centralny ośrodek informatyki

COI-DDI.502.2.2024

Warszawa, dnia 17 maja 2024 roku

Pani
Wioletta Zwara
Sekretarz Komitetu Rady Ministrów
do spraw Cyfryzacji
Ministerstwo Cyfryzacji

Szanowna Pani Sekretarz,

zwracam się z prośbą o przekazanie wniosku o zaopiniowanie projektu informatycznego - System Dokumentacji Prawnej do podpisu Ministra Cyfryzacji oraz przekazanie podpisanego przez Ministra Cyfryzacji wniosku o zaopiniowanie projektu informatycznego - System Dokumentacji Prawnej wraz z załącznikiem „Opis założeń projektu informatycznego - System Dokumentacji Prawnej" do zaopiniowania przez Komitet Rady Ministrów do spraw Cyfryzacji.

Uprzejmie informuję, że Opis Założeń Projektu Informatycznego został uzupełniony na podstawie poniższych uwag Rady Architektury IT wskazanych w Karcie Oceny Projektu nr P376:

- wskazanie potencjalnego wykorzystania produktów projektu oraz modułów w innych obszarach administracji państwowej – uzupełnienie w punkcie 1.1 Identyfikacja problemu i potrzeb,
- brak skalowalnego KPI np. liczby przeprowadzonych spraw/dokumentów; w treści OZPI jest informacja o procesowaniu dziesiątek tysięcy orzeczeń i dokumentacji rocznie co mogłoby być ambitnym KPI – KPI 8. dodany w punkcie 2.1 Cele i korzyści wynikające z projektu,
- doszczegółowienie informacji nt. ponoszonych kosztów UX w wysokości 3,5 mln zł w sytuacji, gdy COI posiada dedykowany zespół UX – uzupełnienie w punkcie 4.2 Wykaz poszczególnych pozycji kosztowych,

- doszczegółowienie informacji nt. wykonanie audytu zgodności z WCAG - uzupełnienie w punktach: 2.4 Produkty końcowe projektu, 3. Kamienie milowe, 4.2 Wykaz poszczególnych pozycji kosztowych.

W uzupełnieniu, przekazujemy również wyjaśnienia adresujące uwagi Rady Architektury IT wskazane w ww. Karcie Oceny Projektu a dotyczące poniższych kwestii:

- przekazania informacji dot. modelu zarządzania informacją (uwzględnienie danych osobowych lub poufnych etc.) zarówno na etapie trenowania modelu, jak również na etapie użytkowania rozwiązania. W opisie prosimy uwzględnić aspekt wykorzystania chmury obliczeniowej (prywatnej lub publicznej),
- doszczegółowienie informacji o planowanym sposobie zabezpieczenia materiałów zastrzeżonych oraz wypracowanych prawach autorskich przy opracowaniu dedykowanych modeli językowych,
- doszczegółowienie informacji o procesie uwierzytelnianie Użytkowników w module logowania (WK czy coś innego).

Wyjaśnienia dot. kwestii AI i danych:

1. Pilotażowa wersja Systemu, na żadnym etapie projektu, nie będzie mieć dostępu do danych objętych tajemnicą skarbową oraz danych niejawnych.
2. Baza wiedzy pilotażowej wersji Systemu zawierać będzie dokumenty (oraz ich metadane) w wersji zanonimizowanej (w tym objęte tajemnicą PGRP, bez danych osobowych), dokumenty dostępne publicznie (np. akty prawne) oraz dokumenty przed poddaniem procesowi anonimizacji (w tym objęte tajemnicą PGRP).
3. Do komponentów AI przekazywane będą jedynie dane zanonimizowane oraz publiczne. Następnie na podstawie otrzymanych z modeli rezultatów, dane te będą uzupełniane/wzbogacane o pełne dane (w tym osobowe) przez aplikację backendową, w celu ich poprawnej prezentacji użytkownikowi poprzez funkcjonalność dekodującą. Dane zawierające dane osobowe nie będą zatem, na żadnym etapie projektu, przekazywane do jakiegokolwiek komponentu AI, a będą przetwarzane w pilotażowej wersji Systemu jedynie w części backendowej i frontendowej w celu zapewnienia założonych funkcjonalności systemu SDP.

4. W ramach projektu nie będziemy trenować modeli językowych LLM. Zakładamy wykorzystanie gotowych modeli, zatem dane PGRP nie będą przekazywane do zewnętrznych podmiotów w celu trenowania modeli.
5. Dopuszczamy, jedynie w razie potrzeby, dotrenowanie modeli LLM na danych dostępnych publicznie, w celu poprawy zdolności językowych modelu w obszarze żargonu prawniczego. Dane partnera pilotażowego, tj. PGRP nie będą w ogóle używane w tym procesie.
6. W celu wykonania rekomendacji dekretacji oraz biegu sprawy może wystąpić konieczność wykorzystania własnych algorytmów statystycznych na danych zanonimizowanych PGRP. Algorytmy oraz powstałe przy ich użyciu modele rezydować będą w ekosystemie aplikacji backendowej - w przestrzeni izolowanej dla PGRP. Model pomocniczy wytrenowany na danych stanowiących tajemnice PGRP używany i zarządzany będzie tylko w kontekście tego jednego podmiotu (prawa autorskie pozostają u klienta).

Wyjaśnienia dot. kwestii uwierzytelnienia użytkowników, bezpieczeństwa danych oraz infrastruktury chmurowej:

1. Pilotażowa wersja Systemu będzie zapewniać bezpieczny sposób logowania oraz zarządzania uprawnieniami użytkowników. Rozważamy możliwość integracji uwierzytelniania z Węzłem Krajowym, który jest rozwijany w COI. Alternatywnie możemy zastosować niezależną autentykację przy użyciu protokołu OAuth2.0 np. używając technologii KeyCloak. Dodatkowo w pilotażowej wersji Systemu będzie zdefiniowanych wiele poziomów uprawnień o różnych granularyzacjach. Możliwość zarządzania użytkownikami będzie odbywać się z poziomu dedykowanego panelu administracyjnego.
2. Bezpieczeństwo danych w przelocie (data in transit) będzie zabezpieczone poprzez wymuszenia komunikacji przez protokół HTTPS pomiędzy infrastrukturą PGRP a pilotażową wersją Systemu.
3. Dane w spoczynku (data at rest) - tj. bazy danych, serwery aplikacyjne itd. będą zabezpieczone poprzez zdefiniowane polityki dostępów pomiędzy komponentami (np. RBAC w Azure).
4. Wszystkie komponenty wewnętrzne pilotażowej wersji Systemu będą się komunikować między sobą w sieci prywatnej - bez możliwości dostępu do nich z poziomu publicznego internetu. Jedynym punktem infrastruktury wyeksponowanym na internet będzie brama,

odpowiednio zabezpieczona przed potencjalnymi atakami hakerskimi poprzez zaporę sieciową.

5. Realizacja projektu w chmurze publicznej umożliwia realizację pilotażowej wersji Systemu w założonej funkcjonalności ze względu na specyfikę obszaru Data & AI – jego dynamikę i stopień skomplikowania. Wiele gotowych, bardzo dużych komponentów np. modeli jest obecnie nieodtwarzalne na infrastrukturze on premise czy private cloud. Chmura publiczna pozwala natomiast bezpiecznie skorzystać z najnowszych, skalowalnych modeli LLM, co przełoży się na wysoką jakość rozwiązania.
6. Istotną korzyścią realizacji projektu w chmurze publicznej jest także uproszczenie realizacji założenia skalowalności i reużywalności pilotażowej wersji Systemu u innych podmiotów wykonujących funkcje publiczne (potencjalnych beneficjentów budowanego rozwiązania), w szczególności z uwagi na brak potrzeby planowania zapotrzebowania dotyczącego zakupu infrastruktury on-premise lub private cloud, czasu potrzebnego na realizację dostaw serwerów, co w konsekwencji przekłada się na czas dostarczenia rozwiązania.
7. Wykorzystanie chmury publicznej umożliwia, niskim nakładem, zapewnienie szyfrowania danych w spoczynku, wysoką dostępność pilotażowej wersji Systemu oraz, przez wykorzystanie backupów i polityk retencyjnych, zabezpieczenie danych przed ich utratą. Dodatkowo mamy pełną kontrolę nad miejscem wdrożenia modelu (wdrożenia na terenie EOG) oraz zarządzania danymi, które przetwarza w celach analitycznych.
8. Dzięki wykorzystaniu w COI granularnego zarządzania na poziomie Active Directory zapobiegamy nieuprawnionemu dostępowi do komponentów rozwiązania wewnątrz organizacji.

Załączniki:

1. Opis założeń projektu informatycznego - System Dokumentacji Prawnej.