



NAJWYŻSZA IZBA KONTROLI

Biuro Bezpieczeństwa i Systemów Informatycznych

BBS-BBA.25.28.2026

Warszawa, 17 lipca 2026 r.

Zapytanie w celu ustalenia szacunkowej wartości zamówienia na dostawę platformy AI

Najwyższa Izba Kontroli w Warszawie (02-056), ul. Filtrowa 57, przekazuje zapytanie w celu ustalenia szacunkowej wartości zamówienia.

Proszę o wycenę przedmiotu zamówienia uwzględniając niżej przedstawione informacje.

I. Przedmiot zamówienia

Zakres zamówienia obejmuje:

1. Dostawa Systemu, będącego zintegrowanym rozwiązaniem sprzętowo-programowym wykorzystującym sztuczną inteligencję, spełniającego wymagania określone w pkt. 6 i 7 zgodnie z wymaganiami określonymi w pkt. 5 OPZ.
2. Wdrożenie dostarczonego Systemu zgodnie z wymaganiami określonymi w pkt. 8 OPZ.
3. Wytworzenie Dokumentacji powykonawczej zgodnie z wymaganiami opisanymi w pkt. 9 OPZ.
4. Przeprowadzenie przez Wykonawcę technicznego warsztatu powdrożeniowego zgodnie z wymaganiami określonymi w pkt. 10 OPZ.
5. Udzielenie gwarancji zgodnie z wymaganiami określonymi w pkt. 11 OPZ.
6. Świadczenie Asysty Technicznej Wykonawcy w ilości 3 000 Roboczogodzin, zgodnie z wymaganiami określonymi w pkt. 12 i 13 OPZ.

II. Termin realizacji zamówienia – harmonogram wstępny

- 4.1. Dostawa Systemu o którym mowa w pkt. I.1. nastąpi w terminie do 90 dni od dnia podpisania umowy.
- 4.2. Wdrożenie o którym mowa w pkt. I.2. nastąpi w terminie do 14 dni od dnia zakończenia dostaw potwierdzonych podpisaniem Protokołów odbioru Dostaw.

- 4.3. Wytworzenie Dokumentacji powykonawczej o której mowa w pkt. I.3. nastąpi w terminie do 14 dni od dnia zakończenia Wdrożenia, o którym mowa w pkt. II.2. potwierdzonego podpisaniem Protokołu odbioru Wdrożenia.
- 4.4. Przeprowadzenie przez Wykonawcę technicznego warsztatu powdrożeniowego o którym mowa w pkt. I.4, zostanie zrealizowane w terminie do 14 dni od dnia zakończenia Wdrożenia, o którym mowa w pkt II.2. potwierdzonego podpisaniem Protokołu odbioru Wdrożenia.
- 4.5. Wykonawca udzieli gwarancji o której mowa w pkt. I.5. na prawidłowe, w pełni zgodne z przeznaczeniem, funkcjonowanie dostarczonego Systemu od momentu dostawy o której mowa w pkt II.1. przez okres 36 miesięcy dla oprogramowania i 60 miesięcy dla sprzętu.
- 4.6. Wykonawca będzie świadczył Asystę Techniczną o której mowa w pkt. I.6. w ramach puli 3 000 Roboczogodzin od momentu zakończenia Wdrożenia, o którym mowa w pkt. II.2. do momentu upływu okresu gwarancji na oprogramowanie o którym mowa w pkt II.5.

III. Wycena

Przesłana wycena powinna zawierać wszystkie koszty, związane z realizacją zamówienia z uwzględnieniem warunków wskazanych w pkt. I. i Załączniku nr 2 (OPZ) do Zapytania, wyrażone w PLN. Wyceny należy dokonać na formularzu stanowiącym Załącznik nr 1 do Zapytania.

IV. Składanie wycen szacunkowych

Odpowiedzi na niniejsze Zapytanie należy udzielić, nie później niż do dnia **31 lipca 2026 r. godz. 14:00**, pocztą elektroniczną na adres email: **szacowanie-zp@nik.gov.pl**.

V. Inne

1. Informujemy, że niniejsze zapytanie nie stanowi zaproszenia do składania ofert w rozumieniu art. 66 Kodeksu cywilnego, nie zobowiązuje Zamawiającego do zawarcia umowy, czy też udzielenia zamówienia.
2. W przypadku wątpliwości i pytań dotyczących przedmiotu zamówienia należy kontaktować się z panem Szymonem Organelem, tel. 22 444 56 63, e-mail: szymon.organek@nik.gov.pl

VI. Klauzule RODO

1.1 Informacja o przetwarzaniu danych osoby, której dane pozyskano bezpośrednio od niej.

Zgodnie z art. 13 ust. 1 i 2 rozporządzenia Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych), (Dz. Urz. UE L 119 z 04.05.2016, str. 1, ze zm.,) dalej „RODO”, informujemy, że:

1. Administratorem Pani/Pana danych osobowych jest Najwyższa Izba Kontroli, 02-056 Warszawa, ul. Filtrowa 57, tel. (22) 444 50 00. Z administratorem można się Pani/ Pan skontaktować listownie na adres: ul. Filtrowa 57, 02-056 Warszawa, lub przez

elektroniczną skrzynkę podawczą na stronie

<https://www.nik.gov.pl/kontakt/elektroniczna-skrzynka-podawcza/>

2. Dane kontaktowe Inspektora Ochrony Danych: tel. 22 444 58 19, e-mail: iod@nik.gov.pl
3. Pani/ Pana dane osobowe będą przetwarzane na podstawie art. 6 ust. 1 lit. b) lub art. 6 ust. 1 lit. c) i f) RODO w związku z art. 44 ust. 3 pkt 1 ustawy z dnia 27.08.2009 r. o finansach publicznych (t.j. Dz.U. z 2024 r. poz. 1530, ze zm.), w celu przeprowadzenia wyceny postępowania o udzielenie zamówienia publicznego i zawarcia, realizacji i archiwizacji umowy o zamówienie oraz dochodzenie ewentualnych roszczeń z tytułu realizacji umowy.
4. Dane osobowe mogą być przekazywane podmiotom upoważnionym na podstawie przepisów prawa. Odbiorcą danych może być również podmiot świadczący usługi IT w zakresie serwisowania i usuwania awarii, na rzecz Administratora danych.
5. Dane osobowe będą przechowywane przez okres niezbędny do przeprowadzenia postępowania o udzielenie zamówienia, a w stosunku do danych osobowych wskazanych przez Wykonawcę, którego oferta została wybrana, przez okres trwania umowy o zamówienie oraz do czasu przedawnienia ewentualnych roszczeń wynikających z umowy. Ponadto dane osobowe będą przechowywane przez okres archiwizacji dokumentów wynikający z przepisów powszechnie obowiązujących oraz przepisów wewnętrznych Administratora Danych.
6. Podanie przez Panią/Pana danych osobowych jest dobrowolne, ale niezbędne do przeprowadzenia wyceny postępowania o udzielenie zamówienia.
7. W odniesieniu do danych osobowych przekazanych NIK, decyzje nie będą podejmowane w sposób zautomatyzowany, stosownie do art. 22 RODO.
8. Posiada Pani/Pan:
 - na podstawie art. 15 RODO prawo dostępu do danych osobowych Pani/Pana dotyczących;
 - na podstawie art. 16 RODO prawo do sprostowania Pani/Pana danych osobowych, które są nieprawidłowe lub ich uzupełnienia jeśli są niekompletne z uwzględnieniem celów ich przetwarzania
 - na podstawie art. 18 RODO prawo żądania od administratora ograniczenia przetwarzania danych osobowych z zastrzeżeniem przypadków, o których mowa w art. 18 ust. 2 RODO,
 - prawo do przenoszenia danych, w przypadkach określonych w art. 20 RODO;
 - prawo sprzeciwu wobec przetwarzania dotyczących Pani/Pana danych osobowych, na zasadach określonych w art. 21 RODO;
 - prawo do wniesienia skargi do Prezesa Urzędu Ochrony Danych Osobowych, gdy uzna Pani/Pan, że przetwarzanie danych osobowych Pani/Pana dotyczących narusza przepisy o ochronie danych osobowych.
9. Nie przysługuje Pani/Panu prawo do usunięcia danych osobowych, w związku z art. 17 ust. 3 lit. b), d) lub e) RODO.
10. Nie przysługuje Pani/ Panu prawo do przenoszenia danych osobowych, o których mowa w art. 20 RODO, z wyjątkiem sytuacji, gdy Pani/ Pana dane osobowe są przetwarzane na podstawie art. 6 ust. 1 lit. b) RODO.

1.2 Zakres informacji przekazywanych przez Wykonawcę osobom, których dane osobowe są wykazywane w związku ze składaną wyceną lub działającym i

realizującym wycenę w jego imieniu, o przetwarzaniu ich danych osobowych, w związku z art. 14 RODO

1. Z chwilą udostępnienia danych osobowych Najwyższej Izbie Kontroli, administratorem tych danych osobowych jest Najwyższa Izba Kontroli, z siedzibą w Warszawie przy ul. Filtrowej 57. Z administratorem może się Pani/Pan kontaktować listownie na adres ul. Filtrowa 57, 02-056 Warszawa, lub przez elektroniczną skrzynkę podawczą <https://www.nik.gov.pl/kontakt/elektroniczna-skrzynka-podawcza/>
2. Dane kontaktowe Inspektora Ochrony Danych: tel. 22 444 58 19, e-mail: iod@nik.gov.pl.
3. Dane osobowe są przetwarzane na podstawie art. 6 ust. 1 lit. c) i f) RODO w związku z art. 44 ust. 3 pkt 1 ustawy z dnia 27.08.2009 r. o finansach publicznych (t.j. Dz.U. z 2024 r. poz. 1530, ze zm.), w celu przeprowadzenia niniejszej wyceny dla postępowania o udzielenie zamówienia w trybie zapytania ofertowego, zawarcia, realizacji i archiwizacji umowy o zamówienie oraz dochodzenia ewentualnych roszczeń z tytułu realizacji umowy.
4. Kategorie danych osobowych zebrane przez Wykonawcę i udostępnione przez niego Administratorowi danych jest następująca: dane osobowe osób, których dotyczą, ujawnione NIK w celu wyceny w niniejszym postępowaniu o udzielenie zamówienia publicznego, zawarcia i realizacji umowy o zamówienie.
5. Dane osobowe mogą być przekazywane podmiotom upoważnionym na podstawie przepisów prawa. Odbiorcą danych może być również podmiot świadczący usługi IT w zakresie serwisowania i usuwania awarii, na rzecz Administratora danych.
6. Dane osobowe będą przechowywane przez okres niezbędny do przeprowadzenia postępowania o udzielenie zamówienia, a w stosunku do danych osobowych wskazanych przez Wykonawcę, którego oferta została wybrana - przez okres trwania umowy o zamówienie oraz do czasu przedawnienia ewentualnych roszczeń wynikających z umowy. Ponadto dane osobowe przechowywane będą przez okres archiwizacji dokumentów wynikający z przepisów powszechnie obowiązujących oraz przepisów wewnętrznych Administratora Danych.
7. W odniesieniu do danych osobowych przekazanych NIK, decyzje nie będą podejmowane w sposób zautomatyzowany, stosownie do art. 22 RODO.
8. Osoby, których dane zostaną przekazane NIK, mają prawo dostępu do danych osobowych ich dotyczących oraz do sprostowania lub uzupełnienia tych danych, w przypadkach określonych w art. 16 RODO.
9. Osoby, których dane zostaną przekazane NIK, mają prawo żądania od administratora ograniczenia przetwarzania danych osobowych, z wyjątkami określonymi w art. 18 ust. 2 RODO oraz mają prawo sprzeciwu wobec przetwarzania dotyczących ich danych osobowych, na zasadach określonych w art. 21 RODO.
10. Osoby, których dane zostaną przekazane NIK, mają prawo do wniesienia skargi do organu nadzorczego, którym w Polsce jest Prezes Urzędu Ochrony Danych Osobowych, gdy przetwarzanie ich danych osobowych narusza przepisy o ochronie danych osobowych.
11. Osobom, których dane zostaną przekazane NIK, nie przysługuje prawo do usunięcia danych osobowych w związku z art. 17 ust. 3 lit. b), d) lub e) RODO oraz prawo do przenoszenia danych osobowych, o którym mowa w art. 20 RODO.

VII. Procedura dokonywania zgłoszeń wewnętrznych naruszeń prawa i podejmowania działań następczych w Najwyższej Izbie Kontroli

Zamawiający informuje, że procedura zgłaszania informacji o naruszeniach prawa i podejmowania działań następczych w Najwyższej Izbie Kontroli dostępna jest na stronie internetowej Zamawiającego, pod adresem:

[Informacje dla sygnalistów wewnętrznych - Najwyższa Izba Kontroli \(nik.gov.pl\)](http://nik.gov.pl)

Załączniki:

Załącznik nr 1 – Formularz Wyceny

Załącznik nr 2 – Opis przedmiotu zamówienia (OPZ)

Stanisław Bartos

Kierownik wydziału
w Biurze Bezpieczeństwa
i Systemów Informatycznych
Najwyższej Izby Kontroli
/podpisano elektronicznie/

Szacowanie zamówienia na Dostawę platformy AI

Wykonawca	
-----------	--

adres ul. _____ kod pocztowy: _____

miasto: kraj:

nr telefonu: _____ adres e-mail: _____

Lp.	Produkt/ Usługa	Rodzaj płatności	Okres wsparcia/realizacji	Ilość	Cena jednostkowa netto w PLN - (<i>należy podać z dokładnością do dwóch miejsc po przecinku</i>)	Cena jednostkowa brutto w PLN - (<i>należy podać z dokładnością do dwóch miejsc po przecinku</i>)	Cena netto w PLN produktu (iloczyn wartości z kolumny 4 i 5 [4x5]) -(<i>należy podać z dokładnością do dwóch miejsc po przecinku</i>)	Cena brutto w PLN produktu (iloczyn wartości z kolumny 4 i 6 [4x6]) -(<i>należy podać z dokładnością do dwóch miejsc po przecinku</i>)
	1	2	3	4	5	6	7	8
1	Dostawa Systemu - Serwer (OPZ pkt 6) wraz z instalacją i konfiguracją Producent:..... Model:..... Uwaga: do wyceny należy załączyć szczegółową specyfikację techniczną proponowanego serwera	jednorazowo po dostawie	Dostawa do 90 dni od daty podpisania umowy gwarancja 60 miesięcy od daty dostawy	1	- zł	- zł	- zł	- zł
2	Dostawa systemu - oprogramowanie (OPZ pkt 7.1-7.11) wraz z wdrożeniem, dokumentacją i warsztatem powdrożeniowym Uwaga: do wyceny należy załączyć wykaz głównych koponentów oprogramowania wchodzących w skład wycenianego/oferowanego Systemu	jednorazowo po zakończeniu wdrożenia, wytworzeniu dokumentacji i przeprowadzeniu warsztatu	Dostawa do 90 dni od daty podpisania umowy gwarancja 36 miesięcy od daty dostawy	1	- zł	- zł	- zł	- zł
3	Asysta Techniczna Wykonawcy (OPZ pkt 13)	kwartalnie z dołu za faktycznie wykorzystaną ilość godzin w okresie rozliczeniowym	od zakończenia wdrożenia do upływu 36 miesięcznej gwarancji na oprogramowanie	3000	- zł	- zł	- zł	- zł
4	OPCJA - rozbudowa serwera (OPZ pkt 15) - sprzęt	po wykonaniu rozbudowy	termin zakończenia gwarancji zgodny z gwarancją dostarczonego serwera	1	- zł	- zł	- zł	- zł
5	OPCJA - rozbudowa serwera (OPZ pkt 15) - oprogramowanie	po wykonaniu rozbudowy	termin zakończenia gwarancji zgodny z gwarancją dostarczonego oprogramowania	1	- zł	- zł	- zł	- zł
Razem							- zł	- zł

Szczegółowy opis przedmiotu zamówienia

Dostawa platformy AI

Spis treści

1.	Cel zamówienia publicznego	3
2.	Plan komunikacji Wykonawcy z Zamawiającym	3
3.	Ogólne wymagania Zamawiającego dotyczące zakresu przedmiotu zamówienia	4
4.	Harmonogram realizacji przedmiotu zamówienia	4
5.	Wymagania Zamawiającego dotyczące dostaw	5
6.	Parametry serwera	5
7.	Wymagania w zakresie oprogramowania	7
7.1.	Oprogramowanie do wirtualizacji	7
7.2.	Oprogramowanie do zarządzania infrastrukturą kontenerową i cyklem życia modeli AI	7
7.3.	Stos do uruchamiania modeli językowych (LLM Stack)	12
7.4.	Portal wewnętrzny z aplikacjami AI (portal AI)	13
7.5.	Aplikacja do agregacji wiedzy i konwersacji z asystentem AI	13
7.6.	Dedykowana aplikacja do pracy z dokumentami z wykorzystaniem OCR, czatu AI i poprawnych operacji na liczbach, umożliwiająca pracę grupową	14
7.7.	Aplikacja do automatycznej transkrypcji i analizy spotkań	15
7.8.	Aplikacja webowa do tłumaczenia tekstu i dokumentów	16
7.9.	System zarządzania bazami wiedzy dla asystentów AI	16
7.10.	Wymagane funkcjonalności oprogramowania zarządzania bazami wiedzy:	17
7.11.	System generowania raportów zbiorczych	21
7.12.	Licencje na oprogramowanie	24
7.13.	Wymagania w zakresie bezpieczeństwa Systemu	25
8.	Wymagania w zakresie Wdrożenia	26
9.	Wymagania dotyczące wytwarzanych i przetwarzanych dokumentów w tym Dokumentacji powykonawczej systemu 27	
10.	Wymagania w zakresie technicznego warsztatu powdrożeniowego	28
11.	Wymagania dotyczące świadczenia usług wynikających z udzielonej gwarancji	29
12.	Wymagania w zakresie świadczenia usług Wsparcia technicznego producenta	31
13.	Wymagania w zakresie świadczenia Asysty Technicznej Wykonawcy	31
14.	Zasady zdalnego dostępu do środowiska informatycznego Zamawiającego	32

1. Cel zamówienia publicznego

Celem zamówienia jest dostawa oraz uruchomienie zaawansowanego, lokalnego systemu sztucznej inteligencji (on-premise AI), działającego w całości w infrastrukturze Zamawiającego, bez konieczności przesyłania danych do zewnętrznych chmur obliczeniowych.

Głównym celem projektu jest podniesienie efektywności realizowanych procesów wewnętrznych w zakresie analizy dokumentów i danych, przy jednoczesnym zapewnieniu maksymalnego poziomu bezpieczeństwa, poufności oraz pełnej kontroli nad przetwarzanymi informacjami.

Wdrożenie lokalnego rozwiązania AI ma na celu realizację następujących założeń:

- **Bezpieczeństwo i suwerenność danych:** Zagwarantowanie, że wszelkie dane wejściowe, zapytania oraz wyniki analiz nie opuszczają bezpiecznej sieci wewnętrznej Zamawiającego.
- **Optymalizacja procesów wewnętrznych:** Skrócenie czasu potrzebnego na realizację zadań takich jak np. analiza dokumentów i danych, wyszukiwanie informacji w wewnętrznych bazach wiedzy, transkrypcja spotkań, tłumaczenie dokumentów.
- **Niezależność technologiczna:** Uniezależnienie Zamawiającego od zewnętrznych dostawców usług chmurowych, zmian w ich cennikach, politykach prywatności oraz potencjalnych przerw w dostępie do tych usług.
- **Skalowalność i rozwój:** Stworzenie fundamentu technologicznego na bazie którego rozwijane będą kolejne obszary wykorzystania rozwiązań z zakresu uczenia maszynowego i sztucznej inteligencji (AI), oraz integracje z istniejącymi systemami dziedzicznymi Zamawiającego.

2. Plan komunikacji Wykonawcy z Zamawiającym

2.1. Możliwe kanały komunikacji to:

2.1.1.wideokonferencja (obsługę zapewnia Zamawiający),

2.1.2.telefon,

2.1.3.e-mail,

2.1.4.spotkanie w siedzibie NIK.

2.2. Wszystkie prace będą realizowane przy udziale lub w konsultacji z pracownikami Zamawiającego.

2.3. Wykonawca będzie konsultował z Zamawiającym wszystkie przyjmowane założenia poczynione w związku z realizacją umowy. W związku z tym, w razie potrzeby dostarczy wszelkich niezbędnych wyjaśnień i materiałów dodatkowych (opisów, Dokumentacji itp.) pracownikom Zamawiającego tak, aby możliwe było jednoznaczne zrozumienie proponowanych przez niego założeń.

- 2.4. Wszystkie ustalenia poczynione za pośrednictwem wideokonferencji, telefonicznie lub w trakcie spotkania muszą zostać niezwłocznie potwierdzone za pośrednictwem wiadomości e-mail. Mogą zostać z nich także sporządzone notatki zgodnie ze wzorem dostarczonym przez Zamawiającego.
- 2.5. Wykonawca zobowiązany jest poinformować Zamawiającego o wszystkich zdarzeniach lub przeszkodach mogących spowodować opóźnienie w wykonaniu Umowy w stosunku do terminów przewidzianych w umowie.

3. Ogólne wymagania Zamawiającego dotyczące zakresu przedmiotu zamówienia

- 3.1. Dostawa Systemu, będącego zintegrowanym rozwiązaniem sprzętowo-programowym wykorzystującym sztuczną inteligencję, spełniającego wymagania określone w pkt 6 i 7 zgodnie z wymaganiami określonymi w pkt 5.
- 3.2. Wdrożenie dostarczonego Systemu zgodnie z wymaganiami określonymi w pkt. 8.
- 3.3. Wytworzenie Dokumentacji powykonawczej zgodnie z wymaganiami opisanymi w pkt. 9.
- 3.4. Przeprowadzenie przez Wykonawcę technicznego warsztatu powdrożeniowego zgodnie z wymaganiami określonymi w pkt. 10.
- 3.5. Udzielenie gwarancji zgodnie z wymaganiami określonymi w pkt. 11.
- 3.6. Świadczenie Asysty Technicznej Wykonawcy w ilości 3 000 Roboczogodzin, zgodnie z wymaganiami określonymi w pkt. 12 i 13.

4. Harmonogram realizacji przedmiotu zamówienia

- 4.1. Dostawa Systemu o którym mowa w pkt 3.1 nastąpi w terminie **do 90 dni** od dnia podpisania umowy¹.
- 4.2. Wdrożenie o którym mowa w pkt 3.2 nastąpi w terminie **do 14 dni** od dnia zakończenia dostaw potwierdzonych podpisaniem Protokołów odbioru Dostaw.
- 4.3. Wytworzenie Dokumentacji powykonawczej o której mowa w pkt. 3.3 nastąpi w terminie **do 14 dni** od dnia zakończenia Wdrożenia, o którym mowa w pkt 4.2. potwierdzonego podpisaniem Protokołu odbioru Wdrożenia.
- 4.4. Przeprowadzenie przez Wykonawcę technicznego warsztatu powdrożeniowego o którym mowa w pkt. 3.4, zostanie zrealizowane w terminie **do 14 dni** od dnia zakończenia Wdrożenia, o którym mowa w pkt 4.2 potwierdzonego podpisaniem Protokołu odbioru Wdrożenia.

¹ Termin dostawy może ulec skróceniu zgodnie z ofertą Wykonawcy (kryterium punktowane)

- 4.5. Wykonawca udzieli gwarancji o której mowa w pkt 3.5 na prawidłowe, w pełni zgodne z przeznaczeniem, funkcjonowanie dostarczonego Systemu od momentu dostawy o której mowa w pkt 4.1 przez okres 36 miesięcy dla oprogramowania i 60 miesięcy dla sprzętu.
- 4.6. Wykonawca będzie świadczył Asystę Techniczną o której mowa w pkt 3.6 w ramach puli **3 000 Roboczogodzin** od momentu zakończenia Wdrożenia, o którym mowa w pkt 4.2 do momentu upływu okresu gwarancji na oprogramowanie o którym mowa w pkt 4.5.

5. Wymagania Zamawiającego dotyczące dostaw

- 5.1. Dostawa Systemu musi zostać zrealizowana w terminie wskazanym w pkt 4.1.
- 5.2. Miejscem dostawy jest siedziba Najwyższej Izby Kontroli przy ul. Filtrowej 57 w Warszawie.
- 5.3. Dostawa będzie awizowana przez Wykonawcę na piśmie lub pocztą elektroniczną kierowaną na adres: bbs@nik.gov.pl.
- 5.4. Sprzęt wchodzący w zakres dostawy zostanie dostarczony Zamawiającemu w opakowaniach zabezpieczających przed uszkodzeniem w czasie transportu.
- 5.5. Dostarczone urządzenia muszą być fabrycznie nowe i nie mogą znajdować się na liście „end-of-sale” lub/oraz „end-of-support” producenta oraz pozyskane z oficjalnego kanału dystrybucyjnego producenta.
- 5.6. Wykonawca zobowiązuje się dostarczyć wymagane urządzenia, oprogramowanie oraz licencje pochodzące z legalnego źródła, zakupione w autoryzowanym kanale sprzedaży producenta na terenie UE i objęte standardowym pakietem usług gwarancyjnych świadczonych przez sieć serwisową producenta na terenie UE.
- 5.7. Potwierdzeniem realizacji Dostawy będzie Protokół odbioru podpisany bez uwag, przez Przedstawicieli Zamawiającego i Wykonawcy wg wzoru określonego w Załączniku nr 3 do Umowy.

6. Parametry serwera

W ramach dostawy Wykonawca dostarczy serwer spełniający określone poniżej wymagania minimalne.

L.p.	Nazwa wymagania / parametru	Wartości minimalne wymagane przez Zamawiającego
1	2	3
1	Obudowa	Maksymalnie 4U z możliwością instalacji w szafie przemysłowej 19", z zestawem szyn do mocowania w szafie i wysuwania do celów serwisowych.

2	Procesor	Procesor klasy x86 przystosowany do pracy w serwerach dwuprocesorowych. Procesor 32-rdzeniowy najnowszej generacji oferowanej przez producenta, osiągający w dwuprocesorowej konfiguracji wynik SPEC CPU 2017 Integer Rate Base results powyżej 700 pkt. Wynik musi być dostępny na stronie https://spec.org
3	Liczba procesorów zainstalowanych	Min. 2 identyczne procesory.
4	Płyta główna	Umożliwiająca instalację minimum 2 procesorów, wykonana i zaprojektowana przez producenta serwera oraz oznaczona jego znakiem firmowym. Wszystkie gniazda procesorowe w serwerze muszą być obsadzone procesorami.
5	Pamięć RAM	Minimum 512 GB ECC DDR5 6400 MHz. Pamięć obsadzona w konfiguracji zbalansowanej, tj. obsadzone wszystkie kanały kontrolerów pamięci procesorów. Oryginalna pamięć producenta serwera, przy czym każdy moduł pamięci musi posiadać tzw. part number producenta serwera. Minimum 24 sloty na pamięć. Możliwość rozbudowy do 3TB RAM.
7	Dyski twarde	4 dyski NVME o pojemności co najmniej 3,8TB każdy. Urządzenia muszą być skonfigurowane w RAID 1 za pomocą sprzętowego kontrolera RAID z możliwością wymiany w przypadku awarii bez wyłączania systemu.
8	Dyski twarde	Możliwość rozbudowy o dodatkowe min 4 dyski twarde typu hot plug 2,5
9	Gniazda rozszerzeń	8 złącz PCIe Gen5 x16 (double-wide, full-height, full-length 600W GPUs) 2 złącza PCIe x16, 1 slot OCP
10	GPU	4 szt NVIDIA RTX PRO 6000 Blackwell Server Edition 96GB PCIe Gen5 Możliwość rozbudowy do 8 szt.
11	Interfejsy sieciowe - wbudowane	Minimum 2 porty 10/100/1000 wbudowane na płycie głównej nie zajmujących gniazd PCI.
12	Interfejsy sieciowe - dodatkowe	1 karta sieciowa 2 x 10/25 GbE SFP28 wraz z wkładkami MultiMode 1 karta sieciowa 2 x 100GbE QSFP28 wraz z wkładkami MultiMode
13	Procedura przewidywania awarii	monitorowanie min.: temperatury, dysków, zasilaczy, płyty głównej, procesorów, pamięci operacyjnej
14	Zarządzanie	Zintegrowany z płytą główną kontroler zdalnego zarządzania zgodny ze standardem IPMI 2.0, SNMP 3.0 umożliwiający zdalny restart serwera i pełne zarządzania serwerem poprzez szyfrowane połączenie w sieci TCP/IP przy użyciu przeglądarki internetowej, pozwalający m.in. na dostęp do konsoli serwera (pracującego zarówno w trybie tekstowym jak i graficznym), włączanie/wyłączanie serwera, reinstalację systemu operacyjnego, zdalnego podłączenia napędów, jeśli do wyżej wymienionej funkcjonalności jest wymagana licencja - należy ją dostarczyć.
15	Grafika	Zintegrowana z płytą główną, pamięć minimum 16MB, wsparcie dla rozdzielczości min 1280x1024, złącze VGA
16	Zasilacze	4 redundantne zasilacze, każdy minimum 2400W Hot-Plug, wystarczające na prawidłowe zasilanie oferowanego serwera w pełnej konfiguracji tj. 8 szt GPU, o wysokiej sprawności (minimum 94 %, klasa Platinum) Kable zasilające, umożliwiające podłączenie do listwy zasilającej w szafie Rack zamawiającego o długości co najmniej 2m
17	Wentylatory	zestaw wentylatorów redundantnych typu hot swap o wysokiej wydajności, pozwalająca na prawidłowe chłodzenie serwera w pełnej konfiguracji tj. 8 szt GPU, przy temperaturze otoczenia do 40°C, przy założeniu pracy w tych warunkach przez nie więcej niż 10% czasu w ciągu roku kalendarzowego zainstalowane dedykowane radiatory odbierające ciepło bezpośrednio z procesorów zainstalowane osłony na niewykorzystane sloty dla pamięci RAM
18	Porty dodatkowe	• zintegrowana karta graficzna ze złączem VGA; • łącznie min 4 porty USB: min 1xUSB 3.0 dostępne na froncie obudowy, min 3xUSB 3.0 dostępne z tyłu serwera (Ilość dostępnych złącz VGA i USB nie może być osiągnięta poprzez stosowanie zewnętrznych przejściówek, rozgałęziaczy czy dodatkowych kart rozszerzeń zajmujących jakikolwiek slot PCI Express serwera);
19	Systemy operacyjne oraz oprogramowanie wirtualizacyjne	Wsparcie dla Red Hat Enterprise Linux, SUSE Linux Enterprise Server, Microsoft Windows Server, VMware ESXi
20	Bezpieczeństwo	Zainstalowany moduł TPM 2.0
21	Gwarancja – warunki	Producenta z częściami i robocizną, w miejscu instalacji przez cały czas obowiązywania gwarancji, świadczona w dni robocze od poniedziałku do piątku z wyłączeniem świąt, w godzinach od 8 do 17 z czasem skutecznej naprawy do następnego dnia roboczego od momentu zgłoszenia.

		W przypadku awarii dyski twarde pozostają u Zamawiającego. Zużycie dysku typu flash również jest postrzegane jako awaria przez Zamawiającego.
22	Gwarancja – czas trwania	Minimum 60 miesięcy od daty dostawy
23	Pochodzenie	Dostarczony sprzęt musi być fabrycznie nowy. Elementy, z których zbudowane są serwery muszą być produktami producenta tych serwerów lub być przez niego certyfikowane oraz muszą być objęte gwarancją producenta Sprzęt nie może się znajdować się na liście (EOS = EndOfSupport / EOL= EndOfLife) producenta
24	Sterowniki	Sterowniki dla danego urządzenia muszą być dostępne na stronie producenta. Prawo do pobierania aktualnego oprogramowania typu firmware/sterowniki przez okres obowiązywania gwarancji dla serwera.
25	Certyfikaty	Wymagane od producenta: ISO 9001 lub dokument równoważny, ISO14001 lub dokument równoważny
26	Infolinia	Możliwość telefonicznego sprawdzenia konfiguracji sprzętowej serwera oraz warunków gwarancji po podaniu numeru seryjnego bezpośrednio u producenta lub jego przedstawiciela. Możliwość zgłaszania awarii 24x7x365 poprzez ogólnopolską linię telefoniczną producenta.

7. Wymagania w zakresie oprogramowania

7.1. Oprogramowanie do wirtualizacji

7.1.1. Zamawiający wymaga dostarczenia oprogramowania do wirtualizacji serwerowej, które stanowić będzie fundament infrastruktury platformy AI. Oprogramowanie to musi umożliwiać efektywne zarządzanie zasobami obliczeniowymi, w szczególności dedykowanymi zasobami GPU niezbędnymi do uruchamiania modeli sztucznej inteligencji.

7.1.2. Wymagane funkcjonalności obejmują wsparcie dla wirtualizacji z pełną izolacją poszczególnych środowisk, możliwość alokowania zasobów procesorowych i pamięci zgodnie z bieżącymi potrzebami obciążenia oraz mechanizmy wysokiej dostępności zapewniające ciągłość działania krytycznych usług. System wirtualizacji musi wspierać migrację „live” działających maszyn wirtualnych między fizycznymi hostami.

7.1.3. Oprogramowanie powinno oferować mechanizmy zabezpieczeń obejmujące izolację sieciową pomiędzy środowiskami, kontrolę dostępu opartą na rolach oraz logowanie operacji administracyjnych. Wymagana jest kompatybilność z pozostałymi komponentami platformy AI, w szczególności z oprogramowaniem do zarządzania warstwą konteneryzacji.

7.2. Oprogramowanie do zarządzania infrastrukturą kontenerową i cyklem życia modeli AI

Dostarczone oprogramowanie musi spełniać niżej wymienione wymagania funkcjonalne:

7.2.1. Zarządzanie rozproszoną infrastrukturą obliczeniową

- Rejestracja i automatyczne wykrywanie węzłów obliczeniowych połączonych z orkiestratorem kontenerów, z rozpoznawaniem specyfikacji sprzętowej (CPU, GPU, pamięć VRAM).
- Monitorowanie w czasie rzeczywistym metryk sprzętowych oraz stanu zdrowia maszyn w jednym, spójnym widoku infrastruktury.
- Obsługa klastrów Docker Swarm jako pełnoprawnych obiektów, wraz z procesem discovery i adopcją węzłów.

- Funkcja integracji z dostawcą maszyn wirtualnych - onboarding maszyn, synchronizacja parametrów sprzętowych oraz stanu zasilania.
- Bezpieczne, dwustopniowe wycofywanie węzła (wycofanie, a następnie trwałe usunięcie) z czytelną informacją o powiązanych obiektach.

7.2.2. Orkiestracja wdrożeń kontenerowych

- Wdrożenia oparte o szablony z parametrami, umożliwiające wielokrotne wykorzystanie konfiguracji.
- Obsługa dwóch trybów pracy: Docker Compose (pojedynczy węzeł) oraz Docker Swarm (wiele węzłów, usługi z wieloma zadaniami).
- Pełny cykl życia wdrożenia: wersja robocza → wdrażanie → wdrożone / błąd → zarchiwizowane, niezależny od bieżącej kondycji aplikacji.
- Interaktywne skalowanie usług w klastrze (zmiana liczby replik w czasie działania).
- Sterowanie stanem aplikacji (uruchom / zatrzymaj) oraz operacje na poszczególnych jednostkach (start, stop, restart, usunięcie).
- Możliwość prowadzenia stanu wdrożenia opartego o repozytorium kontroli wersji - jako źródło prawdy, ścieżka wycofywania zmian i pełna przejrzysta historia.

7.2.3. Potok wdrożeniowy z walidacją

- Każde wdrożenie przechodzi przez wieloetapowy potok walidacji przed faktycznym uruchomieniem, z możliwością obserwacji postępu każdego etapu na żywo.
- Sprawdzenie konfiguracji - weryfikacja dostępności orkiestratora oraz (opcjonalnie) repozytorium Git, obecności i poprawności pliku wdrożenia.
- Sprawdzenie węzła - kontrola dostępności węzła docelowego, obecności wymaganych obrazów i plików modelu oraz zasobów (w tym alokacji GPU i dostępnej pamięci VRAM).
- Pobranie obrazów z podglądem postępu warstwa po warstwie (liczba bajtów, procent ukończenia, status); możliwość przerwania operacji przez operatora.
- Automatyczne zatrzymanie potoku na pierwszym nieudanym etapie z zapisem szczegółów błędu w historii; przy ponownym wdrożeniu (redeploy), jeśli błąd wystąpi przed etapem wykonania, dotychczasowa wersja pozostaje uruchomiona.

7.2.4. Zaawansowany system szablonów i parametryzacji

- Wsparcie dla wielu typów zmiennych: tekst, liczby całkowite i zmiennoprzecinkowe, wartości logiczne, listy wyboru, wyrażenia (obliczane dynamicznie), dane w formacie JSON, wybór urządzeń GPU oraz sekrety.
- Organizacja zmiennych w zakres ogólny (dla całego wdrożenia) oraz zakresy przypisane do poszczególnych usług.
- Walidacja wartości zgodnie z regułami danego typu: wzorce dla tekstu, zakresy min/maks dla liczb, dopuszczalne opcje dla list wyboru.
- Warunkowa widoczność pól - pola pojawiają się lub znikają w zależności od wartości innych pól.
- Symbole zastępcze i kontekst - w treści szablonu można odwoływać się do parametrów oraz do kontekstu środowiska (dane węzła, modelu, szablonu i wdrożenia).
- Bezpieczne odwołania do sekretów (rozwijane do nazwy lub wartości).

7.2.5. Wersjonowanie i porównywanie wersji szablonów

- Niezmiennie wersje szablonów - każda zmiana tworzy nową, zachowaną wersję.

- Strukturalny komparator wersji prezentujący dokładnie, co się zmieniło: dodane, usunięte i zmodyfikowane zakresy oraz zmienne (wraz ze zmianami poszczególnych właściwości - typu, wymagalności, wartości domyślnej, reguł walidacji czy warunków widoczności).
- Różnice w treści szablonu w ujęciu wiersz po wierszu oraz ostrzeżenia o skutkach zmian (np. usunięcie elementu wpływającego na działające wdrożenia).
- Klonowanie szablonów i wdrożeń w celu szybkiego tworzenia wariantów.

7.2.6. Eksport, import i patche konfiguracji

- Eksport pełnego szablonu w przenośnym formacie (z sumą kontrolną i wersją formatu).
- Przebieg próbny (dry-run) importu - diagnostyka przed zatwierdzeniem, w tym wskazanie, które wartości wymagają dostosowania w środowisku docelowym.
- Poprawki z kontrolą ciągłości wersji (lineage) - nakładanie zmian jako kolejnych iteracji istniejącego szablonu, z automatycznym ostrzeganiem o konfliktach w przypadku wykrycia modyfikacji rozbieżnych.
- Redakcja danych wrażliwych przy eksporcie (ukrywanie odwołań do sekretów).

7.2.7. Publikowane punkty końcowe i testowanie

- Automatyczne wykrywanie i prezentacja punktów końcowych (endpoints) udostępnianych przez wdrożenie, wraz z opisem, etykietami oraz poziomem widoczności (wewnętrzny / publiczny).
- Test HTTP - serwerowe sprawdzenie dostępności pod zadaną ścieżką (kod odpowiedzi, czas odpowiedzi, podgląd treści), z zabezpieczeniem przed nadużyciami (SSRF).
- Test łączności TCP - niezależny od protokołu test osiągalności portu (warstwa L4).
- Oznaczanie „osieroconych” endpoint, gdy dany port przestaje być publikowany przez wdrożenie.

7.2.8. Konfiguracja adapterów i integracji

- Testowanie połączeń dla każdego typu adaptera (orkiestrator, monitoring, dostawca modeli, brama proxy modeli, dostawca maszyn wirtualnych) z zwrotem czytelnego wyniku, czasu odpowiedzi oraz dodatkowych informacji diagnostycznych.
- Możliwość testowania konfiguracji zarówno dla zapisanego adaptera, jak i przed jej zapisaniem - bez utrwalania jakichkolwiek danych.
- Wychodzące proxy HTTP dla adaptera - ruch wychodzący każdego adaptera można skierować przez wskazane proxy HTTP (z opcjonalnym uwierzytelnieniem), niezależnie dla każdego adaptera.
- Obsługa wielu równoległych adapterów bramy proxy modeli, każdy z własnymi poświadczeniami i konfiguracją.
- Poświadczenia adapterów przechowywane w postaci zaszyfrowanej, rozszyfrowywane wyłącznie w momencie wywołania integracji.

7.2.9. Zarządzanie sekretami i bezpieczeństwo wbudowane

- Szyfrowanie sekretów w spoczynku; wartość jawna pojawia się wyłącznie w momencie wywołania docelowej integracji.
- Dwuetapowe renderowanie konfiguracji - sekrety i dane wrażliwe są wstrzykiwane do środowiska dopiero na styku z systemem orkiestracji (w trybie just-in-time), co eliminuje ryzyko ich wycieku na wcześniejszych etapach przetwarzania.

- Pełny audyt dostępu do sekretów (podgląd, rotacja, użycie) bez zapisywania wartości jawnych.
 - Integracja z natywnymi sekretami Docker Swarm wraz z wykrywaniem rozbieżności stanu.
 - Uprawnienia chroniące krytyczne operacje (podgląd, tworzenie, aktualizacja, usunięcie, ujawnienie, rotacja) oraz dedykowane role.
- 7.2.10. Monitoring, metryki i telemetria w czasie rzeczywistym
- Cykliczna kontrola dostępności i kondycji węzłów.
 - Metryki na żywo: wykorzystanie GPU i pamięci VRAM, CPU, pamięci operacyjnej.
 - Agregacja kondycji aplikacji - czytelny werdykt (zdrowa / degradacja / niesprawna / uruchamianie / zatrzymana) na podstawie stanu poszczególnych jednostek.
 - Historia metryk wspierająca analizę trendów i planowanie pojemności.
- 7.2.11. Logi i strumień na żywo
- Bezpośredni podgląd logów kontenerów oraz scalonych logów usług w klastrze, na żywo.
 - Strumień aktywności prezentujący zdarzenia w całej platformie (zmiany stanu wdrożeń, pojawianie się i znikanie węzłów).
 - Wydajna architektura strumieniowania (SSE) obsługująca wiele równoczesnych podglądów bez obciążania głównego ruchu API.
- 7.2.12. Kontrola dostępu i wielodostępność (RBAC)
- Rozbudowany model uprawnień (w formacie zasób.akcja) oraz zestaw gotowych ról systemowych.
 - Nadawanie dostępu do konkretnych zasobów (węzłów, modeli, wdrożeń, szablonów) wybranym zespołom.
 - Efektywne uprawnienia wynikające z połączenia ról i przyznanych dostępuów.
 - Egzekwowanie uprawnień „domyślnie zamknięte” oraz automatyczne filtrowanie widoków do zasobów, do których użytkownik ma dostęp.
- 7.2.13. Uwierzytelnianie
- Uwierzytelnianie sesyjne dla interfejsu.
 - Opcjonalna integracja z logowaniem jednokrotnym (SSO) przez Keycloak, z automatycznym zakładaniem kont i mapowaniem grup na role.
 - Ochrona przed atakami na hasła (blokada po przekroczeniu liczby nieudanych prób).
- 7.2.14. Audyt i zgodność
- Wielostrumieniowa oś czasu: migawki konfiguracji, zmiany stanu, operacje na jednostkach oraz zdarzenia infrastrukturalne.
 - Możliwość prześledzenia historii dla konkretnego wdrożenia, węzła lub sekretu.
 - Automatyczne usuwanie poświadczeń z treści logów przed ich zapisem.
- 7.2.15. Zadania systemowe
- Lista zadań cyklicznych wraz ze statusem (włączone / wyłączzone), interwałem uruchamiania, czasem ostatniego i kolejnego uruchomienia, łączną liczbą wykonań oraz średnim czasem trwania.
 - Zarządzanie interwałem z poziomu interfejsu - zmiana częstotliwości uruchamiania zadania oraz jego szybkie włączanie i wyłączanie.

- Historia wykonań - rejestr uruchomień z wynikiem (sukces / błąd / ponowienie), czasem rozpoczęcia i zakończenia, czasem trwania, zwróconym wynikiem oraz śladem błędu w razie niepowodzenia.
 - Podgląd procesów roboczych - stan procesów wykonawczych (dostępność, liczba aktywnych oraz liczba przetworzonych zadań).
- 7.2.16. Brama dostępu do modeli (Proxy)
- Generowanie i zarządzanie kluczami API.
 - Zespoły i polityki dostępu określające, które modele są dostępne dla kogo.
 - Aliasy (przyjazne nazwy wskazujące na konkretne modele lub punkty końcowe).
 - Konfiguracja tras kierujących ruch do właściwych usług.
- 7.2.17. Testowanie i czat z modelami
- Czat bezpośredni z wdrożonym modelem - wysyłanie zapytań wprost do punktu końcowego wdrożenia (w formacie zgodnym z OpenAI), z możliwością wyboru kontenera oraz ustawienia limitu tokenów i temperatury.
 - Czat przez bramę proxy z użyciem klucza API/tokenu podawanego przez użytkownika (klucz nie jest zapisywany ani logowany), z wyborem modelu i typu zapytania oraz weryfikacją poprawności klucza.
 - Wspólny interfejs czatu z historią wiadomości, regulacją limitu tokenów i temperatury oraz podglądem zużycia tokenów.
- 7.2.18. Model Agent - komponent wykonawczy na węzłach
- Bezstanowy, rozproszony agent wykonawczy uruchamiany na węzłach obliczeniowych (w tym z procesorami GPU), odpowiedzialny za operacje plikowe na modelach.
 - Komunikacja z Core protokołem REST + SSE (Server-Sent Events): Core zleca zadania przez REST, a agent strumieniuuje postęp i wyniki przez SSE.
 - Pobieranie modeli od dostawców (np. HuggingFace) oraz z magazynów S3/MinIO na lokalny węzeł; wysyłanie modeli z węzła do magazynów zewnętrznych oraz ich usuwanie.
 - Weryfikacja sum kontrolnych (integralność plików) i automatyczne ponawianie nieudanych operacji; anulowanie zadań w toku oraz idempotentność.
 - Strumieniowanie zdarzeń w czasie rzeczywistym: postęp (procent ukończenia, liczba plików, szacowany czas), sygnały „heartbeat”, zdarzenia wyników (success / błąd / częściowe / anulowane) oraz zdarzenia stanu modeli.
 - Inteligentna synchronizacja i odzyskiwanie: cykliczne skanowanie zasobów na dysku i uzgadnianie stanu ze stanem w Core; wykrywanie plików osieroconych oraz uszkodzonych.
 - Efemeryczne poświadczenia dla każdego zadania - wykorzystywane wyłącznie w pamięci operacyjnej na czas trwania operacji i nigdy niezapisywane trwale po stronie agenta.
 - Odporność na awarie - po ponownym połączeniu stan automatycznie uzgadniany, a utracone zadania mogą zostać wznowione.
- 7.2.19. Oczekiwana architektura integracyjna i niezależność od dostawcy (vendor-agnostic)
- Model wtyczkowy do integracji z usługami zewnętrznymi przez wymienne adaptery: orkiestrator kontenerów (wsparcie przynajmniej dla Portainer), system monitoringowy (wsparcie przynajmniej dla Prometheus), magazyn obiektowy

(wsparcie przynajmniej dla S3, MinIO), rejestr modeli (wsparcie przynajmniej dla HuggingFace), brama proxy inferencji, integracja z przynajmniej jednym dostawcą maszyn wirtualnych.

- Możliwość wymiany poszczególnych komponentów infrastrukturalnych bez migracji platformy ani przepisywania logiki biznesowej.
- Wbudowane mechanizmy testowania połączeń i walidacji konfiguracji dla każdej zintegrowanej usługi.
- Architektura oparta na technologii docker jako warstwie bazowej z obsługą Docker Compose i Docker Swarm, z zachowaniem ścieżki ewolucyjnej bez konieczności refaktoryzacji.

7.2.20. Skalowalność i elastyczność operacyjna

- Progresywna skalowalność: od pojedynczego serwera po rozproszone klastry obsługujące dziesiątki węzłów i wielozespołowe środowiska.
- Standaryzacja procesów wdrożeniowych poprzez szablony, redukująca błędy konfiguracyjne i czas onboardingu nowych inżynierów.
- Wsparcie dla heterogenicznego portfolio modeli AI w ramach jednej platformy, eliminujące potrzebę utrzymania osobnych narzędzi per typ modelu.

7.3. Stos do uruchamiania modeli językowych (LLM Stack)

7.3.1. Stos technologiczny do uruchamiania dużych modeli językowych, musi być zoptymalizowanego pod kątem wydajności inferencji oraz efektywnego wykorzystania zasobów sprzętowych. Stos ten musi umożliwiać lokalne uruchamianie modeli AI bez konieczności korzystania z zewnętrznych usług chmurowych. Dostarczone rozwiązanie musi być agnostyczne pod kątem uruchamianych i wykorzystywanych modeli językowych i musi umożliwiać uruchamianie i pracę z modelami językowymi o otwartych wagach w formatach pełnej precyzji, jak również skwantowanych do formatów optymalnych dla wykorzystywanej architektury akceleratorów GPU.

7.3.2. Wymagane komponenty stosu mają zostać uruchomione na wysokowydajnym serwerze inferencji modeli językowych, które mają obsługiwać wiele modeli jednocześnie z niskimi opóźnieniami odpowiedzi. Oprogramowanie do prowadzenia procesu inferencji modeli LLM powinien wspierać popularne architektury modeli, w tym modele transformerowe różnej wielkości, oraz oferować mechanizmy optymalizacji takie jak kwantyzacja, caching uwagi (KV cache) oraz batching inferencji.

7.3.3. Wykonawca ma dobrać modele językowe adekwatne do zadań biznesowych z przeprowadzeniem badania jakości proponowanych modeli i dostarczeniem raportu motywującego podjęte decyzje co do rekomendacji. Wybrane modele językowe powinny być zoptymalizowane pod kątem zastosowań biznesowych przetwarzania danych i dokumentów. Należy uwzględnić modele skwantyzowane minimalizujące zużycie zasobów GPU przy zachowaniu wysokiej jakości odpowiedzi. Wymagany jest również model multimodalny z obsługą przetwarzania wizualnego, przeznaczony w szczególności do zadań optycznego rozpoznawania znaków (OCR) oraz analizy dokumentów zawierających elementy graficzne.

7.3.4. Stos musi oferować jednolitą warstwę abstrakcji umożliwiającą wywoływanie różnych modeli z poziomu standardowego interfejsu API, która to warstwa pozwala na

komunikację z dowolnymi modelami językowymi hostowanymi lokalnie z serwera inferencji w infrastrukturze Zamawiającego bez ograniczeń licencyjnych co do ilości użytkowników i aplikacji oraz typów i dostawców aplikacji korzystających z interfejsu API REST zgodnego ze standardem OpenAI. Wymagane jest umożliwienie wymiany modeli językowych na inne bez konieczności modyfikacji kodu aplikacji.

7.3.5. Platforma musi udostępniać scentralizowaną bramę dostępu (proxy) do modeli językowych z wbudowanymi mechanizmami generowania i zarządzania kluczami API, obsługą zespołów i polityk dostępu, definiowaniem aliasów wskazujących na konkretne modele lub punkty końcowe oraz konfiguracją tras kierujących ruch do właściwych modeli lub dostawców, stanowiącą jednolity punkt wejścia dla wszystkich aplikacji biznesowych integrujących się z platformą.

7.3.6. Platforma musi udostępniać wbudowane mechanizmy testowania połączeń z modelami językowymi, umożliwiające weryfikację poprawności konfiguracji zarówno przed zapisem (bez utrwalania danych), jak i po zapisie adaptera, ze zwrotem czytelnego wyniku, czasu odpowiedzi oraz dodatkowych informacji diagnostycznych, wspierające procesy zapewnienia ciągłości działania systemu w środowisku izolowanym (air-gapped).

7.4. **Portal wewnętrzny z aplikacjami AI (portal AI)**

7.4.1. Zamawiający wymaga dostarczenia zintegrowanego portalu wewnętrznego, stanowiącego punkt dostępu pracowników do narzędzi opartych na sztucznej inteligencji. Portal webowy powinien oferować estetyczny, nowoczesny i ujednolicony interfejs użytkownika z dostępem do dostarczonych aplikacji biznesowych, z zachowaniem obsługi tożsamości oraz kontroli dostępu przynajmniej na poziomie aplikacji biznesowych. Każda z aplikacji musi zawierać podstronę z jasnym opisem zastosowania, korzyści oraz aspektów prywatności i bezpieczeństwa.

7.4.2. Wszystkie dostarczone aplikacje biznesowe muszą wspierać mechanizm kontroli dostępu opartej na rolach. System RBAC musi umożliwiać korzystanie z predefiniowanych ról dla użytkowników, przypisywanie uprawnień do poszczególnych użytkowników oraz zarządzanie dostępem do baz wiedzy poprzez mechanizmy RBAC w dedykowanej aplikacji z interfejsem webowym do zarządzania bazami wiedzy.

7.4.3. Każda z dostarczonych aplikacji musi posiadać interfejs webowy i dokumentację w języku polskim.

7.5. **Aplikacja do agregacji wiedzy i konwersacji z asystentem AI**

7.5.1. Pierwszym komponentem portalu jest aplikacja umożliwiająca agregację rozproszonej wiedzy organizacji oraz interakcję z asystentem AI poprzez interfejs konwersacyjny. Aplikacja musi integrować dane z różnych źródeł w ramach jednolitego środowiska, umożliwiając użytkownikom zadawanie pytań w języku naturalnym i otrzymywanie odpowiedzi w kontekście zgromadzonej wiedzy. Oczekiwane źródła to pliki płaskie w różnych formatach oraz możliwość integracji z kluczowymi aplikacjami biznesowymi.

7.5.2. Wymagane funkcjonalności obejmują mechanizm agregacji dokumentacji wewnętrznej z różnych źródeł systemowych, interfejs czatu umożliwiający prowadzenie konwersacji w języku naturalnym, system generowania odpowiedzi z wykorzystaniem techniki

Retrieval-Augmented Generation (RAG) oraz prezentację źródeł wykorzystanych do wygenerowania odpowiedzi w trzech trybach:

- hybrydowa baza wektorowa (wyszukiwanie semantyczne) połączona z silnikiem FullTextSearch i FuzzySearch, zapewniająca odporność na literówki, fleksję oraz różnice językowe,
 - tryb agentowy RAG (Agentic RAG) - rozszerzenie systemu RAG o tryb agentowy z pętlą typu tool-calling,
 - indeksowanie oparte na rozumowaniu - page Index,
- oraz prezentację źródeł wykorzystanych do wygenerowania odpowiedzi.

7.5.3. Wszystkie odpowiedzi generowane przez system muszą być weryfikowalne poprzez wskazanie fragmentów dokumentów źródłowych, na których zostały oparte. System musi umożliwiać użytkownikom zgłaszanie przypadków nieprawidłowych odpowiedzi w celu doskonalenia jakości działania asystenta.

7.5.4. Dostarczona aplikacja musi integrować się z systemem zarządzania bazami wiedzy dla asystentów AI opisanym w pkt 7.9, a tworzone w tym systemie bazy wiedzy i agentowe bazy wiedzy muszą być dostępne dla użytkowników natychmiast po ich utworzeniu lub udostępnieniu – działając zgodnie z nadanymi uprawnieniami dostępu w mechanizmie RBAC.

7.6. **Dedykowana aplikacja do pracy z dokumentami z wykorzystaniem OCR, czatu AI i poprawnych operacji na liczbach, umożliwiająca pracę grupową**

7.6.1. Drugim komponentem portalu jest aplikacja webowa umożliwiająca przesyłanie, analizę i przetwarzanie dokumentów z wykorzystaniem technologii optycznego rozpoznawania znaków oraz interakcji z asystentem AI. Aplikacja musi obsługiwać różnorodne formaty plików, w tym dokumenty tekstowe, skany oraz fotografie dokumentów.

7.6.2. Wymagane funkcjonalności obejmują funkcję OCR zdolną do ekstrakcji tekstu z dokumentów o różnej jakości, w tym ze skanów o niższej rozdzielczości oraz fotografii, interfejs umożliwiający zadawanie pytań dotyczących zawartości przesłanych dokumentów, mechanizm generowania streszczeń treści dokumentów oraz możliwość jednoczesnej analizy wielu dokumentów i syntetyzowania informacji z nich w spójną odpowiedź.

7.6.3. Funkcja OCR musi zapewniać funkcję opisu obrazków, schematów, wykresów i diagramów, łącznie z danymi na nich zawartymi i znaczeniem prezentacji danych w sposób użyteczny do prowadzenia wnioskowania przez LLM.

7.6.4. Wymagana jest funkcjonalność eksportu wygenerowanych streszczeń i prowadzonych konwersacji z asystentem AI do formatów umożliwiających dalsze wykorzystanie (minimum: docx, pdf, md).

7.6.5. Aplikacja musi zachowywać strukturę logiczną dokumentów podczas przetwarzania, w szczególności rozróżnianie poszczególnych sekcji, tabel i załączników.

7.6.6. Aplikacja musi umożliwiać pracę grupową i udostępnianie projektów w przestrzeniach roboczych z nadawaniem uprawnień do podglądu lub edycji.

7.6.7. Aplikacja musi posiadać dedykowane mechanizmy zintegrowane w w dostarczonym rozwiązaniu w postaci wydzielonej usługi MCP umożliwiające poprawne realizowanie działań matematycznych opisywanych językiem naturalnym na podstawie informacji zagregowanych w przesłanych do aplikacji dokumentach, wykluczając realizowanie działań

matematycznych zleconych przez LLM obsługującego konwersacje w sposób niedeterministyczny. Wymagane jest aby spełnione zostały następujące:

- Funkcja analizy danych w języku naturalnym: Oprogramowanie umożliwia użytkownikowi zadawanie pytań dotyczących wczytanych zbiorów danych w sposób naturalny (sformułowanych językiem naturalnym dla człowieka), bez konieczności posługiwania się poleceniami programistycznymi lub składnią zapytań bazodanowych. System przetwarza pytanie, wykonuje niezbędne obliczenia na danych źródłowych i zwraca czytelny wynik.
- Wczytywanie i przetwarzanie danych ze źródeł w formatach arkuszy kalkulacyjnych: Oprogramowanie zapewnia możliwość wczytywania danych z plików w powszechnie stosowanych formatach kalkulacyjnych (w tym format obsługujący wiele zakładki w jednym pliku), przynajmniej: xls,xlsx, csv.
- Prezentacja wyników w formatach czytelnych dla użytkownika końcowego: Wyniki przeprowadzonych analiz są zwracane użytkownikowi w formie dostosowanej do charakteru odpowiedzi: w przypadku obliczeń pojedynczych – jako wartość liczbową lub tekstową; w przypadku zestawień – jako tabela w formacie umożliwiającym łatwe odczytanie danych; w przypadku analiz wymagających wizualizacji – jako wykres obrazkowy wbudowany w odpowiedź. System poprzez parametry konfiguracji pozwala na ujawnianie zastosowanego sposobu obliczeń, umożliwiając weryfikację poprawności wyniku.
- Izolacja danych między sesjami oraz mechanizmy kontroli dostępu: Oprogramowanie zapewnia separację danych pomiędzy równoległymi sesjami użytkowników – dane wczytane w jednej sesji nie są dostępne w innej. Dostęp do usługi powinien być zabezpieczony mechanizmem uwierzytelniania na podstawie tokenu.
- Prowadzenie wielokrokowej analizy w ramach spójnej rozmowy: Oprogramowanie umożliwia użytkownikowi prowadzenie analizy w formie rozmowy obejmującej wiele kroków, w której kolejne pytania dotyczą tego samego zestawu danych i uwzględniają kontekst wcześniejszych pytań oraz uzyskanych odpowiedzi. Użytkownik może w dowolnym momencie wyczyścić kontekst prowadzonej rozmowy bez utraty wczytanych danych. System zapewnia dodatkowo podgląd struktury danych oraz listę aktywnych zbiorów danych w ramach bieżącej sesji.

7.7. **Aplikacja do automatycznej transkrypcji i analizy spotkań**

7.7.1. Trzecim komponentem portalu jest aplikacja umożliwiająca automatyczne przetwarzanie nagrań audio ze spotkań na tekst, wraz z funkcjonalnościami analitycznymi i wspomagającymi zarządzanie. Aplikacja musi umożliwiać zarówno nagrywanie spotkań bezpośrednio w aplikacji przeglądarkowej, jak i import plików audio z zewnętrznych źródeł.

7.7.2. Wymagane funkcjonalności obejmują automatyczną transkrypcję mowy na tekst z identyfikacją poszczególnych rozmówców, generowanie streszczeń przebiegu spotkań z wyróżnieniem kluczowych punktów i wniosków, automatyczne tagowanie tematów omawianych podczas spotkań oraz możliwość eksportu zapisów i streszczeń do formatów umożliwiających dalsze wykorzystanie.

- 7.7.3. System musi oferować funkcjonalność automatycznego generowania list zadań wynikających ze spotkań, z przypisaniem odpowiedzialności i terminów realizacji. Wymagana jest możliwość udostępniania zapisów uczestnikom spotkania z możliwością weryfikacji i komentowania treści.
- 7.7.4. Aplikacja musi zapewniać możliwość przeszukiwania treści zapisów spotkań według słów kluczowych, tematów oraz uczestników oraz możliwość eksportu transkrypcji, podsumowania spotkania i wytypowanych zadań do formatu tekstowego. Wymagana jest funkcjonalność nawigacji do konkretnych fragmentów nagrania na podstawie treści transkrypcji.
- 7.7.5. Aplikacja musi szyfrować każde nagranie oddzielnym, losowo wygenerowanym kluczem. Klucz musi być przechowywany na serwerze w postaci zaszyfrowanej i udostępniany wyłącznie uwierzytelnionym procesom oraz użytkownikom posiadającym dostęp do danego spotkania. Do czasu zakończenia szyfrowania dostęp do niezaszyfrowanej wersji pliku musi być ograniczony wyłącznie do procesu przetwarzającego, a po utworzeniu i zweryfikowaniu zaszyfrowanej wersji plik źródłowy musi zostać usunięty.

7.8. Aplikacja webowa do tłumaczenia tekstu i dokumentów

- 7.8.1. Wielojęzyczność - obsługa wielu kombinacji językowych (min. 50+ języków źródłowych i docelowych) z możliwością automatycznego wykrywania języka wejściowego.
- 7.8.2. Formaty dokumentów - obsługa różnych typów plików do tłumaczenia (PDF, DOCX, TXT, MD) z zachowaniem formatowania i struktury dokumentu.
- 7.8.3. Tryby przetwarzania - dwa mechanizmy działania: tryb oparty o klasyczne modele ML (CPU) do prostych zastosowań oraz tryb zaawansowany z wykorzystaniem dużych modeli językowych (LLM/GPU) dla lepszej jakości tłumaczeń.
- 7.8.4. Prywatność danych - całkowicie lokalne przetwarzanie tekstów i dokumentów; żadne dane nie są wysyłane do zewnętrznych serwerów.
- 7.8.5. Interfejs webowy - responsywny panel użytkownika z widokiem obok siebie (side-by-side) umożliwiający porównanie tekstu źródłowego i tłumaczenia, edycję wyniku.
- 7.8.6. Pamięć tłumaczeniowa - przechowywanie historii w lokalnej dla użytkownika korzystającego z aplikacji bazie danych z możliwością wyszukiwania poprzednich tłumaczeń i ponownego ich wykorzystania.

7.9. System zarządzania bazami wiedzy dla asystentów AI

- 7.9.1. Zamawiający wymaga dostarczenia systemu umożliwiającego tworzenie, zarządzanie i utrzymywanie baz wiedzy, które stanowią źródło informacji dla oprogramowania analitycznego i innych aplikacji biznesowych dostarczanych w ramach portalu AI. System musi zapewniać centralne repozytorium dokumentów z automatycznym przetwarzaniem i indeksowaniem.
- 7.9.2. Wymagane funkcjonalności obejmują interfejs dla użytkownika biznesowego (udostępniony jako osobna aplikacja biznesowa w portalu AI) do przesyłania dokumentów w formatach PDF, DOCX oraz TXT, automatyczne przetwarzanie i indeksowanie przesłanych dokumentów, mechanizm definiowania reguł zachowania asystenta AI podczas odpowiadania na pytania (instrukcje systemowe) oraz możliwość

podglądu zawartości baz wiedzy bez konieczności uruchamiania pełnej analizy dokumentów.

7.9.3. System musi oferować funkcjonalność automatycznego generowania streszczeń dokumentów, umożliwiającą szybką weryfikację zawartości załadowanych materiałów. Wymagana jest możliwość konfiguracji zachowań asystenta w zakresie stylu odpowiedzi, uwzględnienia kontekstu organizacyjnego oraz stosowania specyficznej terminologii branżowej.

7.9.4. Bazy wiedzy muszą być zarządzane na poziomie uprawnień, z możliwością definiowania, które bazy są dostępne dla poszczególnych użytkowników lub grup użytkowników. System musi umożliwiać aktualizację zawartości baz wiedzy przez użytkowników biznesowych bez konieczności angażowania zespołów technicznych.

7.10. **Wymagane funkcjonalności oprogramowania zarządzania bazami wiedzy:**

7.10.1. Architektura i przetwarzanie danych:

- Implementacja architektury RAG (Retrieval-Augmented Generation) z izolacją danych i przetwarzaniem w środowisku lokalnym (on-premise / air-gapped), bez zależności od zewnętrznych API ani chmur publicznych.
- Przetwarzanie dokumentów wejściowych (PDF, DOCX, TXT, Markdown) z zastosowaniem algorytmów chunking oraz generowania embeddingów przez lokalne modele specjalistyczne.
- Przetwarzanie OCR zapewniające funkcję opisu obrazków, schematów, wykresów i diagramów, łącznie z danymi na nich zawartymi i znaczeniem prezentowanych danych w sposób użyteczny do prowadzenia wnioskowania przez LLM.
- Wielowarstwowy indeks danych: hybrydowa baza wektorowa (wyszukiwanie semantyczne) połączona z silnikiem FullTextSearch i FuzzySearch, zapewniająca odporność na literówki, fleksję oraz różnice językowe.
- Automatyczne generowanie metadanych i podsumowań dokumentów w fazie indeksowania, wykorzystywanych do optymalizacji kontekstu i trafności wyszukiwania.
- Tryb agentowy RAG (Agentic RAG) - rozszerzenie systemu RAG o tryb agentowy z pętlą typu tool-calling, w której LLM samodzielnie decyduje czego szukać, jak szukać i kiedy ma wystarczający kontekst do udzielenia odpowiedzi. Tryb agentowy ma działać obok (nie zamiast) klasycznego trybu, wykorzystując tę samą bazę wiedzy projektu (wektorowa + pełnotekstowa), ale zapewniając wyższą jakość odpowiedzi w przypadku pytań złożonych, wymagających doprecyzowania, wyszukiwań po nazwach własnych oraz obsługi sprzecznych źródeł.
- PageIndex - indeksowanie oparte na rozumowaniu - mechanizm indeksowania dokumentów odwracający klasyczne podejście wektorowe: zamiast bazy wektorowej buduje hierarchiczny indeks dokumentu (drzewo tematyczne) i nawiguje po nim rozumowaniem LLM. Indeks zachowuje naturalną strukturę dokumentu (rozdziały, sekcje, podsekcje) bez arbitralnego chunkingu, zapewniając precyzyjne mapowanie treści na konkretne strony.
- Architektura dwupoziomowego procesu retrievalu - system wspiera dwie równoległe strategie wyszukiwania:

- Klasyczny RAG (tryb workflow) - jednorazowy retrieval + odpowiedź, oparty o hybrydowe wyszukiwanie w bazie wektorowej i pełnotekstowej,
- Tryb agentowy - iteracyjna pętla tool-calling z wieloma turami wyszukiwania, analizy i syntezy.

7.10.2. Mechanizmy wyszukiwania i generowania odpowiedzi:

- Dynamiczne składanie kontekstu z odnalezionych fragmentów dokumentacji, przekazywane do lokalnych LLM w celu generowania odpowiedzi w języku naturalnym.
- System realizuje mechanizm „Source Grounding”, zapewniający jednoznaczne powiązanie każdej generowanej odpowiedzi z konkretnymi fragmentami dokumentów źródłowych, umożliwiając użytkownikowi weryfikację merytoryczną poprzez cytowania lub odnośniki do oryginalnych sekcji.
- Wsparcie dla wielojęzyczności i specjalizacji domenowej poprzez możliwość przypisania dedykowanego modelu LLM do każdego projektu/bazy wiedzy (np. obsługa przez modele Bielik, Llama, Nemotron, Gemma, gpt-oss, Mistral).
- Strategia wyszukiwania w trybie agentowym - agent LLM operuje na zestawie narzędzi o różnej granularności:
 - wyszukiwanie wzorcem po plikach skonwertowanych z dokumentów źródłowych, pierwszy wybór dla literałów (nazwy własne, identyfikatory, akronimy); pozwala na bezpośrednie przejście do fragmentu,
 - semantyczne/hybrydowe wyszukiwanie w bazie wektorowej z ewaluacją chunków przez dedykowany model; eskalacja wyszukiwania naiwnego lub podstawowa technika dla zapytań użytkownika,
 - odczyt pełnej treści pliku z zakresem linii,
 - odnajdywanie dokumentów w projekcie w przypadku posiadania przez dokumenty metadanych (content_type, daty, język, podsumowanie).
- Mechanizm RAG w trybie agentowym realizuje adaptacyjne strategie wyszukiwania wzorców z eskalacją do retrievalu semantycznego tylko w ściśle określonych przypadkach,
- Dekompozycja pytań złożonych - tryb agentowy rozkłada złożone pytanie na podzadania realizowane w ściśle określonej strategii. Lista podzadań ma stanowić kontrakt zakończenia: odpowiedź jest renderowana dopiero, gdy każde podzadanie ma pokrycie w znalezionym materiale.
- Wielotururowa pętla rozumowania - RAG w trybie agentowym pobiera wąski zakres materiału, przeprowadza analizę, w celu określania dalszych strategii prowadzenia działań. Wymagane są strategie niwelujące pobieranie danych nieadekwatnych i pobieranie danych „na zapas”. Przed syntezą odpowiedzi musi zostać wykonana analiza, czy wszystkie podzadania mają oparcie w materiale.
- Nawigacja przez rozumowanie ma odbywać się poprzez wykorzystanie hierarchicznego indeksu dokumentu. W tym trybie agent RAG na podstawie pytania użytkownika i struktury drzewa rozumuje, które gałęzie mogą zawierać odpowiedź. Aktywnie wykorzystywane mają być strategie optymalizacyjne pobierania

najmniejszych możliwych ilości informacji w pojedynczych przebiegach rozumowania w drzewie hierarchicznym w kolejnych iteracjach pracy agenta. Wynikiem pracy agenta w trybie RAG jest odpowiedź z cytowaniem numerów sekcji i stron, jako źródeł. Agent musi zapewniać precyzyjne odpowiedzi z konkretnymi stronami źródłowymi, nawet z dokumentów liczących setki stron.

- Agent musi posiadać funkcje zarządzanie budżetem kontekstu w trybie RAG agentowego poprzez wielowarstwowy mechanizm ograniczania rozmiaru kontekstu w długich pętlach, w tym m.in. poprzez: Accounting tokenów przed każdym requestem z miękkimi i twardymi limitami, ograniczenia rozmiaru wyników per narzędzie AI, defensive guard na rozmiar z funkcją bezpieczeństwa ograniczającą odpowiedź powyżej ustalonego progu, pruning największych tool results powyżej ustalonego miękkiego limitu z inteligentną strategią zachowania najistotniejszych danych, adaptacyjny sygnał informujący o pozostałym budżecie kontekstu, twarde wymuszenie odpowiedzi bez narzędzi.
- Polityka obsługi błędów w trybie agentowym - klasyfikacja błędów narzędzi z odpowiednią reakcją zapewniającą osiągnięcie poprawności odpowiedzi na zapytanie.
- Mechanizm aktywnego zapobiegania driftowi uwagi modelu w długich pętlach, redagujący historię konwersacji i aktualizujący agent flow w oryginalnym pytaniu użytkownika. Strategia częstotliwości ma być kalibrowana w procesie dostrajania, które zrealizuje Wykonawca.
- Integracja z załącznikami konwersacji - tryb agentowy obsługuje pliki załączone do bieżącej wiadomości jako odrębny zakres danych.

7.10.3. Bezpieczeństwo, prywatność i kontrola dostępu:

- Pełna suwerenność danych: wszystkie modele językowe, bazy wektorowe oraz silniki wyszukiwania działają w zamkniętym środowisku Zamawiającego.
- Implementacja kontroli dostępu RBAC na poziomie warstwy bazodanowej i aplikacyjnej, zapewniająca logiczną separację danych między projektami, jednostkami organizacyjnymi lub użytkownikami.
- Możliwość audytu i logowania procesu generowania odpowiedzi, w tym rejestracja wykorzystanych fragmentów źródłowych, parametrów modelu oraz ścieżki decyzyjnej systemu.
- Osobna autoryzacja dla trybu agentowego na poziomie projektu.
- Tracing i audyt pętli agentowej - pełne logowanie przebiegu pętli tool-calling w dedykowanym systemie observability, umożliwiające weryfikację jakości decyzji narzędziowych i optymalizację działania systemu. Trace obejmuje wszystkie istotne elementy procesu.

7.10.4. Konfiguracja, personalizacja i zarządzanie wiedzą:

- Izolacja logiczna baz wiedzy (projektów) z niezależną konfiguracją: wybór modelu LLM, ton wypowiedzi, reguły budowania zapytań, udzielania odpowiedzi oraz kontekst działania asystenta.
- Automatyczna analiza zawartości zindeksowanych dokumentów i na tej podstawie generowanie propozycji optymalnej konfiguracji systemu, modyfikowalnej przez administratora z kontrolą wersji.

- Możliwość ręcznego lub automatycznego definiowania ogólnych zasad działania asystenta, ograniczeń tematycznych oraz procedur walidacji odpowiedzi przed ich wyświetleniem.
- Konfiguracja trybu agentowego - dedykowane parametry sterujące zachowaniem agenta, w tym przynajmniej: określenie modelu prowadzącego reasoning i podejmujący decyzje narzędziowe, określenie temperatury modelu, określenie liczby par wiadomość-odpowiedź w historii konwersacji, określenie maksymalnej liczby tur wywołań narzędzi dla każdego z zapytań z osobna. Niezależne od modelu prowadzącego reasoning określenie modelu do ewaluacji danych z baz wektorowych i semantycznych.
- Specjalizacja asystenta - dedykowane pola konfiguracji projektu wpływające na zachowanie agenta przynajmniej poprzez określenie: roli asystenta w domenie projektu oraz styl, ograniczenia i preferencje finalnej odpowiedzi.
- Indeksowanie dokumentów w specjalizowanym trybie hierarchicznym - automatyczny proces przekształcenia dokumentu w hierarchiczną strukturę drzewiastą realizowane przez przynajmniej: ekstrakcję tekstu z unikalnym oznaczeniem każdej strony, wykrywanie/sprawozdawanie spisu treści, weryfikację poprawności mapowania sekcji na strony z cyklem naprawczym, dekompozycję dużych węzłów rekurencyjnie na mniejsze podwęzły, możliwość wzbogacania o przynajmniej: unikalne ID węzłów, streszczenia, opis dokumentu.
- Wynik indeksowania ma dostarczać ustrukturyzowane dane, zawierające drzewiastą strukturę dokumentu, gdzie każdy węzeł reprezentuje sekcję z precyzyjnym zakresem stron, streszczeniem i rekurencyjnymi podsekcjami. Wynik zachowuje naturalną strukturę dokumentu - granice węzłów odpowiadają granicom tematycznym, nie arbitralnym podziałom.

7.10.5. Integracja, skalowalność i wdrożenie:

- Interfejsy API oraz kompatybilność z popularnymi frameworkami i interfejsami użytkownika (np. openwebui), umożliwiającą przyszłą integrację ze stosami aplikacyjnymi i systemami specjalistycznymi.
- Architektura modułowa pozwalająca na wykorzystanie platformy jako źródła kontekstu dla innych systemów AI (w tym: narzędzia wspierające wytwarzanie oprogramowania, oprogramowanie do pracy z dokumentami, agentyczne pętle decyzyjne).
- Pipe Function dla Open WebUI - dedykowana integracja trybu agentowego z interfejsem Open WebUI poprzez odrębny pipe. Pipe obsługuje przynajmniej: protokół SSE ze strumieniowaniem zdarzeń, renderowanie wywołań narzędzi z sygnalizacją statusu i podsumowaniem, sanitację historii przed wysłaniem do backendu, mechanizm cancel przez dedykowany endpoint.
- Konfigurowalność limitów si timeoutów – w tym możliwość ustawienia niezależnie od siebie przynajmniej: maksymalnej liczby tur wywołań narzędzi, timeout wywołania LLM, timeout dla narzędzi AI, całkowitego timeoutu pętli, budżetu kontekstu (progi soft/hard tokenów).

7.11. Generowanie raportów zbiorczych²

7.11.1. Architektura systemowa i model przetwarzania

- Architektura oparta na deklaratywnym modelu przetwarzania: System realizuje przetwarzanie dokumentów źródłowych w oparciu o zdefiniowany graficzny model przepływu zadań z jawnymi zależnościami między etapami, umożliwiającą modyfikację procesu generowania raportu bez zmiany mechanizmu wykonawczego.
- Rozdzielenie warstw odpowiedzialności: System implementuje trójwarstwową architekturę z wyraźnym podziałem na warstwę definicji procesu, warstwę wykonawczą orkiestrującą przebieg zadań oraz warstwę realizacji pojedynczych zadań redakcyjnych - każda warstwa operuje niezależnie bez wiedzy o wewnętrznych szczegółach pozostałych warstw.
- Izolacja dostępu do materiałów źródłowych: System zapewnia rozdzielenie katalogu dokumentów źródłowych (wyłącznie do odczytu, współdzielony między przebiegami) od katalogu wyników pośrednich (tworzony na czas pojedynczego przebiegu, zapis wyłącznie przez warstwę wykonawczą), z możliwością dynamicznego zawężania widoczności plików per zadanie.
- Komunikacja między zadaniami wyłącznie przez artefakty: Wszystkie zadania porozumiewają się poprzez pliki pośrednie zapisane w dedykowanym katalogu wyników; żadne zadanie nie komunikuje się w pamięci z innymi, a istnienie artefaktu oznacza ukończenie zadania.

7.11.2. Silnik redakcyjny i przetwarzanie dokumentów

- Sekwencyjne, wyczerpujące przetwarzanie dokumentów źródłowych: System realizuje tryb przetwarzania, w którym pojedyncze zadanie redakcyjne czyta przypisany materiał od początku do końca w sposób ciągły, zapewniając kompletność opracowania bez pomijania fragmentów.
- Emisja przyrostowa wyników: System buduje artefakty pośrednie w trybie przyrostowym, dopisując do bufora wyników w miarę postępu czytania materiału, z możliwością domknięcia przebiegu wynikiem częściowym przy przekroczeniu limitów kontekstowych.
- Tryb ukierunkowanego wyszukiwania dla precyzyjnych zadań: Dla zadań o jednoznacznie zdefiniowanym zakresie system wspiera tryb wybiórczego przeszukiwania materiału pod kątem określonych elementów, optymalizując czas przetwarzania w przypadku wąskich kryteriów wyszukiwania.
- Automatyczna klasyfikacja materiału źródłowego: System dynamicznie wyprowadza z korpusu dokumentów źródłowych taksonomię obszarów tematycznych i kategorii problemowych, która jest przekazywana do kolejnych etapów przetwarzania - bez wpisywania klasyfikacji na stałe.
- Standaryzacja numeracji i identyfikacji pozycji: System nadaje każdej wyodrębnionej pozycji (nieprawidłowości, ustalenia, dane finansowe, wnioski) jednoznaczny identyfikator zbudowany z oznaczenia źródła i typu pozycji, umożliwiając precyzyjne odwoływanie się w kolejnych etapach.

7.11.3. Kontrola jakości i wiarygodność wyników

² Kryterium punktowane

- Zasada faktycznej wyłączności źródeł: System gwarantuje, że wszystkie twierdzenia w generowanym raporcie pochodzą wyłącznie z materiałów źródłowych; brak danych w źródłach jest oznaczany jawną formułą, bez uzupełniania treści z wiedzy wewnętrznej systemu.
- Domknięty łańcuch atrybucji: Każde twierdzenie w raporcie jest przechodnio sprowadzalne do konkretnego miejsca w dokumencie źródłowym; warstwy pośrednie zachowują identyfikatory pozycji i oznaczenia źródeł tak, by ścieżka pochodzenia nie ulegała przerwaniu.
- Mechanizm walidacji strukturalnej: System weryfikuje zgodność artefaktów pośrednich z deklarowanym kontraktem wyjściowym (obecność wymaganych pól nagłówka, ciągłość numeracji pozycji w obrębie typu, poprawność parsowania na pozycje), kierując niezgodne zadania do wznowienia.
- Redakcja z zachowaniem integralności danych liczbowych: System realizuje podział odpowiedzialności, w którym ekstrakcja i opis treści wykonywane są przez modele językowe, natomiast operacje arytmetyczne (zliczenia, odsetki, sumy kwot, macierze zestawień) realizowane są deterministycznie przez kod.
- Rejestr kontrolny pokrycia: System generuje na zakończenie przebiegu rejestr umożliwiający weryfikację, czy deklarowane pozycje zostały faktycznie uwzględnione w wynikowym raporcie.

7.11.4. Zarządzanie przebiegiem i mechanizmy stabilności

- Gwarancja ukończenia przebiegu: System zapewnia, że każdy przebieg generowania raportu dochodzi do końca; błędy w pojedynczych zadaniach nie przerywają całkowitego procesu, lecz skutkują wynikiem zdegradowanym z odpowiednią adnotacją.
- Wznawialność zamiast powtarzania: System umożliwia wznowienie przerwane przebiegu z zachowaniem dotychczasowych artefaktów pośrednich.
- Idempotentność uruchomień: Wielokrotne uruchomienie tego samego przebiegu daje przewidywalny, stabilny wynik; zapis nowego artefaktu automatycznie usuwa poprzedni, eliminując duplikację.
- Mechanizm artefaktów zastępczych: Gdy brakuje podzbioru danych do zadania, system tworzy jawny artefakt zastępczy, przerywający kaskadę braków i umożliwiający kontynuację dalszych zadań mimo niekompletnego zestawu.
- Naprawa strukturalna przed wznowieniem: Przed ponownym uruchomieniem zadania system wykonuje beztreściową korektę strukturalną artefaktów, minimalizując potrzebę kosztownych ponownych przebiegów modeli.
- Szukanie zgodności w dopasowaniu etykiet: System toleruje drobne niedokładności w nazewnictwie etykiet (obszarów, jednostek) poprzez wielopoziomowe – przynajmniej 3 poziome, dopasowanie.
- Konfigurowalny tryb pracy: System wspiera tryb miękki (ostrzeżenia zamiast przerw, degradacja zamiast niepowodzenia) oraz tryb ścisły (naruszenie kontraktu odrzuca, niepowodzenie kończy przebieg), z możliwością wyboru zależnie od kontekstu.

7.11.5. Optymalizacja zasobów i konfiguracja modeli

- Dobór modeli zgodnie z charakterem zadania: System umożliwia przypisanie różnych modeli językowych do różnych typów zadań redakcyjnych - zadania

równoległe i specjalizowane otrzymują specjalizowane i dopasowane do zadania modele językowe z odpowiednio dobranymi formatami i technikami optymalizacji; zadania wymagające głębszego osądu (scalenia, oceny) delegowane są do modeli z większymi wymaganiami pamięciowymi i obliczeniowymi oraz inną strategią zarządzania kontekstem.

- Szeregowy łańcuch scalający: Zadania łączące materiał z wielu źródeł wykonywane muszą być zawsze szeregowo, niezależnie od ustawionego poziomu równoległości, zapewniając stabilność zasobów i spójność wyników.
- Równoległość zadań rozgałęzionych: Zadania niepowiązane ze sobą zależnościami wykonują się równoległe zgodnie z limitem konfigurowanym w planie przebiegu; zadania rozgałęzione (osobne instancje dla poszczególnych dokumentów jednostek) pracują niezależnie w ramach swojej grupy.
- Modyfikacja parametrów bez restartu: Możliwość modyfikacji kryteriów selekcji, wag i konfiguracji bez konieczności restartu usług lub migracji schematu bazy danych; zmiany obowiązują od kolejnego uruchomienia.

7.11.6. Bezpieczeństwo, izolacja i audyt

- Zawężanie widoczności materiałów per zadanie: System egzekwuje dla każdego zadania jawną listę plików, do których ma dostęp, osobno dla każdego rodzaju operacji (odczyt, wyszukiwanie); plik spoza listy jest nieodróżnialny od nieistniejącego, bez komunikatu o polityce dostępu.
- Izolacja materiału wzorcowego: Materiał wzorcowy (np. opublikowane raporty tej samej klasy) może służyć jako źródło struktury i stylu, lecz nie jako źródło faktów ani liczb - odcięcie realizowane strukturalnie przez kontrolę dostępu.
- Brak przeszukiwania semantycznego w katalogu wyników: System nie wykonuje wyszukiwania treści w katalogu wyników pośrednich - agent nie może pobrać własnej wcześniejszej syntezy jako „źródła”.
- Pełna ścieżka audytowa przebiegu: System generuje raport z przebiegu obejmujący status każdego zadania, liczbę prób i tur, czas wykonania oraz liczbę ostrzeżeń, umożliwiając pełną weryfikację i diagnostykę procesu.

7.11.7. Wymagania нефункциональные и ограничения системные

- Skalowalność przetwarzania rozgałęzionego: System obsługuje dowolną liczbę dokumentów źródłowych poprzez automatyczne rozgałęzienie zadań ekstrakcji i redakcji - każdy dokument przetwarzany jest przez niezależną instancję, a skalowalność regulowana jest granulacją zadań w modelu przepływu.
- Pamięciochłonność zadania jako budżet: Maksymalna objętość materiału przetwarzanego w pojedynczym zadaniu mieści się w budżecie kontekstowym modelu językowego; przekroczenie skutkuje automatycznym podziałem na mniejsze fragmenty z automatycznym scalaniem przez warstwę wykonawczą.
- Determinizm warstwy danych: Operacje arytmetyczne i selekcyjne wykonywane przez kod dają ten sam wynik przy tych samych danych wejściowych; konfiguracja operacji agregacji jest deklaratoryjna, a nie wpisana w kod na stałe.
- Obserwowalność przebiegów: System zapewnia pełen wgląd w przebieg generowania raportu, w tym: ślad przebiegu z zapisem odcinków odpowiadających kolejnym próbom, liczbą naruszeń i informacją o postępie; zapis odrzuconych prób w osobnym katalogu do wglądu; raport agregacyjny według statusów zadań.

- Zgodność ze środowiskiem Zamawiającego: System działa w zamkniętym środowisku bez zależności od zewnętrznych usług chmurowych; możliwość funkcjonowania w pełni on-premises oraz w infrastrukturze izolowanej (air-gapped).
- Elastyczność definicji procesu: Nowy typ raportu oznacza nową definicję procesu w modelu przepływu, bez konieczności modyfikacji mechanizmu wykonawczego ani refaktoringu systemu.
- Dedykowane mechanizmy zintegrowane w dostarczonym rozwiązaniu w postaci usługi MCP umożliwiające poprawne realizowanie działań matematycznych opisywanych językiem naturalnym na podstawie informacji zagregowanych w przesłanych do aplikacji dokumentach, wykluczając realizowanie działań matematycznych w sposób niedeterministyczny z funkcjonalnościami opisanymi w pkt 7.6

7.12. Licencje na oprogramowanie

- 7.12.1. Wykonawca dostarczy Zamawiającemu wszystkie niezbędne do działania Systemu licencje dla komponentów firm trzecich wykorzystanych w systemie (zarówno Open Source, jak i komercyjne).
- 7.12.2. Wraz z licencjami komercyjnymi Wykonawca zapewni wsparcie techniczne producentów tego oprogramowania zgodnie z wymaganiami określonymi w pkt 12 na okres udzielonej gwarancji na System.
- 7.12.3. Dla oprogramowania własnego Wykonawcy, wymagane jest dostarczenie licencji bezterminowych w ilości niezbędnej do prawidłowego działania dostarczonego Systemu zgodnie z wymaganiami określonymi w OPZ.
- 7.12.4. Wykonawca zobowiązany jest do przekazania Zamawiającemu kodu źródłowego oprogramowania własnego wraz z dokumentacją instalacyjną i instrukcją kompilacji w formie tzw. depozytu bezpośredniego.
- 7.12.5. Kod źródłowy będzie przechowywany przez Zamawiającego w sposób poufny. Zamawiający uzyska prawo do jego użycia oraz modyfikacji wyłącznie w przypadku zaprzestania działalności przez Wykonawcę lub zaprzestania świadczenia usług wsparcia technicznego lub rozwoju dla tego oprogramowania.
- 7.12.6. W terminie do 14 dni od dnia podpisania Protokołu odbioru Wdrożenia, Wykonawca prześle Zamawiającemu aktualny kod źródłowy oprogramowania własnego na zabezpieczonym, zaszyfrowanym nośniku danych (lub: wgra do dedykowanego, repozytorium wskazanego przez Zamawiającego).
- 7.12.7. Wraz z kodem źródłowym Wykonawca dostarczy szczegółową instrukcję kompilacji oraz wykaz narzędzi, kompilatorów i bibliotek (wraz z ich wersjami) niezbędnych do poprawnego zbudowania i uruchomienia wersji produkcyjnej oprogramowania.
- 7.12.8. Wykonawca zobowiązuje się do aktualizacji kodu źródłowego w depozycie po każdej istotnej modyfikacji Systemu, wdrożeniu nowej wersji lub usunięciu błędów krytycznych, nie rzadziej niż raz na 6 miesięcy.
- 7.12.9. Zamawiający zobowiązuje się do zachowania przekazanego kodu w poufności i zabezpieczenia go przed dostępem osób trzecich.
- 7.12.10. Zamawiający staje się uprawniony do jego pełnego wykorzystania i modyfikacji w przypadku wystąpienia któregośkolwiek z niżej wymienionych zdarzeń:
 - likwidacji, ogłoszenia upadłości lub faktycznego zaprzestania prowadzenia działalności gospodarczej przez Wykonawcę.

- zakończenia przez Wykonawcę rozwoju lub wsparcia technicznego dla oprogramowania własnego (tzw. End of Life / End of Support).
 - uchylania się przez Wykonawcę od realizacji obowiązków gwarancyjnych lub asysty technicznej przez okres dłuższy niż 60 dni od zgłoszenia, pomimo pisemnego wezwania ze strony Zamawiającego.
- 7.12.11. Z chwilą wystąpienia któregośkolwiek ze zdarzeń, o których mowa powyżej, Wykonawca udziela Zamawiającemu bezterminowej, niewyłącznej i nieograniczonej terytorialnie licencji na korzystanie z kodu źródłowego. Licencja ta uprawnia Zamawiającego (oraz działające na jego zlecenie podmioty trzecie) do:
- dowolnego modyfikowania, adaptacji, tłumaczenia, rozwoju oraz naprawiania błędów w kodzie źródłowym w celu zapewnienia ciągłości działania Systemu.
 - zwielokrotniania kodu źródłowego w zakresie niezbędnym do jego modyfikacji i uruchamiania.
- 7.12.12. Wynagrodzenie z tytułu licencji i praw, o których mowa powyżej, zawiera się w całkowitym wynagrodzeniu Wykonawcy określonym w Umowie.

7.13. Wymagania w zakresie bezpieczeństwa Systemu

- 7.13.1. Dostarczony System AI, w tym wykorzystywane modele oraz wszelkie komponenty towarzyszące, musi być zdolny do działania w środowisku całkowicie odizolowanym od sieci Internet (środowisko typu air-gapped). Niedopuszczalne jest wymaganie przez oprogramowanie połączenia zewnętrznego do weryfikacji licencji, pobierania aktualizacji w locie, czy wysyłania jakichkolwiek danych telemetrycznych, statystycznych lub diagnostycznych do Wykonawcy lub podmiotów trzecich.
- 7.13.2. System musi zapewniać ścisłą izolację uprawnień i rozliczalność operacji.
- Każde zapytanie użytkownika do modelu, a także każde automatyczne wywołanie narzędzi zewnętrznych (np. baz danych czy zasobów poprzez serwery MCP/RAG), musi być wykonywane wyłącznie w granicach uprawnień przypisanych do tożsamości zalogowanego użytkownika.
 - System musi dynamicznie filtrować dane zasilające kontekst modelu (bazy wiedzy, indeksy wektorowe) tak, aby sztuczna inteligencja w generowanej odpowiedzi nie wykorzystwała informacji, do których dany użytkownik nie ma jawnego prawa wglądu w systemach źródłowych. Ponadto wywołanie jakichkolwiek akcji modyfikujących (np. zapis, edycja, usunięcie danych) przez agentów AI wymaga wdrożenia mechanizmu Human-in-the-loop – każda taka operacja musi być jawnie autoryzowana i zatwierdzana przez człowieka.
- 7.13.3. Wykonawca musi zagwarantować, że dostarczone oprogramowanie oraz zawarte w nim modele AI są w pełni zgodne z wymaganiami Rozporządzenia Parlamentu Europejskiego i Rady ustanawiającego zharmonizowane przepisy dotyczące sztucznej inteligencji (Akt o sztucznej inteligencji / AI Act) adekwatnie do przypisanej systemowi klasy ryzyka oraz ról pełnionych przez Strony. Wykonawca dostarczy pełną dokumentację techniczną Systemu, podsumowanie strategii przestrzegania unijnego prawa autorskiego przez dostawców modeli bazowych oraz deklarację zgodności CE.
- 7.13.4. System musi być zbudowany zgodnie z zasadami Privacy by Design i Privacy by Default, gwarantując pełną kontrolę nad lokalnym przepływem danych oraz

techniczne wsparcie dla realizacji praw użytkowników wynikających z RODO (w tym prawa do bycia zapomnianym w bazach wiedzy i indeksach wektorowych).

- 7.13.5. Wykonawca zobowiązuje się dostarczyć pełną dokumentację opisującą architekturę przepływu danych wewnątrz Systemu oraz aktywnie uczestniczyć i wspierać Zamawiającego w przeprowadzeniu oceny skutków dla ochrony danych (DPIA – Data Protection Impact Assessment) przed produkcyjnym wdrożeniem Systemu.
- 7.13.6. Wykonawca dostarczy Raport z Testów Bezpieczeństwa (AI Red Teaming), potwierdzający odporność Systemu na 10 kluczowych podatności według standardu OWASP Top 10 for LLM Applications. Testy muszą zostać przeprowadzone przez certyfikowanych, niezależnych audytorów bezpieczeństwa (zewnętrznych wobec zespołu projektowego Systemu Wykonawcy). Raport musi wykazać brak podatności o krytycznym (Critical) i wysokim (High) poziomie istotności według metodologii CVSS). Testy muszą być przeprowadzane nie rzadziej niż 1 raz w roku a ich wynik dostarczany Zamawiającemu do upływu 12,24,36 miesięcy od zakończenia wdrożenia.
- 7.13.7. System musi posiadać wbudowane mechanizmy zarządzania zasobami obliczeniowymi i kolejkami zapytań (Rate Limiting / Concurrency Limits), zapobiegające przeciążeniu lokalnej infrastruktury sprzętowej Zamawiającego (zgodnie z podatnością OWASP LLM10). Administrator Zamawiającego musi mieć możliwość definiowania twardych limitów (maksymalna liczba jednocześnie przetwarzanych zapytań, maksymalna długość promptu wejściowego i wyjściowego wyrażona w tokenach) per użytkownik lub per rola systemowa.
- 7.13.8. Wszystkie wbudowane oraz dostarczone w ramach Wdrożenia modele AI (w tym modele językowe LLM, modele osadzeń / Embeddings oraz modele przeszeregowujące / Rerankers) muszą być dostarczone wyłącznie w bezpiecznych formatach niepozwalających na zdalne wykonanie kodu w systemie operacyjnym podczas ich ładowania (wymagany standard formatu: .safetensors lub GGUF; zakaz stosowania formatów bazujących na serializacji kodu, takich jak .pickle).

8. Wymagania w zakresie Wdrożenia

- 8.1. Instalacja dostarczonego sprzętu w serwerowni Zamawiającego, w siedzibie NIK w Warszawie.
- 8.2. Konfiguracja wszystkich niezbędnych komponentów oprogramowania Systemu.
- 8.3. Integracja Systemu z:
 - 8.3.1.usługą katalogową AD w zakresie logowania użytkowników,
 - 8.3.2.Systemem pocztowym Exchange w zakresie wysyłania powiadomień e-mail,
 - 8.3.3.Systemem SIEM w zakresie przekazywanie logów,
- 8.4. Parametryzacja Systemu w celu uzyskania wymaganych funkcjonalności jak i optymalnej wydajności zgodnie z najlepszymi praktykami oraz wiedzą Wykonawcy.

- 8.5. Konfiguracja wykonania kopii zapasowych Systemu oraz jej odtworzenie w celu sprawdzenia poprawności działania mechanizmu kopii zapasowych³.
- 8.6. Potwierdzeniem zakończenia Wdrożenia będzie Protokół odbioru podpisany bez uwag przez Przedstawicieli Zamawiającego i Wykonawcy, wg wzoru określonego w Umowie.

9. Wymagania dotyczące wytwarzanych i przetwarzanych dokumentów w tym Dokumentacji powykonawczej systemu

- 9.1. Zamawiający wymaga, aby wszystkie dokumenty tworzone w ramach realizacji projektu charakteryzowały się wysoką jakością, w szczególności:
- 9.1.1. czytelną i zrozumiałą strukturą zarówno poszczególnych dokumentów jak i całej dokumentacji z podziałem na rozdziały podrozdziały i sekcje,
 - 9.1.2. zachowaniem standardów oraz sposobu pisania, rozumianych jako zachowanie jednolitej i spójnej struktury, formy i sposobu prezentacji treści poszczególnych dokumentów, oraz fragmentów tego samego dokumentu jak również całej dokumentacji.
- 9.2. Dokumentacja powykonawcza Systemu, powinna być sporządzona w języku polskim i zawierać w szczególności:
- 9.2.1. szczegółowy opis przeprowadzonych prac instalacyjnych i konfiguracyjnych,
 - 9.2.2. wykaz wszystkich komponentów firm trzecich (zarówno Open Source, jak i komercyjnych) wykorzystanych w Systemie, który musi zawierać dla każdego komponentu:
 - Dokładną nazwę i wersję komponentu,
 - Typ licencji (np. MIT, Apache 2.0, GNU GPL, komercyjna),
 - Informację, czy komponent został zmodyfikowany przez Wykonawcę,
 - Określenie roli komponentu w architekturze Systemu.
 - 9.2.3. procedury eksploatacyjne i awaryjne w szczególności:
 - uruchamiania, zatrzymywania i restartu środowiska AI.
 - monitorowania wydajności infrastruktury sprzętowo-systemowej (w szczególności użycia GPU/VRAM) oraz poprawności pracy środowiska i wszystkich jego komponentów,
 - zarządzania uprawnieniami, dostępem użytkowników oraz rolami (RBAC),
 - aktualizacji, wersjonowania i wgrywania nowych modeli AI,
 - wykonywania i weryfikacji kopii zapasowych (Backup),
 - postępowania w przypadku awarii sprzętowej komponentów obliczeniowych,
 - bezpiecznego restartu systemu i weryfikacji spójności danych po nagłym zaniku zasilania.

³ Wymagana jest integracja z posiadanym przez Zamawiającego systemem kopii zapasowych Veeam Backup lub konfiguracja innego równoważnego mechanizmu backupu. Niezbędną przestrzeń na kopie zapasowe zapewni Zamawiający.

- obsługi błędów,
- reagowania na incydenty bezpieczeństwa, próby nadużyć i nieuprawnionego dostępu,
- awaryjnego odtworzenia środowiska,
- obsługi anomalii w działaniu modeli AI.

9.3. Wraz z dokumentacją powykonawczą Systemu Wykonawca dostarczy:

9.3.1. Dokumentację techniczną i administracyjną Systemu.

9.3.2. Strategię przestrzegania unijnego prawa autorskiego przez dostawców modeli bazowych oraz deklarację zgodności CE.

9.3.3. Raport z Testów Bezpieczeństwa (AI Red Teaming), potwierdzający odporność Systemu na 10 kluczowych podatności według standardu OWASP Top 10 for LLM Applications.

9.4. Dokumentacja musi być weryfikowana i w razie potrzeby aktualizowana podczas prac serwisowych Wykonawcy realizowanych w ramach Gwarancji lub Asysty Technicznej Wykonawcy.

9.5. Zamawiający wymaga, aby cała dokumentacja, o której mowa powyżej, podlegała jego akceptacji.

9.6. Przekazanie Dokumentacji powykonawczej w formie elektronicznej będzie potwierdzone Protokołem odbioru dokumentacji zgodnie ze wzorem określonym w Umowie.

10. Wymagania w zakresie technicznego warsztatu powdrożeniowego

10.1. Wykonawca przeprowadzi techniczny warsztat powdrożeniowy dla jednej grupy pracowników Zamawiającego, składającej się maksymalnie 10 osób, na terenie Warszawy, np. w siedzibie Zamawiającego.

10.2. Termin i miejsce technicznego warsztatu powdrożeniowego zostaną uzgodnione z Zamawiającym z co najmniej tygodniowym wyprzedzeniem.

10.3. Techniczny warsztat powdrożeniowy musi trwać minimum 3 Dni Robocze, minimum 5 godzin efektywnych zajęć każdego dnia prowadzonych w języku polskim.

10.4. Warsztat musi być prowadzony przez inżyniera Wykonawcy lub producenta dostarczonego Systemu, posiadającego wiedzę i doświadczenie w jego zakresie.

10.5. Program technicznego warsztatu powdrożeniowego musi obejmować całość zagadnień związanych z czynnościami administracyjnymi dostarczonego Systemu, w tym:

10.5.1. omówienie przeprowadzonych prac instalacyjno-konfiguracyjnych,

10.5.2. omówienie konfiguracji poszczególnych komponentów Systemu,

- 10.5.3. omówienie czynności administracyjnych Systemu,
- 10.5.4. omówienie czynności w zakresie monitorowanie obciążenia i bezpieczeństwa Systemu,
- 10.5.5. omówienie najczęściej występujących Awarii i Usterek oraz sposobów ich usuwania i zabezpieczania się przed nimi.
- 10.5.6. omówienie funkcjonalności oprogramowania użytkowego.
- 10.6. Program technicznego warsztatu powdrożeniowego musi być zgodny z wykorzystywaną przez Zamawiającego wersją Systemu.
- 10.7. Potwierdzeniem przeprowadzenia warsztatu powdrożeniowego będzie podpisany przez obie strony Protokół odbioru warsztatu zgodny ze wzorem określonym w Umowie.

11. Wymagania dotyczące świadczenia usług wynikających z udzielonej gwarancji

- 11.1. Wykonawca udziela gwarancji na prawidłowe, w pełni zgodne z przeznaczeniem, funkcjonowanie dostarczonego Systemu (wszystkich jego komponentów sprzętowych i programowych) w okresie wskazanym w pkt. 4.5.
- 11.2. Wykonawca ponosi pełną i wyłączną odpowiedzialność za prawidłowe działanie całego Systemu, w tym za działanie wszystkich zintegrowanych w nim komponentów Open Source oraz komercyjnych firm trzecich.
 - 11.2.1. Wykonawca nie może zwolnić się z odpowiedzialności z tytułu Gwarancji lub SLA (usuwania błędów) z powołaniem się na fakt, iż źródłem błędu lub awarii jest komponent dostarczony przez podmiot trzeci lub społeczność Open Source.
 - 11.2.2. W przypadku wykrycia luki bezpieczeństwa (podatności CVE) lub błędu w użytym komponencie, Wykonawca zobowiązany jest w ramach gwarancji do niezwłocznej aktualizacji tego komponentu do wersji bezpiecznej, zastosowania obejścia (workaround) lub zastąpienia go innym, równoważnym komponentem na własny koszt.
- 11.3. Naprawa gwarancyjna zostanie dokonana po uprzedniej nieodpłatnej ocenie zgłoszonej Awarii lub Usterki. Ocena zgłoszenia musi zostać dokonana przez wykwalifikowanego przedstawiciela Wykonawcy, w miejscu użytkowania Systemu.
- 11.4. Wykonawca niezwłocznie po otrzymaniu zgłoszenia przystąpi do analizy zaistniałej sytuacji i podejmie działania zmierzające do usunięcia zgłoszonych nieprawidłowości w działaniu Systemu.
- 11.5. W celu przystąpienia do naprawy przedstawiciel służb serwisowych Wykonawcy zgłosi się do miejsca użytkowania Systemu. Jeśli naprawa w siedzibie

Zamawiającego nie jest możliwa, Wykonawca odbierze uszkodzony element Systemu i dostarczy po naprawie na własny koszt i na własną odpowiedzialność,

- 11.5.1. po zwrocie naprawionego elementu Systemu nastąpi sprawdzenie poprawności funkcjonowania całego Systemu,
- 11.5.2. czasy naprawy (przywrócenia prawidłowego działania Systemu) to:
 - w przypadku Awarii sprzętu lub oprogramowania – do końca następnego dnia roboczego od momentu zgłoszenia,
 - w przypadku Usterki sprzętu lub oprogramowania do końca 4 dnia roboczego od momentu zgłoszenia.
- 11.6. Zamawiający dopuszcza dostarczenie na czas naprawy urządzenia zastępczego o parametrach nie gorszych niż urządzenie podlegające naprawie. W takim przypadku czas naprawy zostaje wydłużony do 10 dni roboczych pod warunkiem że urządzenie zastępcze zostanie dostarczone w terminie wskazanym w poprzednim punkcie. Wykonawca zobowiązany jest do odtworzenia konfiguracji na urządzeniu zastępczym oraz jej usunięcia w momencie zwrotu.
- 11.7. W przypadku Awarii dysków Wykonawca wymieni uszkodzone w ramach gwarancji dyski, jednocześnie uszkodzone dyski pozostają u Zamawiającego i nie podlegają zwrotowi.
- 11.8. W trakcie trwania umowy Zamawiającego z Wykonawcą usługi gwarancyjne producenta, usługi Wsparcia technicznego producenta dla posiadanych przez Zamawiającego urządzeń i oprogramowania, wchodzących w skład Systemu, świadczone będą za pośrednictwem Wykonawcy.
- 11.9. W przypadku wystąpienia wad w opracowanych dokumentach, o których mowa w pkt 9, Zamawiający ma prawo żądać ich usunięcia w terminie 5 Dni Roboczych od daty zawiadomienia Wykonawcy (naniesienie uzupełnień i poprawek na wszystkich egzemplarzach dostarczonych Zamawiającemu), pod rygorem naliczania kar umownych na zasadach określonych w Umowie. Uprawnienia z tytułu rękojmi za wady dokumentacji powykonawczej wygasają wraz z upływem okresu gwarancji.
- 11.10. Potwierdzeniem wykonania usługi gwarancyjnej będzie podpisany przez obie strony protokół odbioru zgodny ze wzorem określonym w Umowie.
- 11.11. Udzielona gwarancja nie może ograniczać praw Zamawiającego do użytkowania Systemu, zgodnie z zasadami sztuki, przez wykwalifikowany personel Zamawiającego.
- 11.12. W ramach Gwarancji Wykonawca zapewni wsparcie w postaci konsultacji telefonicznych i e-mailowych dla administratorów, dotyczących Systemu, w Dni Robocze pod podanym w umowie nr telefonu i adresem e-mail w średnim miesięcznym wymiarze 8 roboczogodzin, przy czym jeden e-mail traktowany jest jak 15 min rozmowy.

12. Wymagania w zakresie świadczenia usług Wsparcia technicznego producenta

- 12.1. Zapewnienie dostępu do nowych wersji Oprogramowania oraz Aktualizacji i poprawek.
- 12.2. Zapewnienie dostępu do programów dodatkowych objętych licencją danego produktu.
- 12.3. Zapewnienie dostępu do dokumentacji i bazy wiedzy o produkcie.
- 12.4. Dostęp w trybie 5x8 (przez 5 dni w tygodniu, 8 godzin dziennie) do inżynierów pomocy technicznej, w języku polskim.

13. Wymagania w zakresie świadczenia Asysty Technicznej Wykonawcy

- 13.1. Asysta Techniczna będzie rozliczana zgodnie ze złożoną ofertą Wykonawcy w zależności od faktycznie przepracowanych godzin, rozliczanych zgodnie ze stawką godzinową zawartą w ofercie.
- 13.2. Możliwa do wykorzystania ilość godzin w ramach Asysty Technicznej wynosi 3 000 godzin.
- 13.3. W ramach Asysty Technicznej będzie zobowiązany do realizacji prac zleconych przez Zamawiającego, w szczególności w zakresie instalacji i konfiguracji nowych wersji oprogramowania składającego się na System, realizacji warsztatów i szkoleń, prowadzenia prac analitycznych, projektowych i programistycznych związanych z konfiguracją i rozwojem Systemu lub innych prac zleconych przez Zamawiającego, dotyczących dostarczonego Systemu.
- 13.4. Zlecenie na usługi będzie wysyłane do Wykonawcy z wyprzedzeniem 3 dni roboczych w stosunku do terminu rozpoczęcia realizacji usługi. W zleceniu określony zostanie oczekiwany zakres usługi.
- 13.5. Zamawiający w uzasadnionych przypadkach, ze względu m.in. na zakres i złożoność, może zlecić wcześniejsze opracowanie projektu dla prac realizowanych w ramach Asysty Technicznej.
- 13.6. W przypadku realizacji zleconych zadań, których realizacja w ocenie Wykonawcy będzie przekraczała 16 godzin roboczych, konieczne jest przed rozpoczęciem tych prac oszacowanie czasu ich trwania, uzyskanie zgody Zamawiającego na ich rozpoczęcie w postaci zlecenia zgodnie ze wzorem określonym w Umowie.
- 13.7. Wzory protokołów zlecenia i odbioru usług Asysty Technicznej są załącznikami do umowy.

- 13.8. W ramach Asysty Technicznej Strony każdorazowo będą ustalały terminy realizacji usługi. W uzasadnionych przypadkach uzgodniony termin realizacji będzie mógł zostać zmieniony za zgodą Stron.
- 13.9. Zamawiający umożliwi Wykonawcy zdalny lub fizyczny dostęp do elementów Systemu w celu wykonania prac realizowanych w ramach Asysty Technicznej.
- 13.10. Czas związany z obsługą Awarii i Usterek Systemu jak i kontakty z producentem rozwiązania w kwestii usuwania Awarii czy Usterek nie wlicza się do płatnych godzin Asysty Technicznej.
- 13.11. Potwierdzeniem wykonania zleconych prac i wykorzystanego czasu w danym okresie rozliczeniowym będzie protokół odbioru prac wykonanych na rzecz Zamawiającego w ramach Asysty Technicznej Wykonawcy podpisany na koniec każdego trzymiesięcznego okresu rozliczeniowego przez obie strony, zgodnie ze wzorem określonym Umowie.

14. Zasady zdalnego dostępu do środowiska informatycznego Zamawiającego

- 14.1. Zdalny dostęp będzie realizowany za pośrednictwem rozwiązania udostępnianego przez Zamawiającego i na zasadach przez niego określonych.
- 14.2. Zdalny dostęp do Systemu będzie przyznany wyłącznie w celu wykonywania prac wynikających z umowy.
- 14.3. Wykonawca przekaże Zamawiającemu listę pracowników wraz z niezbędnymi danymi określonymi każdorazowo przez Zamawiającego (w szczególności: imię, nazwisko, adres e-mail, nr telefonu komórkowego oraz dane jednoznacznie identyfikujące komputer z którego będzie uzyskiwany dostęp) do konfiguracji zdalnego dostępu.
- 14.4. Minimalne warunki techniczne jakie muszą spełniać komputery Wykonawcy wykorzystywane do zdalnego dostępu to:
 - 14.4.1. system operacyjny wspierany przez klienta VPN (np. Windows 10/11, CentOS/RHEL, Ubuntu/Debian, MacOS) z instalowanymi na bieżąco aktualizacjami bezpieczeństwa,
 - 14.4.2. na bieżąco aktualizowane oprogramowanie zabezpieczające w zakresie ochrony przed wirusami i malware,
 - 14.4.3. zainstalowane oprogramowanie VPN w wersji wskazanej przez Zamawiającego.Warunki te mogą być przez Zamawiającego aktualizowane w trakcie realizacji umowy po uprzednim powiadomieniu Wykonawcy.
- 14.5. Zamawiający zastrzega sobie możliwość ograniczenia ilości osób, którym przyznany zostanie zdalny dostęp.

- 14.6. Pracownicy Wykonawcy, którym przyznany został zdalny dostęp zobowiązani są do nie przekazywania danych umożliwiających jego uzyskanie (w szczególności są to: adres systemu, login, hasło, kody jednorazowe) osobom trzecim.
- 14.7. Zamawiający zastrzega sobie możliwość nagrywania wszystkich czynności realizowanych przez pracowników Wykonawcy za pośrednictwem zdalnego dostępu.
- 14.8. Zabrania się wykonywania jakichkolwiek działań, których efektem będzie ukrycie czynności wykonywanych przez Wykonawcę za pomocą zdalnego dostępu. Tego rodzaju postępowanie będzie skutkowało odebraniem uprawnień do zdalnego dostępu.
- 14.9. W przypadku czasowej niedostępności zdalnego dostępu z przyczyn niezależnych od Wykonawcy lub Zamawiającego prace wynikające z umowy będą realizowane przez pracowników Wykonawcy, w siedzibie Zamawiającego.
- 14.10. Wykonawca zobowiązany jest każdorazowo do wcześniejszego uzgodnienia z upoważnionymi pracownikami Zamawiającego zakresu prac realizowanych za pośrednictwem zdalnego dostępu.
- 14.11. Wykonawca zobowiązuje się do podjęcia wszelkich niezbędnych działań, w tym organizacyjnych i technicznych, mających na celu zabezpieczenie sprzętu za pomocą, którego pracownicy Wykonawcy będą realizowali prace za pośrednictwem zdalnego dostępu.

15. Opcja – rozbudowa serwera

- 15.1. Dostawy 4 szt. Kart GPU NVIDIA RTX PRO 6000 Blackwell Server Edition 96GB PCIe Gen5 oraz innych niezbędnych do rozbudowy komponentów serwera i licencji dla dostarczonego oprogramowania wraz z ich instalacją i konfiguracją w dostarczonym serwerze i aktualizacją dokumentacji powykonawczej.
- 15.2. Zamawiający skorzysta z opcji w przypadku zaistnienia potrzeby zwiększenia wydajności dostarczonego Systemu.
- 15.3. Zlecenie Opcji może nastąpić w terminie do 12 miesięcy od dnia podpisania umowy.
- 15.4. Dostawa nastąpi w terminie do 90 dni od daty zamówienia.
- 15.5. Instalacja i konfiguracja nastąpi w terminie do 14 dni od daty dostawy.
- 15.6. Realizacja zamówienia opcjonalnego będzie odbywać się na podstawie informacji (Zlecenia Opcji), którą Zamawiający przekaże Wykonawcy w formie pisemnej lub elektronicznej pod rygorem nieważności. Zamówienie opcjonalne nie będzie stanowiło zobowiązania Zamawiającego do jego udzielenia, jak również nie będzie stanowiło podstawy do dochodzenia przez Wykonawcę roszczeń odszkodowawczych z tytułu niezłożenia tego zamówienia.