

Wyjaśnienia uzupełniające do Opisu założeń przedsięwzięcia informatycznego „Platforma Usług Inteligentnych RF” (PUIRF)

Wyjaśnienia dotyczące:

- klas stosowanych modeli i sposobu realizacji wyszukiwania semantycznego
- zasad trenowania i dostrajania modeli oraz danych wykorzystywanych w tych procesach
- pełnego cyklu życia modeli
- kryteriów wyboru modeli
- odpowiedzialności za wyniki
- zasad nadzoru człowieka.

1. Klasy modeli i architektura przetwarzania AI

Zasada nadrzędna. PUIRF przyjmuje podejście „RAG-first”: wiedza resortu jest dostarczana modelom w warstwie wyszukiwania (Retrieval-Augmented Generation), a nie zapisywana w wagach modeli. Modele nie będą trenowane od podstaw ani domyślnie dostrajane na danych resortu. Decyzja ta ma charakter architektoniczny i ma trzy konsekwencje: dane wrażliwe (w tym objęte tajemnicą skarbową) nie są utrwalane w parametrach modeli, realizacja praw RODO (np. usunięcia danych) sprowadza się do usunięcia dokumentu z indeksu bez konieczności ponownego treningu, a wymiana modelu na nowszy nie powoduje utraty wiedzy organizacyjnej.

1.1. Klasy modeli w architekturze PUIRF

Klasa modelu	Zastosowanie i założenia techniczne
LLM „General” (open-weight)	Analiza, streszczanie, tłumaczenie i generowanie dokumentów dla pracowników RF (grupa A). Modele open-weight uruchamiane lokalnie w Module Silniki LLM, m.in. PLLuM, Bielik oraz modele wielojęzyczne; docelowo min. 3 modele lokalne, wybierane w analizie porównawczej
LLM „Coder” (open-weight)	Weryfikacja i generowanie kodu źródłowego oraz dokumentacji technicznej dla pracowników IT (grupa B). Model wyspecjalizowany w zadaniach programistycznych, oceniany na wewnętrznym zbiorze zadań (przegląd kodu, wykrywanie błędów, zgodność ze standardami CIRF).
Modele embeddingowe + reranker	Wektoryzacja fragmentów dokumentów i zapytań (model wielojęzyczny ze wsparciem j. polskiego) na potrzeby wyszukiwania semantycznego; reranker (cross-encoder) porządkuje wyniki przed przekazaniem do LLM. Wersja modelu embeddingowego jest stabilizowana — jej zmiana wymaga planowanej reindeksacji.
OCR / (opcjonalnie) HTR	Warstwa digitalizacji w Module ETL: konwersja skanów i PDF do tekstu z kontrolą jakości rozpoznania przed indeksacją. HTR (pismo odręczne) zostanie wdrożony wyłącznie, jeżeli pilotaż baz wiedzy wykaże istotny wolumen takich dokumentów.
Modele pomocnicze (NLP)	Klasyfikatory wspierające routing zapytań do właściwego modelu, detekcję danych osobowych/wrażliwych (DLP, anonimizacja) oraz guardrails wejścia i wyjścia (m.in. wykrywanie prompt injection i treści niedozwolonych itp).

1.2. Realizacja wyszukiwania semantycznego

Dokumenty pozyskiwane konektorami (SharePoint MF, EZD PUW MF, repozytorium plików, e-mail itp.) są dzielone na fragmenty (chunking z zachowaniem struktury dokumentu), opatrywane metadanymi oraz uprawnieniami źródłowymi (ACL) i indeksowane w bazie wektorowej. Zapytanie użytkownika jest realizowane jako wyszukiwanie hybrydowe (pełnotekstowe BM25 + wektorowe kNN/HNSW) z filtrowaniem uprawnień przed wyszukaniem (pre-filtering multi-tenant — użytkownik przeszukuje wyłącznie zasoby, do których ma dostęp), następnie reranking i przekazanie najlepszych fragmentów do LLM. Odpowiedź zawiera obowiązkowe odwołania do dokumentów źródłowych; przy braku wystarczających trafień system informuje o braku danych zamiast generować odpowiedź (przeciwdziałanie konfabulacji). Całość ruchu do modeli przechodzi przez Moduł Silniki LLM / Hub LLM, który realizuje routing, limity, wersjonowanie promptów systemowych i pełne logowanie audytowe.

2. Zasady trenowania i dostrajania modeli oraz dane

Przyjęto czteropoziomową hierarchię adaptacji modeli do zadań resortu — od najtańszej i najbezpieczniejszej do najbardziej ingerencyjnej:

Poziom adaptacji	Zasady stosowania	Dane wykorzystywane
1. Inżynieria promptów i szablony	Podstawowy mechanizm specjalizacji usług (szablony pism, instrukcje systemowe, style urzędowe). Prompty systemowe są wersjonowane i testowane	Wzory i szablony dokumentów, instrukcje redakcyjne — bez danych osobowych.

	regresyjnie (aby potwierdzić, że wcześniej działające funkcje nadal działają poprawnie).	
2. RAG (domyślny)	Wiedza merytoryczna dostarczana wyłącznie w czasie zapytania z baz wiedzy 40 departamentów; brak utrwalania danych w modelu.	Dokumenty departamentów pozyskiwane przez ETL po walidacji jakości, deduplikacji i odwzorowaniu ACL.
3. Dostrajanie parametryczne (opcjonalne)	Wyłącznie techniki PEFT (LoRA/QLoRA) i wyłącznie, gdy ewaluacja wykaże lukę jakości niedomykalną poziomami 1–2. Każde dostrojenie wymaga zgody w ramach AI Governance, DPIA oraz ewaluacji przed wdrożeniem. Proces w całości on-premise.	Wyselekcjonowane, zanonimizowane korpusy wewnętrzne (np. pary instrukcja–wzorcowa odpowiedź zatwierdzone przez ekspertów merytorycznych). Wykluczone: dane objęte tajemnicą skarbową, dane osobowe bez anonimizacji.
4. Trening od podstaw	Wykluczony — nie będzie realizowany	—

Informacje zwrotne od użytkowników (oceny odpowiedzi, zgłoszenia błędów) zasilają wyłącznie procesy ewaluacji i doskonalenia promptów/indeksów; nie są wykorzystywane do automatycznego treningu modeli. Dane użytkowników i treści zapytań nie opuszczają infrastruktury resortu.

3. Pełny cykl życia modelu (LLMOps)

Każdy model w PUIRF przechodzi ustandaryzowany, audytowalny cykl życia zarządzany w Module AI Governance i rejestrowany w rejestrze modeli (model registry). Wdrażana do środowiska produkcyjnego jest zawsze kompletna, wersjonowana „konfiguracja odpowiedzi”: model + prompt systemowy + wersja indeksu — co zapewnia odtwarzalność każdej odpowiedzi systemu na potrzeby audytu.

Faza	Zakres i kryteria przejścia
1. Identyfikacja potrzeby	Zdefiniowanie zadania (grupa A/B), wymagań jakościowych, językowych i wydajnościowych; przypisanie właściciela modelu po stronie CIRF.
2. Preselekcja	Przegląd dostępnych modeli open-weight: licencja dopuszczająca użycie w administracji, karta modelu, pochodzenie, wymagania sprzętowe (VRAM, kwantyzacja).
3. Ewaluacja offline	Benchmarki dla j. polskiego oraz wewnętrzny „golden set” zadań resortu; porównanie kandydatów na identycznym zestawie testów.
4. Walidacja bezpieczeństwa	Testy odporności (jailbreak, prompt injection, wyciek danych z kontekstu), przegląd zgodności (RODO, AI Act) w ramach AI Governance.
5. Wdrożenie	Rejestracja w model registry, wdrożenie kanarkowe/równoległe (shadow) na ograniczonej grupie, monitorowanie przed pełnym udostępnieniem.
6. Eksploatacja i monitoring	Ciągły monitoring jakości (próbki odpowiedzi, oceny użytkowników), wydajności (czas odpowiedzi, przepustowość tokenów, użycie GPU) i kosztów; wykrywanie dryfu jakości.
7. Aktualizacja / re-ewaluacja	Nowe wersje modeli przechodzą pełną ścieżkę 2–5; zmiana promptu lub indeksu wymaga testów regresyjnych na golden secie.
8. Wycofanie	Plan deprecjacji z okresem równoległego działania; dla modeli embeddingowych — zaplanowane okno reindeksacji baz wiedzy; archiwizacja konfiguracji na potrzeby audytu.

4. Kryteria wyboru modeli

Analiza porównawcza i wybór modeli (odrębnie dla grup General i Coder oraz dla modeli embeddingowych) zostaną przeprowadzone w fazie PoC (do 30.04.2027) według następujących, ważonych kryteriów:

- licencja i suwerenność — warunki licencji open-weight dopuszczające eksploatację w administracji publicznej bez przekazywania danych licencjodawcy;
- jakość dla języka polskiego — wyniki na benchmarkach dla języka polskiego oraz na wewnętrznym golden secie zadań resortu (jakość merytoryczna, styl urzędowy, terminologia podatkowo-celna);
- wydajność i koszt — przepustowość i opóźnienie inferencji na posiadanych serwerach GPU (tokeny/s, czas do pierwszego tokenu) przy docelowej współbieżności, z uwzględnieniem kwantyzacji (np. INT8/FP8) bez istotnej utraty jakości;
- okno kontekstu — długość okna kontekstu wystarczająca dla dokumentów urzędowych i wyników wyszukiwania RAG;
- bezpieczeństwo i zgodność — odporność na ataki (jailbreak, prompt injection) oraz dostępność dokumentacji wymaganej dla modeli ogólnego przeznaczenia zgodnie z AI Act;

- dojrzałość i utrzymanie — aktywność społeczności/dostawcy, przewidywalność wydań, dostępność wersji poprawkowych (istotne dla 5-letniego okresu trwałości).

4.1. Podstawa doboru modeli open-weight — benchmarki referencyjne

Preselekcja modeli odbywa się na podstawie publicznych, powtarzalnych benchmarków (tabela poniżej); wyniki są w miarę możliwości odtwarzane lokalnie na infrastrukturze CIRF w celu weryfikacji. Publiczne rankingi służą wyłącznie do wyłonienia krótkiej listy kandydatów (top-N), ponieważ obarczone są ryzykiem kontaminacji zbiorów treningowych (model mógł „widzieć” zadania testowe podczas treningu) — rozstrzygająca jest zawsze punktacja na wewnętrznym golden secie zadań resortu (pkt 5).

Obszar oceny	Benchmark (źródło)	Metryka i sposób pomiaru
Jakość języka polskiego (LLM „General”)	Open PL LLM Leaderboard (SpeakLeash) huggingface.co/spaces/speakleash/open_pl_llm_leaderboard KLEJ (NLU dla j. polskiego) klejbenchmark.com	Wynik zbiorczy (Avg) — średnia dokładność i F1 w zadaniach rozumienia i generowania dla j. polskiego; dokładność: $acc = \frac{n_{poprawne}}{n}$
Wiedza „egzaminacyjna” PL	LLMzSzł — egzaminy państwowe (CKE) huggingface.co/spaces/enelpol/llmzszl_leaderboard	Dokładność odpowiedzi na pytania polskich egzaminów państwowych (miara wiedzy realioznawczej i prawno-administracyjnej w j. polskim).
Wiedza ogólna i rozumowanie	Open LLM Leaderboard: MMLU-Pro, IFEval, BBH, MATH, GPQA, MuSR huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard	Dokładność / exact-match, normalizowana względem wyniku losowego; IFEval — odsetek poleceń wykonanych zgodnie z instrukcją (istotne dla przestrzegania promptów systemowych).
Jakość konwersacyjna i preferencje użytkowników	MT-Bench-PL (SpeakLeash) huggingface.co/spaces/speakleash/mt-bench-pl LMArena (ranking Elo) lmarena.ai	MT-Bench: ocena sędziego (LLM-as-judge) w skali 1–10 dla dialogów wieloturuowych. Arena: ranking Elo z porównań parami wg modelu Bradleya–Terry’ego: $P(A > B) = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}}$
Generowanie i weryfikacja kodu (LLM „Coder”)	HumanEval github.com/openai/human-eval BigCodeBench bigcode-bench.github.io SWE-bench swebench.com LiveCodeBench livecodebench.github.io	Metryka funkcjonalnej poprawności kodu: $pass@k = \mathbb{E}_{zad.} [1 - \frac{C(n-c, k)}{C(n, k)}]$ gdzie n — liczba wygenerowanych prób, c — próby przechodzące testy jednostkowe, C — symbol Newtona; raportowane pass@1. SWE-bench: odsetek rzeczywistych zgłoszeń naprawionych z zaliczeniem testów.
Modele embeddingowe i retrieval	MTEB / PL-MTEB (zadania polskie) huggingface.co/spaces/mteb/leaderboard BEIR (retrieval zero-shot) github.com/beir-cellar/beir	nDCG@10 (wzór — pkt 5.1) na zadaniach retrieval dla j. polskiego oraz Recall@100; dodatkowo wymiar wektora i długość wejścia (koszt indeksu).
Dopasowanie do domeny resortu	Wewnętrzny korpus walidacyjny (dokumenty jawne, bez danych wrażliwych)	Perpleksja na korpusie domenowym: $PPL = \exp(-\frac{1}{N} \sum_{i=1}^N \ln p(t_i t_{<i}))$ im niższa, tym lepsze dopasowanie językowe modelu do tekstów urzędowych resortu (t_i — kolejne tokeny korpusu).
Bezpieczeństwo modelu	HarmBench harmbench.org + wewnętrzna biblioteka ataków (pkt 5.3)	ASR — skuteczność ataków (wzór — pkt 5.3) oraz odsetek prawidłowych odmów dla treści niedozwolonych; testy powtarzane po kwantyzacji modelu.

4.2. Procedura punktacji i wyboru

- **kryteria bramkowe (0/1)** — przed punktacją model musi spełnić warunki eliminacyjne: licencja open-weight dopuszczająca użycie i modyfikację w administracji publicznej bez przekazywania danych licencjodawcy, wykonalność uruchomienia on-premise (zapotrzebowanie VRAM po kwantyzacji mieszczące się w pojemności posiadanych GPU), brak wbudowanej telemetrii zewnętrznej oraz dostępność dokumentacji modelu ogólnego przeznaczenia wymaganej przez AI Act; niespełnienie któregośkolwiek warunku eliminuje model niezależnie od wyników jakościowych;

- **normalizacja i agregacja** — wyniki surowe x_i z benchmarków i golden setu są sprowadzane do skali [0,1] normalizacją min-max: $s_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$, a wynik łączny modelu jest sumą ważoną: $S = \sum_{i=1}^m w_i \cdot s_i$, $\sum_{i=1}^m w_i = 1$. Wagi w_i są zatwierdzane przez Komitet AI Governance przed rozpoczęciem ewaluacji (co wyklucza dopasowywanie wag do znanych wyników).
- **wstępne wagi kryteriów (do walidacji w PoC)** — LLM „General”: jakość j. polskiego 0,30; golden set resortu 0,25; rozumowanie i przestrzeganie instrukcji 0,10; wydajność/koszt inferencji 0,15; bezpieczeństwo 0,10; dojrzałość i utrzymanie 0,10. LLM „Coder”: benchmarki kodu 0,35; golden set CIRF (przegląd kodu) 0,25; wydajność 0,15; bezpieczeństwo 0,10; dojrzałość 0,15. Modele embeddingowe: PL-MTEB/nDCG@10 0,45; golden set retrieval 0,30; koszt indeksowania i wymiar wektora 0,25;
- **ścieżka decyzyjna** — z rankingu publicznego wyłaniana jest krótka lista (top-N, wstępnie $N = 8$ dla General, $N = 5$ dla Coder i embeddingów), która przechodzi pełną ewaluację lokalną (golden set, wydajność na docelowych GPU, red teaming); wybierane są modele o najwyższym S z zachowaniem wymogu min. 3 modeli lokalnych. Karta wyboru (wyniki, wagi, uzasadnienie) jest archiwizowana w rejestrze modeli i podlega audytowi;
- **aktualizacja rankingu** — ranking jest odtwarzany przy każdej istotnej publikacji nowych modeli open-weight oraz obowiązkowo przy re-ewaluacji w cyklu życia (pkt 3, faza 7), na tych samych zasadach punktacji.

5. Jakość wyników — ewaluacja i przeciwdziałanie halucynacjom

Dla każdego modułu usługowego (generowanie dokumentów, wyszukiwanie semantyczne, weryfikacja kodu) zostanie zbudowany i utrzymywany „golden set” — reprezentatywny zbiór przypadków testowych opracowany z departamentami merytorycznymi. Stosowane będą metryki właściwe dla systemów RAG: wierność odpowiedzi względem źródeł (groundedness/faithfulness), trafność odpowiedzi względem pytania, precyzja i pokrycie wyszukanego kontekstu, a dla kodu — poprawność wskazań weryfikacji. Ewaluacja prowadzona jest w dwóch trybach: offline (obowiązkowe testy regresyjne przy każdej zmianie modelu, promptu lub indeksu) oraz online (próbkowanie odpowiedzi produkcyjnych z oceną ekspercką i analiza ocen użytkowników).

Mechanizmy ograniczające halucynacje:

wymuszone cytowanie źródeł w każdej odpowiedzi opartej na wiedzy resortu; odmowa odpowiedzi przy braku wystarczających trafień w bazach wiedzy (progi podobieństwa) zamiast generowania treści nieopartej źródłami; guardrails na wejściu i wyjściu (filtrowanie prompt injection, danych osobowych, treści niedozwolonych); oznaczanie wszystkich treści wygenerowanych przez AI. Degradacja jakości odpowiedzi jest objęta rejestrem ryzyk z pętlą informacji zwrotnej.

5.1. Metryki jakości wyszukiwania (retrieval)

Jakość warstwy wyszukiwania jest mierzona na golden secie z adnotacjami istotności (D_{rel} — zbiór fragmentów istotnych dla zapytania, D_k — k fragmentów zwróconych przez system, Q — zbiór zapytań testowych). Wartości docelowe stanowią założenia projektowe i zostaną zwalidowane w PoC technologicznym (KM2):

Metryka	Wzór	Interpretacja i wartość docelowa
Precision@k	$P@k = \frac{ D_{rel} \cap D_k }{k}$	Odsetek istotnych fragmentów wśród k przekazanych do LLM (po rerankingu). Cel: $\geq 0,80$ przy $k = 5$.
Recall@k	$R@k = \frac{ D_{rel} \cap D_k }{ D_{rel} }$	Pokrycie wszystkich fragmentów istotnych dla zapytania (przed rerankiem). Cel: $\geq 0,90$ przy $k = 20$.
MRR (Mean Reciprocal Rank)	$MRR = \frac{1}{ Q } \sum_{i=1}^{ Q } \frac{1}{rank_i}$	Średnia odwrotność pozycji pierwszego istotnego fragmentu ($rank_i$). Cel: $\geq 0,85$.
nDCG@k	$nDCG@k = \frac{DCG@k}{IDCG@k},$ $DCG@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)}$	Jakość uporządkowania rankingu ze stopniowaną istotnością rel_i (0–3). Cel: $\geq 0,85$.
RRF (Reciprocal Rank Fusion)	$RRF(d) = \sum_{r \in R} \frac{1}{k_0 + rank_r(d)}$	Fuzja rankingu leksykalnego (BM25) i wektorowego; R — zbiór rankingów, stała $k_0 = 60$ (strojona w PoC).
Próg podobieństwa (abstencja)	$sim(q, c) = \frac{E(q) \cdot E(c)}{\ E(q)\ \cdot \ E(c)\ } \geq \tau$	Warunek dopuszczenia fragmentu c do kontekstu; gdy żaden fragment nie przekracza progu τ , system

		odmawia odpowiedzi. τ kalibrowane na zbiorze walidacyjnym (pkt 5.3).
--	--	---

5.2. Metryki jakości generowania (odpowiedzi LLM)

Odpowiedzi są oceniane na poziomie pojedynczych stwierdzeń (claims) wyekstrahowanych z odpowiedzi i weryfikowanych względem dostarczonego kontekstu źródłowego:

Metryka	Wzór	Interpretacja i wartość docelowa
Faithfulness (wierność źródłom)	$F = \frac{ Z_{poparte} }{ Z_{ogółem} }$	Odsetek stwierdzeń odpowiedzi (Z) mających oparcie w dostarczonym kontekście — ocena zdanie po zdaniu. Cel: $\geq 0,95$.
Answer relevancy (trafność)	$AR = \frac{1}{N} \sum_{i=1}^N \cos(E(q), E(q'_i))$	Zbieżność semantyczna pytania q i odpowiedzi; q'_i — pytania zrekonstruowane z odpowiedzi, E — model embeddingowy. Cel: $\geq 0,85$.
Context precision / context recall	$CP = \frac{ C_{istotne} \cap C_{uzyte} }{ C_{uzyte} }$	CP — odsetek istotnych fragmentów wśród przekazanych do LLM; analogicznie CR — pokrycie informacji wymaganych przez odpowiedź wzorcową. Cel: CP $\geq 0,75$; CR $\geq 0,90$.
Wskaźnik halucynacji (HR)	$HR = \frac{n_{hal}}{n}$	Odsetek odpowiedzi zawierających co najmniej jedno stwierdzenie bez oparcia w źródłach (n_{hal} z n ocenianych). Cel: $\leq 0,02$ offline; monitorowany na próbie produkcyjnej.
Poprawność cytowań (CA)	$CA = \frac{n_{cyt, popr}}{n_{cyt}}$	Odsetek cytowań wskazujących właściwy dokument i fragment źródłowy. Cel: $\geq 0,95$.
F1 dla weryfikacji kodu (Coder)	$F1 = \frac{2 \cdot P \cdot R}{P + R}$	Średnia harmoniczna precyzji (P) i czułości (R) wykrywania defektów na wewnętrznym zbiorze zadań CIRF. Cel: $\geq 0,80$.

5.3. Zaawansowane mechanizmy sterowania jakością

- dywersyfikacja kontekstu (MMR)** — przy doborze fragmentów kontekstu stosowany jest algorytm Maximal Marginal Relevance: $MMR = \arg \max_{c \in C \setminus S} [\lambda \cdot \text{sim}(q, c) - (1 - \lambda) \cdot \max_{s \in S} \text{sim}(c, s)]$, gdzie C — kandydaci, S — fragmenty już wybrane, $\lambda \approx 0,7$; ogranicza redundancję kontekstu i zwiększa pokrycie informacyjne przy stałym budżecie tokenów.
- kontrolowane dekodowanie** — dla zadań urzędowych stosowane są parametry deterministyczne (temperatura $\leq 0,3$, top-p $\leq 0,9$, stały seed); komplet parametrów dekodowania jest częścią wersjonowanej „konfiguracji odpowiedzi” (pkt 3), co zapewnia odtwarzalność wyników na potrzeby audytu.
- abstencja kalibrowana (selective prediction)** — próg podobieństwa τ (pkt 5.1) jest dobierany na zbiorze walidacyjnym tak, aby odsetek odpowiedzi błędnych wśród udzielonych nie przekraczał przyjętego poziomu ryzyka α (podejście konformalne); poniżej progu system odpowiada „brak wystarczających danych w bazach wiedzy” zamiast generować treść.
- ewaluacja LLM-as-judge z kalibracją** — metryki faithfulness i answer relevancy są wyliczane automatycznie przez model-sędziego; sędzia jest dopuszczany do użycia wyłącznie po kalibracji względem ocen eksperckich (korelacja rang Spearmana $\rho \geq 0,8$). Zgodność ocen eksperckich jest kontrolowana współczynnikiem Cohena $\kappa = \frac{p_o - p_e}{1 - p_e}$ (p_o — zgodność obserwowana, p_e — oczekiwana losowo); wymagane $\kappa \geq 0,7$.
- ciągły red teaming** — odporność mierzona wskaźnikiem skuteczności ataków $ASR = \frac{n_{ataki udane}}{n_{ataki}}$ na utrzymywanej bibliotece ataków (jailbreak, prompt injection, próby wycieku kontekstu i danych osobowych); cel: $ASR \leq 0,01$. Biblioteka jest rozszerzana po każdym incydencie oraz w ramach cyklicznych testów bezpieczeństwa.

5.4. Monitoring operacyjny i wykrywanie dryfu

- dryf danych i zapytań** — rozkłady cech zapytań i embeddingów w bieżącym oknie (q) są porównywane z oknem referencyjnym (p) w B przedziałach wskaźnikiem stabilności populacji: $PSI = \sum_{i=1}^B (p_i - q_i) \cdot \ln \frac{p_i}{q_i}$. Wartość $PSI > 0,2$ uruchamia alert oraz obowiązkową re-ewaluację konfiguracji na golden secie.
- wydajność inferencji** — czas do pierwszego tokenu (TTFT) i czas pełnej odpowiedzi w percentylu P95, przepustowość (tokeny/s) oraz utylizacja GPU przy docelowej współbieżności; progi SLA zostaną potwierdzone Raportem z testów wydajności (KM4).
- sprzężenie z AI Governance** — metryki jakości (pkt 5.1–5.2), oceny użytkowników i wskaźniki operacyjne zasilają telemetrię platformy (m.in. KPI 7) oraz kwartalne przeglądy Komitetu AI Governance; przekroczenie progów skutkuje działaniami naprawczymi zgodnie z cyklem życia modelu (pkt 3).

6. Odpowiedzialność za wyniki i zasady nadzoru człowieka

PUIRF jest narzędziem wspomagającym pracę (human-in-the-loop). System nie podejmuje żadnych decyzji automatycznych i nie wywołuje samodzielnie skutków prawnych ani faktycznych wobec obywateli lub przedsiębiorców. Każdy rezultat działania AI (projekt pisma, odpowiedź z bazy wiedzy, raport z weryfikacji kodu) ma status materiału roboczego, który przed jakimkolwiek użyciem służbowym podlega weryfikacji, korekcie i zatwierdzeniu przez pracownika. Odpowiedzialność merytoryczna za treść dokumentu pozostaje — tak jak dotychczas — po stronie pracownika i jego przełożonych, zgodnie z obowiązującymi w resorcie zasadami podpisywania i akceptacji dokumentów.

6.1. Podział odpowiedzialności

Rola	Odpowiedzialność
Pracownik (użytkownik)	Weryfikacja i zatwierdzenie każdego rezultatu przed użyciem; zgłaszanie błędnych odpowiedzi (wbudowany mechanizm feedbacku).
Departament (właściciel bazy wiedzy)	Aktualność, jakość i klasyfikacja dokumentów zasilających bazę wiedzy; decyzje o zakresie udostępnienia (ACL).
CIRF (właściciel techniczny modeli i platformy)	Cykl życia modeli (pkt 4), monitoring jakości i bezpieczeństwa, SLA, reagowanie na incydenty AI.
Komitet AI Governance (z udziałem IOD)	Polityki użycia AI, zatwierdzanie modeli i przypadków użycia, rejestr ryzyk AI, przeglądy zgodności (RODO, AI Act), decyzje o dostrajaniu.

6.2. Środki nadzoru człowieka i zgodność z AI Act

Nadzór człowieka jest realizowany przez: (1) obowiązkową weryfikację rezultatów przed użyciem (brak ścieżki automatycznego wykorzystania wyników), (2) oznaczanie treści wygenerowanych przez AI, (3) pełną audytowalność — dla każdej odpowiedzi rejestrowane są: użytkownik, czas, wersja modelu i promptu, wykorzystane źródła, (4) możliwość wyłączenia modelu lub usługi przez CIRF w trybie natychmiastowym, (5) program szkoleń budujący kompetencje AI pracowników (AI literacy, zgodnie z art. 4 AI Act) — ujęty w KPI 4 i KPI 5. Zgodnie ze wstępną oceną usługi PUIRF, jako wewnętrzne narzędzia wspomagające, nie mieszczą się w katalogu systemów wysokiego ryzyka załącznika III AI Act; klasyfikacja ta zostanie formalnie zweryfikowana i udokumentowana w ramach Środowiska AI Governance (KM5, do 30.11.2028), z uwzględnieniem rozporządzenia (UE) 2026/1744 oraz wymogów dla modeli ogólnego przeznaczenia. Niezależnie od wyniku klasyfikacji projekt dobrowolnie stosuje środki właściwe dla systemów wyższego ryzyka: dokumentację techniczną modeli, rejestrowanie zdarzeń, nadzór człowieka i zarządzanie ryzykiem AI.

7. Compliance, bezpieczeństwo i ochrona danych

Przetwarzanie danych wyłącznie w infrastrukturze resortu (on-premise, serwery GPU), izolacja danych departamentów (multi-tenant), uwierzytelnianie AD (SSO, MFA) i autoryzacja RBAC (Role-Based Access Control), szyfrowanie w spoczynku i tranzycie, DLP z anonimizacją, guardrails, pełny audyt zapytań i odpowiedzi oraz cykliczne testy bezpieczeństwa. Przed produkcyjnym uruchomieniem usług zostanie przeprowadzona ocena skutków dla ochrony danych — jej podstawą będzie raport z inicjalnego testu prywatności. Retencja logów audytowych oraz zasady postępowania z treściami zapytań zostaną określone w politykach AI Governance zgodnie z zasadą minimalizacji danych (RODO) itp.