procesu dostaw.
Ataki typu backdoor	Backdoor Attacks		Umieszczanie ukrytych funkcji w modelu AI, które mogą zostać aktywowane przez atakującego.	Atak na integralność	Model	Aktywacja ukrytych złośliwych funkcji, które mogą całkowicie przejąć kontrolę nad systemem lub sabotować jego działanie.	Regularne audyty modelu, analiza kodu pod kątem obecności backdoorów oraz stosowanie testów weryfikacyjnych przed wdrożeniem modelu.
Ataki typu trojan	Trojan Attacks		Instalowanie złośliwego kodu, który aktywuje się w określonych warunkach i przejmuje kontrolę nad działaniem modelu.	Atak na integralność	Model	Przejęcie kontroli nad systemem w określonych warunkach, umożliwiające atakującym pełną kontrolę nad modelem.	Zabezpieczanie procesu trenowania i walidacji, testowanie modelu w różnych warunkach operacyjnych.
Kradzież modelu	Model Theft	LLM10 ML05: 2023	Nieautoryzowane uzyskanie dostępu do modelu, jego parametrów lub wagi, aby odtworzyć go.	Atak na integralność	Model	Kradzież własności intelektualnej, utrata przewagi konkurencyjnej i ujawnienie poufnych danych.	Zabezpieczenie dostępu do modelu przez mechanizmy uwierzytelniania, licencjonowanie oraz monitorowanie nadużyć.
Atak typu odmowa usługi (DoS)	Model Denial of Service	LLM04	Przeciążenie modelu AI operacjami wymagającymi dużych zasobów, prowadzące do przerw w działaniu.	Atak na dostępność	Model	Zablokowanie dostępu do zasobów modelu, co prowadzi do przerw w jego działaniu, utraty dostępności i wzrostu kosztów.	Stosowanie mechanizmów ochrony przed atakami DoS (np. rate limiting) oraz monitorowanie zużycia zasobów w czasie rzeczywistym.
Ataki bocznokanałowe	Side-Channel Attacks		Wykorzystanie informacji pośrednich, takich jak czas reakcji, do uzyskania danych o modelu AI.	Atak na poufność	Model	Wykorzystanie informacji pośrednich do wycieku danych lub rekonstrukcji danych wejściowych, co może naruszać prywatność.	Zastosowanie technik maskowania i ochrony przed atakami bocznokanałowymi, takich jak ukrywanie informacji o czasie przetwarzania danych.
Ataki na transfer uczenia	Transfer Learning Attacks	ML07: 2023	Wykorzystanie złośliwych modeli wstępnie wytrenowanych, które następnie są adaptowane do innych zadań.	Atak na integralność	Model	Wprowadzenie złośliwych funkcji do wstępnie wytrenowanych modeli, co może wpłynąć na późniejsze, krytyczne operacje.	Walidacja i testowanie modeli przed ich wdrożeniem oraz stosowanie modeli od zaufanych dostawców.
Ataki przeprogramowujące	Reprogramming Attacks		Zmiana przeznaczenia modelu AI do realizacji innych zadań niż pierwotnie zamierzono.	Atak na integralność	Model	Zmiana działania modelu, co może prowadzić do wykonywania nieautoryzowanych zadań i destabilizacji operacji.	Ograniczanie dostępu do modelu oraz wprowadzenie mechanizmów monitorowania jego działania.
Zatrucie modelu	Model Poisoning	ML10: 2023	Manipulowanie parametrami modelu w celu wprowadzenia błędnych zachowań	Atak na integralność	Model	Model zwraca błędne wyniki, co może prowadzić do utraty poufnych danych i manipulacji decyzji	Regularizacja, zabezpieczenia kryptograficzne, projektowanie odpornych modeli
Ujawnienie poufnych informacji	Sensitive Information Disclosure	LLM06	Niewłaściwe zarządzanie wynikami AI, które może prowadzić do wycieku poufnych informacji.	Atak na poufność	Wyniki	Ujawnienie poufnych danych użytkowników lub wyników, co może skutkować naruszeniami prywatności i potencjalnymi sankcjami.	Maskowanie danych, stosowanie mechanizmów anonimizacji oraz bezpieczne przetwarzanie i usuwanie wyników modeli.

Inwersja modelu	Model Inversion Attack	ML03: 2023	Odtwarzanie modelu lub jego parametrów na podstawie generowanych wyników.	Atak na poufność	Wyniki	Uzyskanie dostępu do danych wejściowych, które mogą zawierać informacje wrażliwe o architekturze modelu, wykorzystanych algorytmach lub ich parametrach.	Techniki prywatności różnicowej oraz szyfrowanie danych wyjściowych modelu, które zmniejszają ryzyko wycieku informacji.
Niebezpieczne przetwarzanie wyników	Insecure Output Handling		Brak walidacji wyników AI, co może prowadzić do naruszenia bezpieczeństwa, np. przez wprowadzenie złośliwego kodu.	Atak na integralność	Wyniki	Nieprawidłowe przetwarzanie wyników modelu, co może prowadzić do wykonania złośliwego kodu i naruszenia integralności systemu.	Walidacja wyników generowanych przez model oraz wprowadzenie mechanizmów kontroli danych wyjściowych.
Nadmierna zależność	Overreliance	LLM09	Poleganie na wynikach AI bez odpowiedniej weryfikacji, co może prowadzić do błędnych decyzji i naruszeń bezpieczeństwa.	Atak na integralność	Wyniki	Błędne decyzje oparte na niezweryfikowanych wynikach modelu mogą prowadzić do poważnych konsekwencji finansowych lub operacyjnych.	Weryfikacja wyników modelu przez człowieka oraz ograniczenie automatycznego podejmowania decyzji w krytycznych procesach.
Atak na integralność wyników	Output Integrity Attack	ML09: 2023	Manipulowanie wynikami modelu w celu wprowadzenia błędów lub nieprawidłowych decyzji	Atak na integralność	Wyniki	Model zwraca fałszywe wyniki, co może prowadzić do nieprawidłowych decyzji lub wycieków danych	Walidacja wyników, monitorowanie wyjść modelu, zastosowanie filtrów i detekcji anomalii