Załącznik nr 1 do zapytania ofertowego nr 1/2026/FEPW

SZCZEGÓŁOWY OPIS PRZEDMIOTU ZAMÓWIENIA

1. Kontekst

Rudder AI jest platformą SaaS, w której meta-orkiestrator zarządza pracą wyspecjalizowanych agentów AI (analitycznego, prawnego, interpretacyjnego, decyzyjnego, walidacyjnego i raportującego). Platforma przetwarza dokumenty finansowe i podatkowe klientów, w tym dane osobowe, dlatego kluczowe są: lokalizacja danych w UE lub EOG, brak retencji i trenowania modeli na danych Zamawiającego, audytowalność oraz wysoka dostępność środowiska produkcyjnego.

2. Obszar A – Infrastruktura chmurowa

Element	Wymaganie minimalne
Lokalizacja	Wszystkie zasoby, dane, kopie zapasowe i logi wyłącznie w regionach UE lub EOG
Środowiska	3 odseparowane środowiska: produkcyjne, testowe, deweloperskie
Uruchamianie aplikacji	Kontenery z automatycznym skalowaniem horyzontalnym (min. 2 instancje na produkcji), kolejka zadań asynchronicznych
Baza danych	PostgreSQL w usłudze zarządzanej: na produkcji w konfiguracji wysokiej dostępności (replika w innej strefie dostępności), automatyczne kopie zapasowe z retencją min. 14 dni i odtwarzaniem do punktu w czasie; osobna instancja dla środowiska testowego
Wyszukiwanie semantyczne	Usługa wyszukiwania wektorowego i pełnotekstowego z indeksem o pojemności min. 2 GB
Magazyn danych	Min. 1 TB magazynu obiektowego z kopiami geo-redundantnymi w obrębie UE lub EOG, szyfrowanie danych w spoczynku i w tranzycie
Sieć i bezpieczeństwo	Brama aplikacyjna z zaporą WAF i siecią CDN, certyfikaty TLS, DNS, ruch wychodzący w zakresie niezbędnym do działania platformy
Klucze i sekrety	Usługa sejfu kluczy (key vault), skanowanie podatności zasobów chmurowych
Monitoring	Monitoring metryk i logów z alertami, retencja logów min. 30 dni
Poczta transakcyjna	Usługa wysyłki wiadomości e-mail z aplikacji (powiadomienia, raporty) z obsługą SPF/DKIM

3. Obszar B – Modele AI przez API

Wymagania wspólne: przetwarzanie danych w UE lub EOG; brak retencji danych wejściowych i wyjściowych po przetworzeniu (z wyjątkiem krótkotrwałego monitoringu nadużyć, jeżeli jest wymagany przez dostawcę i nie dłużej niż 30 dni); brak wykorzystania danych do trenowania modeli; umowa powierzenia przetwarzania danych.

Model	Wymagania funkcjonalne	Wolumen miesięczny
Model główny	Duży model językowy ogólnego przeznaczenia najwyższej klasy dostępnej u danego dostawcy; okno kontekstu min. 128 tys. tokenów; wywoływanie funkcji (tool use) i odpowiedzi w ustrukturyzowanym formacie (JSON); wysoka jakość pracy na tekstach prawnych w języku polskim	100 mln tokenów wejściowych i 20 mln tokenów wyjściowych

Model	Wymagania funkcjonalne	Wolumen miesięczny
Model pomocniczy	Szybszy model językowy niższej klasy do klasyfikacji, ekstrakcji i zadań pomocniczych; okno kontekstu min. 128 tys. tokenów	200 mln tokenów (łącznie wejście i wyjście)
Model wektoryzacji	Model embeddingów obsługujący język polski	20 mln tokenów miesięcznie oraz jednorazowo ok. 1,6 mld tokenów na początkowe zasilenie bazy wiedzy

Zamawiający dopuszcza dostarczenie modeli różnych dostawców w ramach jednego dostępu. Wykonawca zapewnia możliwość zmiany modelu na nowszą wersję o parametrach nie gorszych bez zmiany wynagrodzenia (pkt 15.2 zapytania).

4. Obszar C – Narzędzia środowiskowe

Narzędzie	Wymaganie minimalne
Repozytoria kodu z CI/CD	Prywatne repozytoria, przeglądy kodu, potoki CI/CD – dla 6 użytkowników
Monitorowanie błędów aplikacji	Zbieranie i grupowanie błędów front-end i back-end, alerty
Skanowanie bezpieczeństwa kodu i zależności	Statyczna analiza kodu i skanowanie podatności zależności – dla 4 użytkowników
Menedżer haseł i sekretów	Wspólny sejf z kontrolą dostępu – dla 8 użytkowników
Tożsamość	Jednokrotne logowanie (SSO) lub logowanie dwuskładnikowe do wszystkich narzędzi i konsol chmurowych

Licencje w planach biznesowych, z zapewnieniem, że dane Zamawiającego nie są wykorzystywane do trenowania modeli.

5. Obszar D – Wdrożenie i utrzymanie

1. **Wdrożenie (do 14 dni od zawarcia umowy):** konfiguracja środowisk w podejściu infrastruktury jako kod, potoki CI/CD, migracja istniejącego MVP i danych, konfiguracja dostępu do modeli AI i narzędzi, dokumentacja techniczna i operacyjna. Odbiór protokołem uruchomienia.
2. **Utrzymanie:** co najmniej **14 godzin** pracy inżyniera chmury miesięcznie (lub więcej zgodnie z ofertą) na zmiany konfiguracji, optymalizację kosztów i wydajności, aktualizacje bezpieczeństwa, wsparcie zespołu Zamawiającego; niewykorzystane godziny nie przechodzą na kolejne miesiące.
3. **Poziom usług:** dostępność środowiska produkcyjnego min. **99,9%** miesięcznie; czas reakcji na awarię krytyczną (niedostępność produkcji) **maks. 4 godziny** (lub krócej zgodnie z ofertą), 24/7.
4. **Zarządzanie subskrypcjami i licencjami:** rozliczenia z dostawcami chmury, modeli i licencji, pilnowanie limitów zużycia z alertem przy 80% wolumenu miesięcznego.
5. **Zakończenie umowy:** eksport danych, przekazanie kodu, konfiguracji, kont i licencji Zamawiającemu bez przerwy w działaniu środowiska produkcyjnego.