DELL Technologies

10 najważniejszych obaw dotyczących cyberbezpieczeństwa w kontekście GenAI i LLM

Rozpocznij

Wprowadzenie

Sztuczna inteligencja (AI) rewolucjonizuje sposób działania organizacji, a generatywna sztuczna inteligencja (GenAI) i duże modele językowe (LLM) stają się krytycznymi obciążeniami roboczymi w nowoczesnych środowiskach korporacyjnych.

Podobnie jak w przypadku każdego innego obciążenia roboczego, aplikacje te mają własny zestaw złożoności i luk w zabezpieczeniach do rozwiązania. Ponieważ firmy nadal wdrażają sztuczną inteligencję w celu zwiększenia innowacyjności, efektywności i przewagi konkurencyjnej, zapewnienie bezpieczeństwa tych aplikacji staje się podstawową koniecznością. Dobra cyberhigiena jest podstawą zabezpieczenia każdego obciążenia roboczego. Podobnie jak w przypadku wszystkich obciążeń roboczych priorytetowo traktujesz bezpieczeństwo, ważne jest również przestrzeganie dobrej cyberhigieny w przypadku sztucznej inteligencji. Obejmuje to wdrażanie praktyk, takich jak odpowiednie aktualizowanie systemu, uwierzytelnianie wieloskładnikowe, dostęp oparty na rolach i segmentacja sieci. Środki te mają fundamentalne znaczenie, ale kluczem jest zrozumienie, w jaki sposób te możliwości pasują do konkretnej architektury i wykorzystania obciążenia roboczego.

W firmie Dell doskonale rozumiemy obciążenia robocze związane ze sztuczną inteligencją i wyjątkowe wyzwania związane z bezpieczeństwem, przed którymi stoi. Identyfikując potencjalne zagrożenia dla tych obciążeń, Dell może pomóc w opracowaniu skutecznej strategii bezpieczeństwa. Obejmuje to reagowanie na zagrożenia, takie jak zatruwanie danych szkoleniowych, kradzież modelu lub manipulacja nim, rekonstrukcja zestawu danych i inne.

Koncentrujemy się również na zarządzaniu wyzwaniami związanymi z danymi wejściowymi przesyłanymi do Twojego modelu sztucznej inteligencji, takimi jak zapobieganie ujawnianiu poufnych informacji, ograniczanie poruszania niebezpiecznych tematów lub wykazywania się stronniczością oraz zapewnianie zgodności z przepisami. Po stronie wyjściowej pomagamy rozwiązywać takie problemy, jak nadmierna zależność od modelu i ryzyko związane z zgodnością z przepisami.

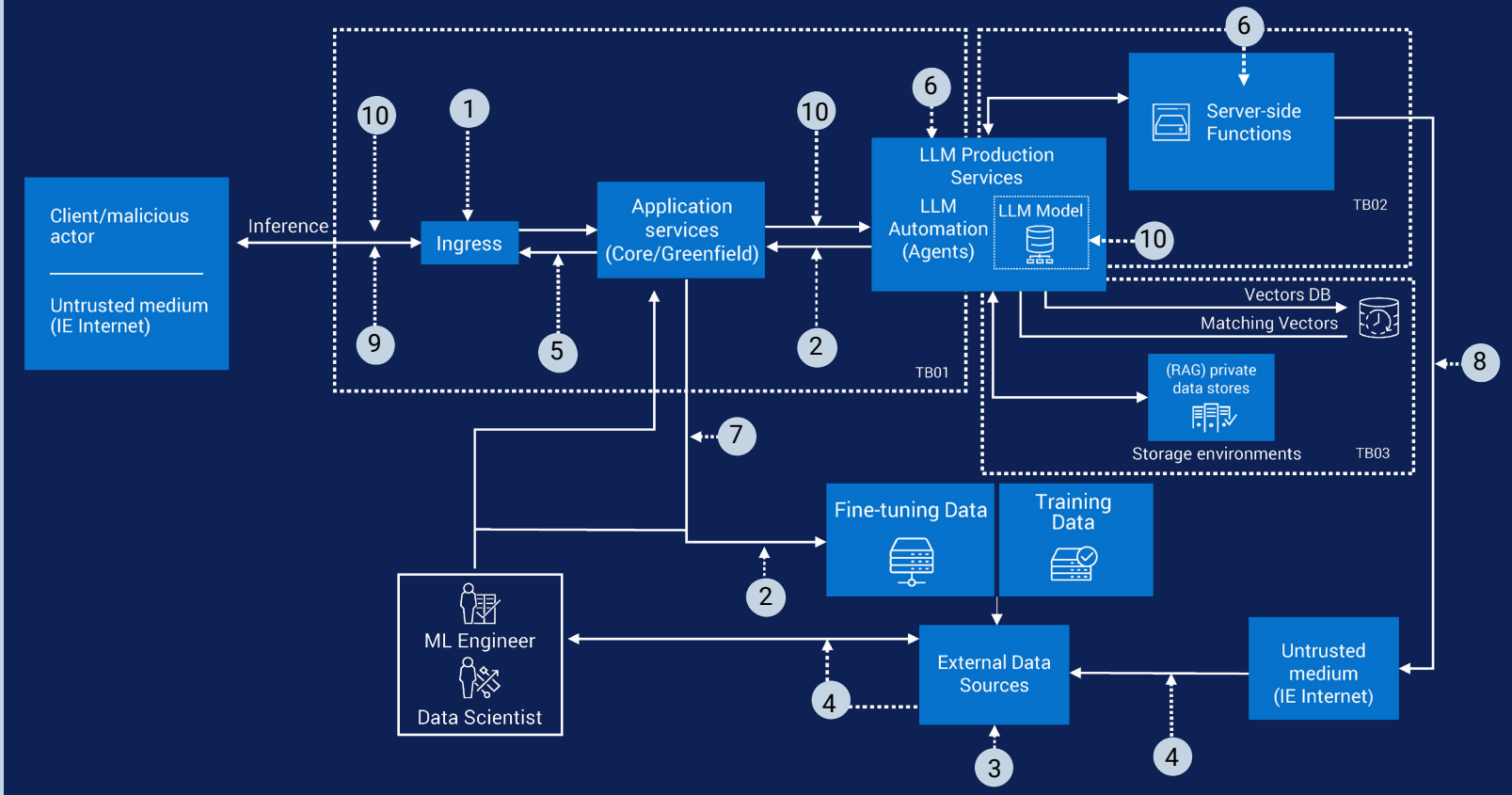
W firmie Dell umożliwiamy przedsiębiorstwom ograniczanie tych zagrożeń poprzez wykorzystanie istniejących rozwiązań w zakresie cyberbezpieczeństwa lub poszukiwanie nowych narzędzi i praktyk w celu ochrony ich systemów. Naszym celem jest upewnienie się, że bezpieczeństwo nie zakłóca innowacji. Dzięki zrozumieniu, w jaki sposób działają obciążenia robocze związane ze sztuczną inteligencją i jakie zagrożenia dla bezpieczeństwa z tego wynikają, możemy pomóc w zbudowaniu solidniejszego stanu bezpieczeństwa, czyniąc Twoje środowisko bardziej odpornym, a jednocześnie umożliwiając wprowadzanie innowacji bez obaw. Dzięki naszej wiedzy specjalistycznej pomagamy Ci bezpiecznie wykorzystać potencjał sztucznej inteligencji przy jednoczesnym zachowaniu solidnych zabezpieczeń.

10 najważniejszych obaw dotyczących cyberbezpieczeństwa w kontekście GenAI i LLM

Są to najważniejsze obawy dotyczące ochrony modeli GenAI/LLM, jak opisano w OWASP.

Kliknij problem, aby dowiedzieć się więcej:

1. Wstrzykiwanie promptów
2. Ujawnianie wrażliwych informacji
3. Łańcuch dostaw
4. Zatruwanie danych modelu
5. Nieprawidłowa obsługa danych wyjściowych
6. Nadmierna agencja
7. Wyciek promptu systemowego
8. Słabości wektora i osadzania
9. Dezinformacja
10. Nieograniczona konsumpcja

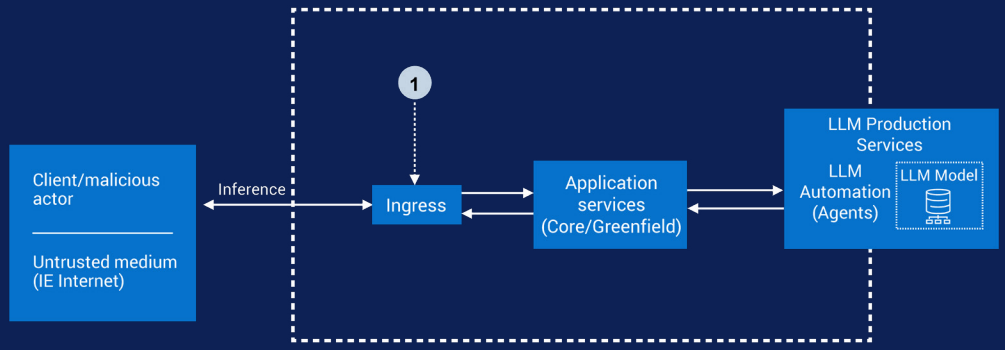


Problem nr 1: Wstrzykiwanie promptów

Strategie ograniczenia problemów ze wstrzykiwaniem promptów:

- **Sanityzacja danych i weryfikacja danych wejściowych:** Przetwarzaj dokładnie dane wejściowe użytkownika, aby usunąć szkodliwą zawartość. Użyj normalizacji i kodowania, aby zapobiec niewłaściwemu użyciu.
- **Przetwarzanie języka naturalnego (NLP) i podejścia oparte na uczeniu maszynowym:** Użyj NLP i uczenia maszynowego do wykrywania i blokowania zmanipulowanych lub złośliwych promptów.
- **Ustalenie wyraźnych granic formatowania danych wyjściowych i kontroli reakcji:** Ustaw ścisłe granice reakcji, aby zapewnić zgodność danych wyjściowych z zamierzonymi formatami i zapobiec nieautoryzowanym działaniom. Użyj filtrowania promptów i weryfikacji odpowiedzi, aby zachować integralność.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Inżynieria bezpiecznych promptów:** Korzystaj z bezpiecznego projektowania i analizy promptów w ramach ogólnego bezpieczeństwa oprogramowania w celu ochrony przetwarzania danych wejściowych.
- **Weryfikacja modelu** Regularnie weryfikuj modele ML, aby upewnić się, że nie zostały one naruszone przed wdrożeniem, chroniąc ich dokładność i integralność.
- **Filtrowanie, ranking i walidacja odpowiedzi:** Analizuj i twórz ranking promptów w celu zapewnienia, że przetwarzane są tylko bezpieczne dane wejściowe. Weryfikuj odpowiedzi, aby zapobiec niewłaściwemu użyciu.
- **Kontrole odporności:** Przeprowadzaj regularne oceny w celu identyfikacji i naprawiania luk w zabezpieczeniach, zapewniając bezpieczeństwo i niezawodność sztucznej inteligencji.

Powrót do schematu architektury



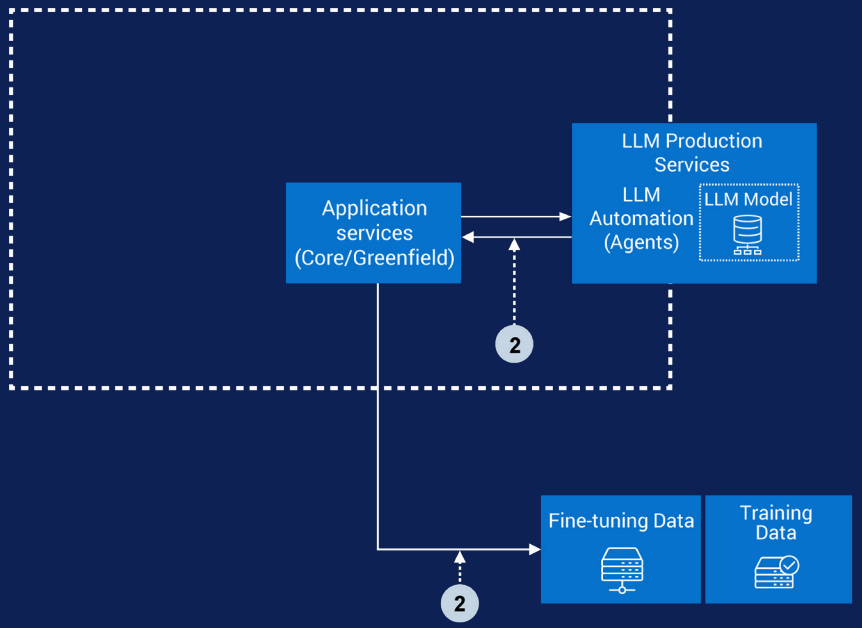
Wstrzykiwanie promptów to nowe wyzwanie w świecie generatywnej sztucznej inteligencji (GenAI), w którym złośliwe dane wejściowe są tworzone w celu manipulowania zachowaniem modelu lub naruszenia jego integralności. Ataki te wykorzystują luki w zabezpieczeniach w sposobie przetwarzania przez systemy AI i reagowania na dane wejściowe użytkowników, co może prowadzić do nieautoryzowanych działań, dezinformacji lub narażenia wrażliwych danych. W miarę jak GenAI staje się coraz bardziej zintegrowana z krytycznymi przepływami pracy, rozwiązanie tych zagrożeń ma kluczowe znaczenie dla utrzymania zaufania i bezpieczeństwa

Problem nr 2: Ujawnianie wrażliwych informacji

Strategie ograniczania ujawniania wrażliwych informacji:

- **Sanityzacja danych i weryfikacja danych wejściowych:** Przetwarzaj dokładnie dane wejściowe użytkownika, aby usunąć szkodliwą zawartość. Użyj normalizacji i kodowania, aby zapobiec niewłaściwemu użyciu.
- **Korzystanie z szyfrowania homomorficznego**, aby bezpiecznie przetwarzać wrażliwe informacje bez ujawniania ich zawartości. Zapewnia to, że nawet podczas korzystania z danych pozostają one zaszyfrowane i chronione przed naruszeniami.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Wykorzystanie bezpiecznych interfejsów API i interfejsów systemowych** do interakcji z danymi sztucznej inteligencji, rutynowo przeglądając konfiguracje w celu zminimalizowania narażenia na zagrożenia i powierzchni ataku.
- **Zabezpieczenie gromadzenia i przechowywania danych oraz powiązanych polityk**, a zarazem egzekwowanie kompleksowych zasad ochrony danych i zarządzania nimi, które gwarantują zgodność z przepisami i minimalizują ryzyko związane z naruszeniem danych.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Bezpieczny rozwój, konfiguracja i audyty:** Stosuj praktyki bezpiecznego kodowania, korzystaj z narzędzi do automatycznego zarządzania konfiguracją oraz przeprowadzaj regularne przeglądy, audyty i aktualizacje w celu zapewnienia bezpieczeństwa i aktualności konfiguracji systemów sztucznej inteligencji.
- **Edukacja użytkowników i świadomość w zakresie bezpieczeństwa:** Zapewnij użytkownikom i administratorom ciągłe szkolenia z zakresu świadomości w zakresie zabezpieczeń właściwych dla sztucznej inteligencji w celu zmniejszenia ryzyka niebezpiecznych przypadków użycia i przypadkowego ujawnienia danych.

Powrót do schematu architektury



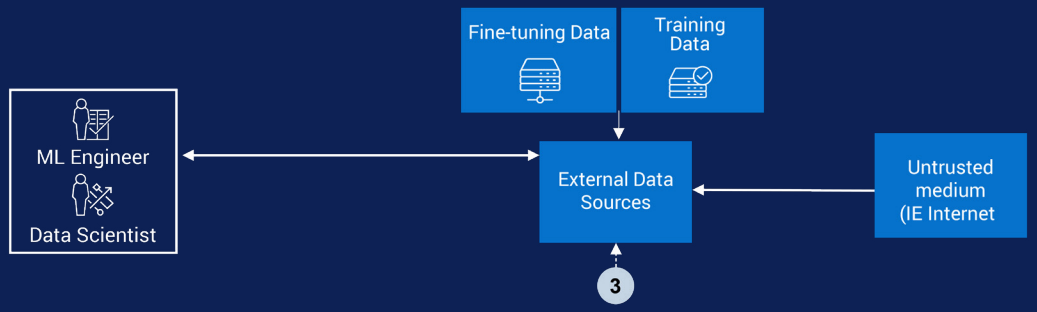
Generatywna sztuczna inteligencja wniosła niesamowity postęp, ale wiąże się również ze znacznym ryzykiem, zwłaszcza niezamierzonym narażeniem wrażliwych informacji. Niezależnie od tego, czy chodzi o dane osobowe, czy o zastrzeżone dane biznesowe, niewłaściwe użycie lub niewłaściwe postępowanie z narzędziami GenAI może prowadzić do wycieków danych, niezgodności z przepisami lub uszczerbku na reputacji. To sprawia, że organizacje muszą zrozumieć te zagrożenia i proaktywnie zająć się nimi, aby zapewnić bezpieczne wdrażanie i korzystanie z systemów sztucznej inteligencji.

Problem nr 3: Luki w łańcuchu dostaw

Strategie ograniczania luk w łańcuchu dostaw:

- **Sprawdzanie dostawców i zagwarantowanie zgodności z praktykami zabezpieczenia łańcucha dostaw:** Oceniaj dostawców i zawieraj umowy, które nadają priorytet bezpieczeństwu łańcucha dostaw.
- **Wdrożenie zestawienia materiałowego oprogramowania:** Monitoruj i weryfikuj pochodzenie komponentów oprogramowania, zapewniając przejrzystość i zmniejszając ryzyko naruszenia kodu.
- **Weryfikacja modelu** Regularnie weryfikuj modele ML, aby upewnić się, że nie zostały one naruszone przed wdrożeniem, chroniąc ich dokładność i integralność.
- **Uruchamianie kontenerów i podów z minimalnymi uprawnieniami** Pozwala to zredukować potencjalne skutki w przypadku naruszenia bezpieczeństwa i ograniczyć nieautoryzowany dostęp.
- **Wdrażanie zapór** Zablokuj niepotrzebną łączność sieciową, zmniejszając narażenie na potencjalne zagrożenia i ograniczając możliwości atakujących.
- **Ochrona danych i adnotacji** Zabezpiecz dane i powiązane adnotacje, aby zapobiec ingerencjom, nieautoryzowanemu dostępowi i uszkodzeniu krytycznych informacji.
- **Bezpieczny sprzęt** Korzystaj ze sprzętu zweryfikowanego pod kątem bezpieczeństwa, aby zapobiec lukom w zabezpieczeniach, które mogą wynikać z ataków sprzętowych, zapewniając solidny fundament infrastruktury.
- **Bezpieczne komponenty oprogramowania uczenia maszynowego** Wykorzystuj zaufane i zweryfikowane komponenty oprogramowania uczenia maszynowego, aby zminimalizować luki w zabezpieczeniach i zwiększyć ogólne bezpieczeństwo przepływów pracy związanych z uczeniem maszynowym.
- **Bezpieczny rozwój, konfiguracja i audyty:** Stosuj praktyki bezpiecznego kodowania, korzystaj z narzędzi do automatycznego zarządzania konfiguracją oraz przeprowadzaj regularne przeglądy, audyty i aktualizacje w celu zapewnienia bezpieczeństwa i aktualności konfiguracji systemów sztucznej inteligencji.

Powrót do schematu architektury



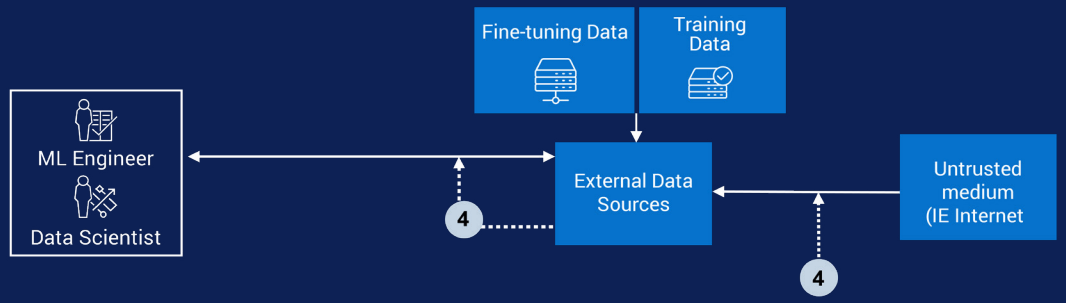
Poznaj luki w zabezpieczeniach łańcucha dostaw rozwiązań LLM, które mogą mieć wpływ na kluczowe elementy, takie jak wstępnie wytrenowana integralność modelu i adaptery innych firm. Systemy sztucznej inteligencji polegają zarówno na sprzęcie, jak i oprogramowaniu, które mogą zostać naruszone na długo przed wdrożeniem. Przeciwnicy mogą wykorzystać słabe punkty na różnych etapach łańcucha dostaw uczenia maszynowego, atakując sprzęt GPU, dane i ich adnotacje, elementy oprogramowania ML, a nawet sam model. Naruszając te unikatowe części, atakujący mogą uzyskać początkowy dostęp do systemów, co stwarza poważne ryzyko dla bezpieczeństwa i integralności. Zrozumienie tych luk w zabezpieczeniach i ich ograniczenie ma kluczowe znaczenie dla tworzenia solidnych, bezpiecznych rozwiązań AI.

Problem nr 4: Zatruwanie danych modelu

Strategie łagodzenia skutków zatrucia danymi modelu:

- **Wykrywanie anomalii i weryfikacja danych podczas szkolenia** pozwala identyfikować i rozwiązywać niespójności danych oraz zapewnić, że do trenowania modelu wykorzystywane są wyłącznie czyste, wysokiej jakości dane.
- **Izolowanie środowisk podczas faz dostrajania** w celu uniemożliwienia uzyskania nieautoryzowanego dostępu lub zanieczyszczenia modelu w trakcie krytycznych etapów jego rozwoju.
- **Weryfikacja modelu** Regularnie weryfikuj modele ML, aby upewnić się, że nie zostały one naruszone przed wdrożeniem, chroniąc ich dokładność i integralność.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Sanityzacja danych i sprawdzanie danych wejściowych:** Przetwarzaj dokładnie dane wejściowe użytkownika, aby usunąć szkodliwą zawartość. Użyj normalizacji i kodowania, aby zapobiec niewłaściwemu użyciu.
- **Bezpieczny rozwój, konfiguracja i audyty:** Stosuj praktyki bezpiecznego kodowania, korzystaj z narzędzi do automatycznego zarządzania konfiguracją oraz przeprowadzaj regularne przeglądy, audyty i aktualizacje w celu zapewnienia bezpieczeństwa i aktualności konfiguracji systemów sztucznej inteligencji.
- **Kontrole odporności:** Przeprowadzaj regularne oceny w celu identyfikacji i naprawiania luk w zabezpieczeniach, zapewniając bezpieczeństwo i niezawodność sztucznej inteligencji.
- **Wdrożenie segmentacji sieci** w celu ograniczenia dostępu do niezabezpieczonych interfejsów i krytycznych komponentów systemu.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.

Powrót do schematu architektury



Zatruwanie danych modelu to zagrożenie bezpieczeństwa w cyklu eksploatacji sztucznej inteligencji, w którym atakujący celowo zanieczyszczają dane szkoleniowe uszkodzonymi, wprowadzającymi w błąd lub złośliwymi danymi wejściowymi. Ryzyko to może mieć wpływ na kluczowe elementy — od gromadzenia surowych danych i adnotacji po tworzenie i integrację zestawów danych używanych do uczenia maszynowego lub dużych modeli językowych. Niezawodność systemów sztucznej inteligencji zależy od integralności ich źródeł danych, które mogą być narażone na manipulacje przed szkoleniem, podczas przetwarzania wstępnego lub za pośrednictwem zewnętrznych kanałów przepływu danych.

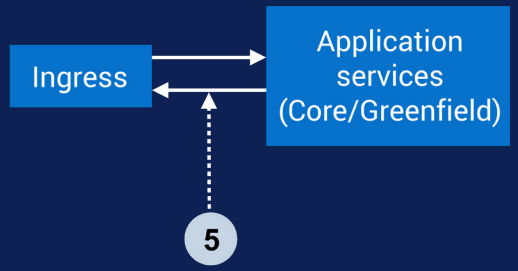
Atakujący wykorzystują zatruwanie danych w celu obniżenia dokładności modelu, wprowadzenia luk w zabezpieczeniach lub wywołania szkodliwych danych wyjściowych. Atakując słabe punkty w zakresie pochodzenia danych, jakości adnotacji lub procesów przyswajania zestawów danych, przeciwnicy mogą osłabić bezpieczeństwo, zaufanie i odporność. Rozpoznawanie i ograniczanie tych zagrożeń zorientowanych na dane ma kluczowe znaczenie dla tworzenia solidnych, niezawodnych rozwiązań AI.

Problem nr 5: Nieprawidłowa obsługa danych wyjściowych

Strategie ograniczania nieprawidłowej obsługi danych wyjściowych:

- **Kodowanie danych wyjściowych zgodne z kontekstem:** Aby zapobiec lukom w zabezpieczeniach, które mogą np. umożliwić przeprowadzenie ataku polegającego na wstrzyknięciu, należy zawsze stosować techniki kodowania i znaków modyfikacji, które są specjalnie dostosowane do konkretnego kontekstu, w którym dane wyjściowe będą używane, takie jak HTML, SQL lub API.
- **Sanityzacja danych wyjściowych:** Postępuj zgodnie z rygorystycznymi praktykami weryfikacji i sanityzacji dla danych wyjściowych modeli zgodnie z wytycznymi Open Web Application Security Project (OWASP) Application Security Verification Standard (ASVS), aby zagwarantować bezpieczne, dalsze stosowanie i ograniczyć zagrożenie bezpieczeństwa.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Automatyczne testowanie bezpieczeństwa danych wyjściowych:** Przeprowadzaj regularne testy bezpieczeństwa przy użyciu zautomatyzowanych narzędzi w celu identyfikacji zagrożeń w danych wyjściowych, takich jak tworzenie skryptów między witrynami (XSS) lub luki umożliwiające przeprowadzenie ataku polegającego na wstrzyknięciu kodu, i proaktywnego rozwiązywanie takich zagrożeń.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Przegląd w modelu „człowiek w pętli”:** W przypadku zastosowań wysokiego ryzyka, takich jak finanse lub opieka zdrowotna, wymagany jest nadzór człowieka i przeprowadzony przez niego przegląd wyników generowanych przez modele w celu zapewnienia dokładności, zabezpieczeń i bezpieczeństwa.
- **Prywatność i zgodność z przepisami:** Zintegruj techniki ochrony prywatności w procesie wyjściowym i zapewnij zgodność z odpowiednimi przepisami i normami dotyczącymi bezpiecznego korzystania z poufnych informacji.

Powrót do schematu architektury



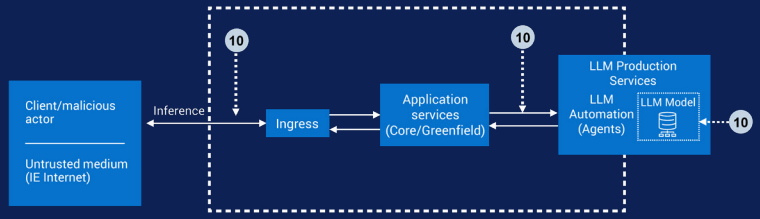
Niewystarczająca weryfikacja lub oczyszczanie danych wyjściowych modelu sztucznej inteligencji może prowadzić do poważnych zagrożeń bezpieczeństwa, w tym eskalacji uprawnień i naruszeń danych. Gdy modele sztucznej inteligencji generują dane wyjściowe, które nie są prawidłowo sprawdzane lub filtrowane, złośliwe podmioty mogą wykorzystać te luki w zabezpieczeniach w celu uzyskania nieautoryzowanego dostępu lub eskalacji uprawnień w systemie. Ten brak nadzoru może skutkować naruszonymi danymi, nieautoryzowanymi działaniami i znaczącymi naruszeniami bezpieczeństwa, co podkreśla znaczenie wdrożenia solidnych procesów walidacji i sanityzacji dla wszelkich danych wyjściowych generowanych przez sztuczną inteligencję.

Problem nr 6: Nadmierna agencja

Strategie łagodzenia nadmiernej agencji

- **Egzekwowanie najmniejszych uprawnień:** Przyznaj LLM i agentom podsystemowym tylko minimalne uprawnienia wymagane do wykonywania zamierzonych operacji i regularnego sprawdzania kontroli dostępu.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Ustalenie granic operacyjnych:** Jasno określ zakres LLM/agentów, czyli to, do czego mogą uzyskać dostęp lub co mogą wykonać.
- **Przegląd w modelu „człowiek w pętli”:** W przypadku zastosowań wysokiego ryzyka, takich jak finanse lub opieka zdrowotna, wymagany jest nadzór człowieka i przeprowadzony przez niego przegląd wyników generowanych przez modele w celu zapewnienia dokładności, zabezpieczeń i bezpieczeństwa.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Ograniczanie autonomii:** Ogranicz możliwości LLM, aby uniknąć nieograniczonego dostępu lub kontroli.
- **Bezpieczny rozwój, konfiguracja i audyty:** Stosuj praktyki bezpiecznego kodowania, korzystaj z narzędzi do automatycznego zarządzania konfiguracją oraz przeprowadzaj regularne przeglądy, audyty i aktualizacje w celu zapewnienia bezpieczeństwa i aktualności konfiguracji systemów sztucznej inteligencji.
- **Wdrażanie zapór** Zablokuj niepotrzebną łączność sieciową, zmniejszając narażenie na potencjalne zagrożenia i ograniczając możliwości atakujących.
- **Kontrole odporności:** Przeprowadzaj regularne oceny w celu identyfikacji i naprawiania luk w zabezpieczeniach, zapewniając bezpieczeństwo i niezawodność sztucznej inteligencji.

Powrót do schematu architektury



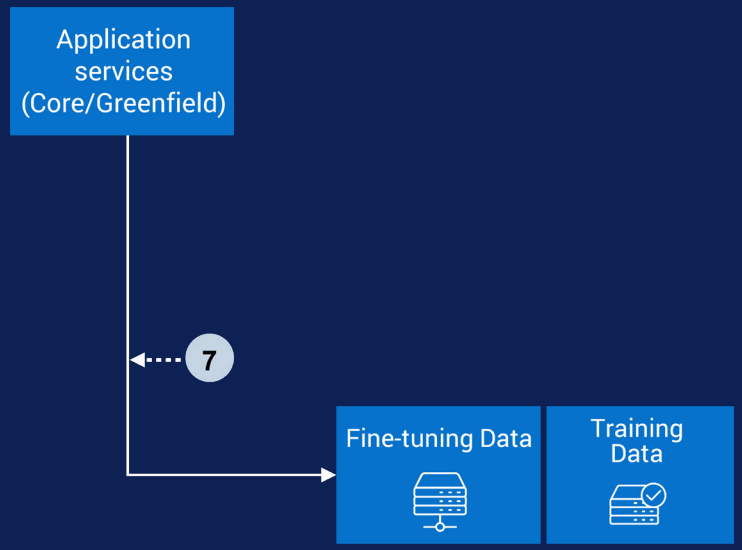
Przyznanie agentom lub wtyczkom sztucznej inteligencji nadmiernej autonomii lub niepotrzebnej funkcjonalności w przepływach pracy może stanowić poważne ryzyko. Gdy system sztucznej inteligencji otrzyma uprawnienia lub możliwości wykraczające poza to, co jest wymagane, zwiększa to prawdopodobieństwo niezamierzonych konsekwencji. Może się to zdarzyć, gdy systemy oparte na dużych modelach językowych (LLM) są projektowane z nadmiernymi uprawnieniami, umożliwiając im podejmowanie działań lub dostęp do informacji, których nie powinny. Takie przekroczenie może prowadzić do błędów, niewłaściwego wykorzystania danych lub nawet luk w zabezpieczeniach, podkreślając znaczenie ostrożnego ograniczania i monitorowania możliwości AI w celu zapewnienia bezpiecznego i odpowiedzialnego użytkowania.

Problem nr 7: Wyciek promptów

Strategie ograniczania wycieku promptów

- **Unikanie osadzania poufnych informacji w promptach** Nigdy nie podawaj poświadczeń, kluczy API ani zastrzeżonej logiki w promptach — zarządzaj nimi bezpiecznie poza systemem.
- **Oddzielenie elementów sterujących zabezpieczeniami od promptów** Obsługuj uwierzytelnianie, autoryzację i zarządzanie sesjami w logice aplikacji, a nie w promptach.
- **Weryfikacja danych wejściowych i wyjściowych:** Przeprowadź sanityzacje promptów i odpowiedzi dzięki solidnej weryfikacji, aby blokować podejrzane wzorce lub próby manipulacji.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Szyfrowanie i zabezpieczanie promptów** Zapisuj prompty i konfiguracje w zaszyfrowanej, bezpiecznej pamięci masowej, aby zapobiec nieautoryzowanemu dostępowi.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Regularny przegląd promptów** Okresowo przeglądaj i czyść prompty, aby usunąć wrażliwe dane i zapewnić zgodność z przepisami dotyczącymi bezpieczeństwa.
- **Testowanie z wykorzystaniem „czerwonego zespołu" w celu wykrycia luk:** przeprowadzaj testy z wykorzystaniem symulowanej grupy atakujących w celu zidentyfikowania i naprawienia słabych punktów zabezpieczeń w zarządzaniu promptami lub danymi wyjściowymi.
- **Odizolowanie promptów pochodzących z danych wejściowych użytkowników:** Projektuj systemy tak, aby zapobiec manipulacji lub ujawniania promptów przez zapytania użytkowników.
- **Egzekwowanie limitów prędkości** Ogranicz użycie interfejsu API, ogranicz podejrzaną aktywność i blokuj automatyczne ataki z promptami.

Powrót do schematu architektury



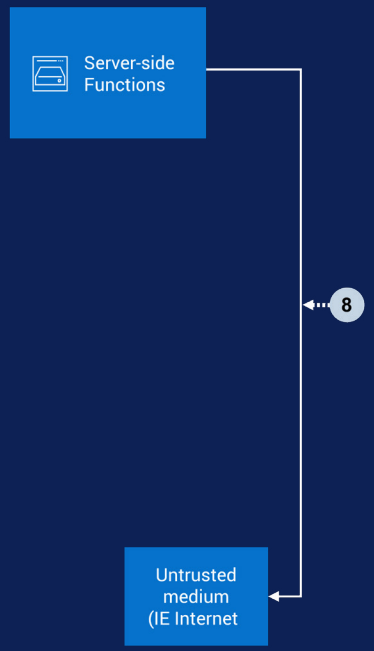
Atak polegający na wycieku promptu systemowego w dużym modelu językowym (LLM) lub systemie sztucznej inteligencji występuje, gdy osoba atakująca jest w stanie wyodrębnić lub wywnioskować ukryte instrukcje — „prompty systemowe" — które kierują zachowaniem modelu i ustalają granice operacyjne. Te prompty zazwyczaj nie są widoczne dla użytkowników końcowych, ponieważ zawierają podstawowe zasady, ograniczenia, a czasem wrażliwą logikę operacyjną. Poprzez specjalnie spreparowane dane wejściowe lub wykorzystanie luk w zabezpieczeniach osoba atakująca może nakłonić LLM do ujawnienia promptu systemowego, w całości lub w części. Jeśli prompt wycieknie, informacje te mogą być wykorzystane do odwrócenia ograniczeń, obejścia filtrów bezpieczeństwa lub opracowania nowych ataków ukierunkowanych, co ostatecznie zwiększa ryzyko ataku szybkim wstrzyknięciem kodu, eskalacji uprawnień lub niewłaściwego użycia modelu i systemów podrzędnych, które polegają na integralności danego modelu.

Problem nr 8: Słabości wektora i osadzania

Strategie łagodzenia słabości wektora i osadzania

- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Szyfrowanie** Zabezpieczaj dane wektorowe w transporcie i w spoczynku przy użyciu solidnych standardów szyfrowania, takich jak AES.
- **Zabezpieczona konfiguracja i monitorowanie** Wzmacniaj systemy, konfiguruj je bezpiecznie i stale monitoruj pod kątem błędów konfiguracji, nieautoryzowanego dostępu lub anomalii.
- **Zarządzanie lukami w zabezpieczeniach** Regularnie aktualizuj i wprowadzaj poprawki w całym oprogramowaniu, wszystkich programach zależnych i silnikach magazynu wektorowego w celu wyeliminowania zagrożeń bezpieczeństwa.
- **Sanityzacja danych i sprawdzanie danych wejściowych:** Przetwarzaj dokładnie dane wejściowe użytkownika, aby usunąć szkodliwą zawartość. Użyj normalizacji i kodowania, aby zapobiec niewłaściwemu użyciu.
- **Wykorzystanie bezpiecznych interfejsów API i interfejsów systemowych** do interakcji z danymi sztucznej inteligencji, rutynowo przeglądając konfiguracje w celu zminimalizowania narażenia na zagrożenia i powierzchni ataku.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Bezpieczny sprzęt** Korzystaj ze sprzętu zweryfikowanego pod kątem bezpieczeństwa, aby zapobiec lukom w zabezpieczeniach, które mogą wynikać z ataków sprzętowych, zapewniając solidny fundament infrastruktury.
- **Bezpieczny rozwój, konfiguracja i audyty:** Stosuj praktyki bezpiecznego kodowania, korzystaj z narzędzi do automatycznego zarządzania konfiguracją oraz przeprowadzaj regularne przeglądy, audyty i aktualizacje w celu zapewnienia bezpieczeństwa i aktualności konfiguracji systemów sztucznej inteligencji.

Powrót do schematu architektury



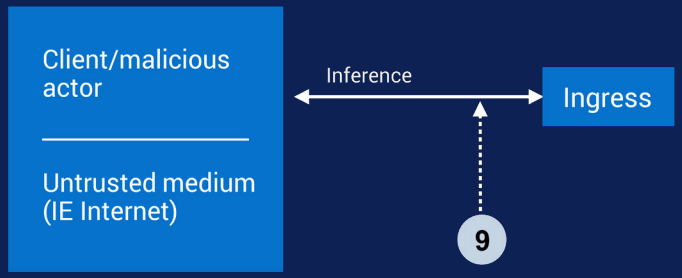
Atak wykorzystujący słabości wektora i osadzania w dużym modelu językowym (LLM) lub systemie sztucznej inteligencji — szczególnie tych korzystających z techniki Retrieval Augmented Generation (RAG) — bierze na cel luki w sposobie kodowania, przechowywania i pobierania informacji jako wektorów numerycznych i osadzeń. Słabe punkty tych mechanizmów mogą być wykorzystywane poprzez złośliwe działania, takie jak osadzanie inwersji (rekonstrukcja wrażliwych danych z osadzeń), zatruwanie danych (wprowadzanie szkodliwych lub stronniczych treści w celu manipulowania zachowaniem modelu), nieautoryzowany dostęp do baz danych wektorowych (prowadzący do wycieków danych) lub manipulację danymi wyjściowymi pobierania. Ataki te zagrażają prywatności, integralności i niezawodności, umożliwiając atakującym ujawnienie poufnych informacji, zmianę danych wyjściowych lub osłabienie zaufania użytkowników do aplikacji opartych na sztucznej inteligencji. Prawidłowa kontrola dostępu, weryfikacja danych, szyfrowanie i bieżące monitorowanie mają kluczowe znaczenie dla ochrony przed tymi ewoluującymi zagrożeniami.

Problem nr 9: Dezinformacja

Strategie ograniczania dezinformacji

- **Retrieval-Augmented Generation (RAG) z autorytatywnymi źródłami:** Używaj techniki RAG do pobierania i integrowania informacji ze zweryfikowanych, zaufanych baz danych oraz repozytoriów wiedzy, aby ograniczyć halucynacje sztucznej inteligencji.
- **Dostrajanie i kalibracja danych wyjściowych modeli:** Dostrajaj modele przy użyciu różnych zestawów danych i stosuj techniki minimalizujące stronniczość czy szerzenie dezinformacji.
- **Zautomatyzowane sprawdzanie faktów:** Porównuj wyniki z wiarygodnymi źródłami i automatycznie oznaczaj fałszywe informacje.
- **Monitorowanie niepewności:** Oznaczaj odpowiedzi o niskiej wiarygodności, aby w krytycznych przypadkach były one poddawane przeglądowi przez człowieka.
- **Przegląd w modelu „człowiek w pętli”:** W przypadku zastosowań wysokiego ryzyka, takich jak finanse lub opieka zdrowotna, wymagany jest nadzór człowieka i przeprowadzony przez niego przegląd wyników generowanych przez modele w celu zapewnienia dokładności, zabezpieczeń i bezpieczeństwa.
- **Opinie użytkowników:** Umożliwiaj użytkownikom zgłaszania błędów w celu ciągłego doskonalenia modelu i szybkiej korekty ścieżek dezinformacji.
- **Ograniczenia dostępu i nadzór człowieka:** W celu ograniczenia dostępu należy zastosować kontrolę dostępu opartą na rolach (RBAC), uwierzytelnianie wieloskładnikowe (MFA) i zarządzanie tożsamością. Wykorzystaj ludzki przegląd przy podejmowaniu krytycznych decyzji.
- **Bezpieczny rozwój, konfiguracja i audyty:** Stosuj praktyki bezpiecznego kodowania, korzystaj z narzędzi do automatycznego zarządzania konfiguracją oraz przeprowadzaj regularne przeglądy, audyty i aktualizacje w celu zapewnienia bezpieczeństwa i aktualności konfiguracji systemów sztucznej inteligencji.
- **Komunikacja o ryzyku:** Edukuj użytkowników na temat ograniczeń związanych ze sztuczną inteligencją i zachęcaj do niezależnej weryfikacji.
- **Zamierzone projektowanie interfejsu użytkownika i interfejsu API:** Wyróżniaj treści generowane przez AI i informuj użytkowników o odpowiedzialnym korzystaniu z nich.

Powrót do schematu architektury



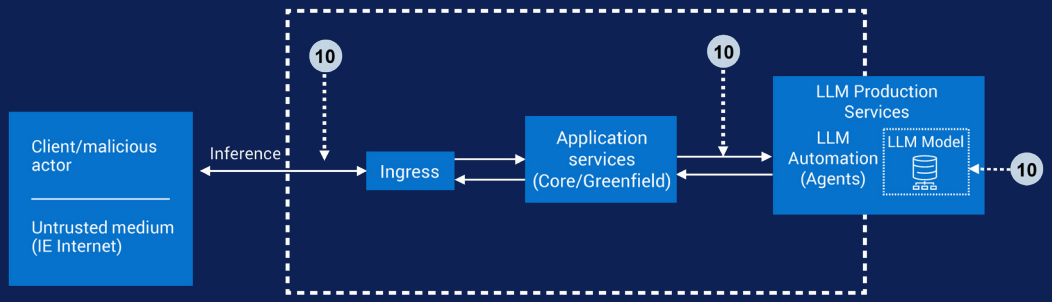
Atak dezinformacyjny na system LLM lub AI to celowe działanie mające na celu skłonienie modelu do generowania lub rozpowszechniania fałszywych, wprowadzających w błąd lub pozornie wiarygodnych, ale nieprawdziwych informacji poprzez jego dane wyjściowe. Luka ta wynika z kilku czynników: tendencji modelu do „halucynacji” (generowania sfabrykowanych, ale wiarygodnie brzmiących treści), uprzedzeń lub luk występujących w danych szkoleniowych oraz wpływu podpowiedzi przeciwników. Halucynacje pojawiają się, ponieważ LLM statystycznie generują tekst, który pasuje do wzorca, zamiast naprawdę rozumieć fakty, co prowadzi do odpowiedzi, które wydają się autorytatywne, ale są w rzeczywistości nieuzasadnione. Ryzyko związane z takimi atakami obejmuje naruszenia zabezpieczeń, uszczerbek na reputacji, a nawet odpowiedzialność prawną, zwłaszcza w środowiskach, w których użytkownicy nadmiernie polegają na odpowiedziach LLM bez weryfikowania ich dokładności lub istotności, potencjalnie osadzając błędy lub niepoprawne informacje w podejmowaniu istotnych decyzji i procesach.

Problem nr 10: Nieograniczona konsumpcja

Strategie nieograniczonej konsumpcji

- **Wymuszanie ograniczania częstotliwości korzystania i limitów na użytkownika:** Ustaw ścisłe limity żądań, tokenów lub danych na użytkownika, a następnie wykorzystaj klucz API lub aplikację, aby zapobiec nadużywaniu.
- **Wymagaj uwierzytelniania i segmentacji użytkowników** Użyj silnego uwierzytelniania (np. kluczy API, OAuth) i przypisz role lub poziomy do przetwarzania tylko autoryzowanych żądań.
- **Weryfikacja danych wejściowych i ograniczenia rozmiaru** Weryfikuj rozmiar i strukturę promptów, blokując lub skracając duże lub zniekształcone zapytania.
- **Stosowanie limitów czasu przetwarzania i ograniczenia zasobów:** Ustaw limity czasu i ograniczenia wykorzystania zasobów dla każdego żądania, aby uniknąć długotrwałych operacji i zużycia zasobów.
- **Wdrażanie inteligentnego buforowania i deduplikacji** odpowiedzi na zduplikowane lub podobne zapytania, aby ograniczyć niepotrzebne przetwarzanie.
- **Monitorowanie, rejestrowanie i wykrywanie anomalii:** Ciągłe monitorowanie i rejestrowanie działań systemu sztucznej inteligencji — przy użyciu rozwiązań takich jak MDR/XDR/SIEM — w celu szybkiego wykrywania, badania i reagowania na nieautoryzowany dostęp, anomalie lub wycieki danych.
- **Śledzenie budżetu i kontrola wydatków** Wykorzystuj pulpity i alerty do monitorowania kosztów i blokowania wykorzystania przy progach budżetowych.
- **Techniki piaskownicy i izolacji** uruchamiają obciążenia robocze w odizolowanych środowiskach z ograniczonymi uprawnieniami w celu zmniejszenia ryzyka.
- **Ograniczenie głębi wywołań i etapów rozmów:** Nałóż limity na rekurencyjne wywołania lub etapy konwersacji, aby zapobiec niewłaściwemu wykorzystaniu.
- **Zastosowanie modelu warstwowego lub przydzielanie zasobów:** Przekieruj żądania o wysokim priorytecie do modeli klasy premium, a ruch o niskim priorytecie do bardziej ekonomicznych modeli.

Powrót do schematu architektury



Zagrożenie związane z nieograniczonym wykorzystaniem treści produkowanych przez modele LLM lub systemy sztucznej inteligencji odnosi się do luki w zabezpieczeniach, w której aplikacja umożliwia użytkownikom — atakującym lub nie — przesyłanie nadmiernych, niekontrolowanych żądań wnioskowania lub promptów bez skutecznego ograniczania częstotliwości ich przesyłania, uwierzytelniania lub ogólnych ograniczeń użytkowania. Wnioskowanie w modelach LLM jest kosztowne pod względem obliczeniowym, więc ten brak kontroli można wykorzystać na kilka sposobów: atakujący mogą powodować odmowę usługi (DoS) przez przeciążenie zasobów systemowych, generować nieprzewidziane straty ekonomiczne we wdrożenia w modelu płatności za wykorzystanie lub wdrożeniach hostowanych w chmurze lub systematycznie wysyłać zapytania do modelu w celu sklonowania jego zachowania, a w konsekwencji — kradzieży własności intelektualnej. Konsekwencje obejmują zakłócenia w świadczeniu usług, obniżoną wydajność innych użytkowników, obciążenie finansowe i zwiększone ryzyko wycieku wrażliwych modeli. Zasadniczo nieograniczone zużycie występuje, gdy wykorzystanie zasobów nie jest prawidłowo zarządzane, co naraża aplikacje oparte na LLM na przypadkowe i celowe wykorzystanie.

Dlaczego warto wybrać rozwiązania Dell w zakresie bezpieczeństwa

Dell pomaga organizacjom zabezpieczyć modele sztucznej inteligencji i LLM dzięki kompleksowemu podejściu obejmującemu sprzęt, oprogramowanie i usługi zarządzane. Zabezpieczenia są wbudowane w łańcuch dostaw, urządzenia, infrastrukturę, dane i aplikacje, wszystkie zgodne z zasadami Zero Trust. W całym portfolio Dell oferuje rozwiązania, które mają na celu poprawę cyberhigieny poprzez takie funkcje jak MFA, RBAC, najmniejszy poziom uprawnień i ciągła weryfikacja. To kompleksowe podejście „bezpieczne z założenia" zapewnia organizacjom możliwość wprowadzania innowacji dzięki sztucznej inteligencji i LLM, minimalizując ryzyko związane z kradzieżą modeli, wyciekiem danych, atakami adwersarskimi i innymi zaawansowanymi zagrożeniami cybernetycznymi.

Łańcuch dostaw

Bezpieczny łańcuch dostaw firmy Dell zapewnia podstawową ochronę modeli sztucznej inteligencji i modeli LLM poprzez osadzenie zabezpieczeń na każdym etapie opracowywania, produkcji i dostawy produktów. Dzięki kryptograficznie podpisanym aktualizacjom systemu BIOS i oprogramowania wewnętrznego, Secured Component Verification (SCV), zestawieniom materiałowym oprogramowania skoncentrowanym na sztucznej inteligencji (SBOM), monitorowaniu linii zestawów danych, zintegrowanemu oprogramowaniu zabezpieczającemu oraz konfiguracji i rygorystycznym ocenom ryzyka dostawców zgodnym z globalnymi standardami firma Dell minimalizuje ryzyko związane z manipulacjami, nieautoryzowanym dostępem i atakami w łańcuchu dostaw. Zapewnia to organizacjom możliwość wdrażania zaufanych, odpornych obciążeń roboczych sztucznej inteligencji przy pełnej przejrzystości, integralności i zachowaniu zgodności z przepisami.

Komputery z AI

Dell oferuje podstawowe zabezpieczenia obciążeń roboczych związanych ze sztuczną inteligencją na urządzeniu. Dell Trusted Devices---najbezpieczniejsze komercyjne komputery AI na świecie*---są zaprojektowane z myślą o bezpieczeństwie. Bezpieczeństwo łańcucha dostaw minimalizuje ryzyko wystąpienia luk w zabezpieczeniach produktów i ingerencji w nie. Unikatowe zabezpieczenia wbudowane bezpośrednio w sprzęt i oprogramowanie wewnętrzne zapewniają ochronę komputera i użytkownika końcowego podczas użytkowania. Dell SafeBIOS zapewnia głęboką widoczność na poziomie systemu BIOS i wykrywanie manipulacji, a Dell SafeID zwiększa bezpieczeństwo poświadczeń i umożliwia uwierzytelnianie bez hasła. Oprogramowanie dla partnerów zapewnia zaawansowaną ochronę w środowiskach punktów końcowych, sieci i chmury.

Cyberodporność

Rozwiązania cyberbezpieczeństwa Dell PowerProtect zabezpieczają dane AI za pomocą zaszyfrowanych, niezmiennych kopii zapasowych, szybkiego przywracania i izolowanych skarbców odzyskiwania cybernetycznego. Funkcje te zapobiegają zniszczeniu, ograniczają wpływ złośliwych aktualizacji oraz zapewniają zgodność z przepisami i umożliwiają odzyskanie danych po ataku.

Serwery

Serwery PowerEdge wyróżniają się obsługą poufnego przetwarzania, które izoluje i zabezpiecza prompty i osadzenia AI/LLM, zaufanymi rozwiązaniami RAG (Retrieval-Augmented Generation) zakotwiczonymi w autorytatywnych źródłach, a także modelami MFA, RBAC, Silicon Root of Trust, podpisanym oprogramowaniem wewnętrznym i ciągłym monitorowaniem w celu ochrony krytycznych obciążeń roboczych związanych ze sztuczną inteligencją.

Pamięć masowa

Portfolio pamięci masowych Dell zapewnia bezpieczne, szyfrowane przechowywanie wrażliwych danych AI z wykorzystaniem solidnego szyfrowania AES-256 dla danych w spoczynku i w trakcie transmisji. Zaawansowane szyfrowanie zaprojektowane z myślą o odporności na przyszłe zagrożenia kwantowe jest dostępne w wybranych ofertach. Oferta obejmuje szybką wydajność NVMe, moduły szyfrowania zgodne ze standardem FIPS do zabezpieczania danych — w tym tych wykorzystywanych w obciążeniach roboczych związanych ze sztuczną inteligencją — niezmienne migawki i odizolowane magazyny Cyber Recovery w celu przeciwdziałania atakom typu ransomware. Architektura w modelu „zero trust", bezpieczeństwo łańcucha dostaw i możliwości inspekcji odpornej na manipulacje usprawniają zarządzanie. Wbudowane wykrywanie anomalii i modele AIOps ML chronią obciążenia robocze bez wykorzystywania danych klienta do szkolenia, minimalizując w ten sposób ryzyko ataku oparte na danych wejściowych.

AIOps

Dell AIOps zapewnia zautomatyzowane, ciągłe monitorowanie w celu wykrywania błędnych konfiguracji, luk w zabezpieczeniach (w tym CVE) i wspiera świadomość ryzyka w łańcuchu dostaw wpływającego na obciążenia robocze związane ze sztuczną inteligencją/LLM. Skanowanie CVE w czasie rzeczywistym, inteligentne alerty i pulpity nawigacyjne oparte na sztucznej inteligencji umożliwiają szybką interwencję poprzez oznaczanie anomalii i śledzenie przepływów pracy związanych z rozwiązywaniem problemów. Wbudowane funkcje zgodności z przepisami, kontrola dostępu oparta na rolach i zautomatyzowane raportowanie pomagają utrzymać bezpieczeństwo operacji w obciążeniach roboczych, podczas gdy bezproblemowa integracja EDR/XDR i analiza operacyjna oparta na sztucznej inteligencji — w tym możliwości generatywne w obsługiwanych rozwiązaniach — jeszcze bardziej zwiększają wydajność rozwiązań IT.

Sieci

Rozwiązania Dell Networking chronią środowiska sztucznej inteligencji/LLM dzięki solidnej segmentacji sieci, minimalizując ruch boczny. Zaszyfrowane ścieżki sieciowe i zintegrowane mechanizmy sterowania zaporą blokują nieautoryzowany dostęp do danych sztucznej inteligencji.

Usługi AI związane z bezpieczeństwem i odpornością

Usługi Dell AI związane z bezpieczeństwem i odpornością są zaprojektowane tak, aby eliminować nowe zagrożenia związane z integracją AI w Twojej organizacji. Nasze usługi zostały opracowane tak, aby współpracowały z zespołami w Twojej firmie już od momentu wdrożenia AI. Dają one dostęp do specjalistycznej wiedzy, która pomoże Ci przeprowadzić strategiczne planowanie, wdrożenie rozwiązania i korzystanie z zarządzanych usług zabezpieczeń, zmniejszając obciążenie operacyjne i pozwalając Ci bezpiecznie wprowadzać innowacje przy pomocy AI. Każda z nich jest dostosowana, aby pomóc organizacjom w radzeniu sobie ze zmieniającym się ryzykiem związanym ze sztuczną inteligencją i optymalizacji bezpiecznych wdrożeń sztucznej inteligencji.

Dell AI Factory

Zintegrowane portfolio specjalnie zaprojektowanych zabezpieczeń, takich jak bezpieczny łańcuch dostaw Dell, możliwości w modelu „zero trust" umożliwiające wymuszenie dostępu o najmniejszym poziomie uprawnień oraz rozwiązania MDR oparte na sztucznej inteligencji — każdy z tych elementów ma na celu zagwarantowanie bezpieczeństwa modelu.

* Na podstawie wewnętrznej analizy firmy Dell, październik 2024 r. (Intel) i marzec 2025 r. (AMD). Dotyczy komputerów z procesorami Intel i AMD. Nie wszystkie funkcje są dostępne na wszystkich komputerach osobistych. W przypadku niektórych funkcji wymagany jest dodatkowy zakup. Komputery z procesorami Intel zostały zweryfikowane przez Principled Technologies, lipiec 2025 r.

Wnioski

Aby zbudować odporne struktury sztucznej inteligencji, niezbędne jest podejście oparte na współpracy między organizacjami a ekspertami ds. bezpieczeństwa. W miarę jak sztuczna inteligencja i modele LLM nadal zmieniają oblicze branż, niezwykle ważne jest uwzględnienie ryzyka, jakie niosą ze sobą, w tym wyzwań związanych z bezpieczeństwem danych, integralnością modeli i zgodnością z przepisami. Organizacje powinny skupić się na strategiach zapobiegawczych, które uwzględniają kwestie bezpieczeństwa na każdym etapie wdrażania sztucznej inteligencji.

Dell Technologies jest zaufanym partnerem w tej misji, oferując kompleksowe dostosowywanie generatywnej sztucznej inteligencji, doradztwo w zakresie bezpieczeństwa i zintegrowane rozwiązania dostosowane do Twoich unikalnych potrzeb. Wykorzystując solidne rozwiązania Dell w zakresie cyberbezpieczeństwa, przedsiębiorstwa mogą skutecznie ograniczyć ryzyko związane ze sztuczną inteligencją i LLM, jednocześnie maksymalizując potencjał istniejących inwestycji w zabezpieczenia. Dell umożliwia organizacjom ochronę infrastruktury AI poprzez bezproblemową integrację zaawansowanych zabezpieczeń z ich obecnymi strukturami, zapewniając bezpieczne środowisko gotowe na przyszłość.

Dowiedz się, w jaki sposób kompleksowe rozwiązania Dell w zakresie sztucznej inteligencji mogą zabezpieczyć Twoje środowiska GenAI i LLM: Dell.com/CyberSecurityMonth