

Załącznik stanowiący odpowiedź na uwagi wniesione przez Radę Architektury IT – Karta oceny założeń projektu informatycznego nr P578, dot. projektu pn. „mojeIKP Junior”

Szczegóły dot. wykorzystania AI w budowanym rozwiązaniu – mojeIKP Junior

Przewidziane jest wykorzystanie poniższych technologii:

NLP (Natural Language Processing) - Rozumienie wypowiedzi pacjenta, ekstrakcja objawów, klasyfikacja intencji, generowanie pytań w języku naturalnym.

LLM (Large Language Model) w architekturze generatywnej - Silnik konwersacyjny prowadzący wywiad (metoda OPQRST/SOAP) i generujący listę pytań

RAG (Retrieval-Augmented Generation) - Wzbogacanie odpowiedzi modelu o zweryfikowaną wiedzę kliniczną (wytyczne PTK/PTD, ICD-10, ATC) oraz kontekst z danych pacjenta w P1, bez konieczności trenowania modelu na danych medycznych

Wyszukiwanie semantyczne (semantic search / embeddings) - Realizowane w ramach RAG. Wektoryzacja bazy wiedzy klinicznej (baza wektorowa, embeddingi tekstu) i wyszukiwanie najbardziej relewantnych fragmentów wiedzy do kontekstu rozmowy

Function calling / structured output - Ustrukturyzowane wywołania do API Platformy P1 w celu pobrania kontekstu klinicznego (e-recepty, e-skierowania, rozpoznania, wyniki badań, rejestracje na wizyty)

1. Zasady trenowania modeli

Kluczowa deklaracja: modele bazowe (LLM) nie będą trenowane od podstaw ani fine-tunowane na danych medycznych pacjentów. Przyjęte podejście architektoniczne opiera się na:

- **Wykorzystaniu modeli pre-trenowanych (foundation models)** dostępnych komercyjnie (np. vLLM → Med42 70B / BioMistral 7B), bez modyfikacji wag modelu
- **Metodzie RAG jako głównym mechanizmie personalizacji i aktualizacji wiedzy.** Baza wiedzy klinicznej (wytyczne, ICD-10, ATC) jest aktualizowana niezależnie od modelu, bez potrzeby retreningu
- **Prompt engineering i system promptów** jako mechanizmie kontroli zachowania modelu (definiowanie ról, ograniczeń, zakresu odpowiedzi)
- Opcjonalnie, w fazie rozwojowej: **fine-tuning na wąskim, zanonimizowanym korpusie** przykładowych dialogów medycznych (bez danych pacjentów rzeczywistych), wyłącznie w celu poprawy jakości formułowania pytań, nie do celów diagnostycznych

Taka architektura minimalizuje ryzyko regulacyjne (uniknięcie klasyfikacji jako wyrób medyczny wymagający pełnej walidacji klinicznej modelu) oraz eliminuje konieczność przetwarzania danych osobowych pacjentów w procesie treningowym.

2. Klasyfikacja regulacyjna

- **AI Act (Rozporządzenie UE 2024/1689).** System klasyfikowany jako **wysokiego ryzyka** (Annex III pkt 5) z uwagi na zastosowanie w obszarze ochrony zdrowia. Wymaga: dokumentacji technicznej, rejestru zdarzeń (logging), oceny zgodności, nadzoru ludzkiego (human oversight) oraz testów dokładności przed wdrożeniem produkcyjnym.
- **MDR 2017/745.** Komponent **nie jest klasyfikowany jako wyrób medyczny**, ponieważ nie generuje diagnoz, rekomendacji terapeutycznych ani nie wpływa na decyzje kliniczne. Wspiera wyłącznie komunikację pacjent-lekarz. Klasyfikacja ta będzie podlegać rewalidacji przy każdej zmianie zakresu funkcjonalnego.
- **RODO (art. 9).** Przetwarzanie danych o zdrowiu wymaga przeprowadzenia **oceny skutków dla ochrony danych (DPIA)** przed wdrożeniem; podstawą prawną przetwarzania jest zgoda pacjenta (art. 9 ust. 2 lit. a) lub przesłanka opieki zdrowotnej (lit. h).
- **Ustawa o systemie informacji w ochronie zdrowia (SIOZ).** Integracja z P1 wymaga zgodności z zasadami udostępniania danych z systemu e-zdrowia oraz uzyskania właściwych uzgodnień z Centrum e-Zdrowia jako operatorem P1.

3. Bezpieczeństwo i ochrona danych - mechanizmy techniczne

- **Pseudonimizacja danych przed przesłaniem do dostawcy modelu LLM.** Identyfikatory pacjenta (PESEL, dane osobowe) nie są przekazywane do API modelu w formie jawnej; kontekst kliniczny jest rekonstruowany lokalnie po stronie backendu CeZ
- **Umowa powierzenia przetwarzania danych (DPA) / Business Associate Agreement** z dostawcą usługi modelowej, zapewniająca brak wykorzystania danych do treningu modeli dostawcy oraz przetwarzanie w regionie UE
- **Szyfrowanie end-to-end (TLS 1.3)** w komunikacji między aplikacją, backendem i API modelu
- **Zgodność z wymogami CSIRT CeZ, CERT Polska, NIS2 oraz ISO 27001** w zakresie zarządzania bezpieczeństwem informacji

4. Jakość wyników i weryfikacja odpowiedzi

- **Warstwa guardrails.** Deterministyczne reguły wykrywające stany alarmowe (np. słowa kluczowe wskazujące na stan zagrożenia życia) uruchamiane przed przetworzeniem przez model LLM, z automatycznym przekierowaniem do numerów alarmowych
- **Walidacja treści generowanych.** Mechanizm wykrywania halucynacji poprzez porównanie odpowiedzi modelu z bazą wiedzy RAG (weryfikacja źródłowa)

- **Obowiązkowy disclaimer** w każdej sesji: informacja, że asystent nie zastępuje porady lekarskiej, a lista pytań ma charakter pomocniczy
- **Ewaluacja okresowa.** Testowanie na zestawie referencyjnych przypadków klinicznych (vignette testing) przygotowanych przez zespół medyczny, z pomiarem wskaźnika trafności i częstości halucynacji przed każdym wdrożeniem aktualizacji modelu

5. Odpowiedzialność za rezultaty generowane przez system

- System pozycjonowany jest jako **narzędzie wspomagające komunikację**, nie jako źródło decyzji klinicznych. Odpowiedzialność za treść porady medycznej pozostaje wyłącznie przy lekarzu prowadzącym
- Wprowadzenie **mechanizmu human-in-the-loop** na poziomie operacyjnym: możliwość zgłoszenia błędnej/nieadekwatnej odpowiedzi przez użytkownika, przegląd zgłoszeń przez zespół medyczny CeZ
- Jasne określenie w regulaminie aplikacji zakresu odpowiedzialności dostawcy systemu (CeZ) oraz ograniczeń funkcjonalnych narzędzia, zgodnie z wymogami AI Act dotyczącymi transparentności wobec użytkownika systemu wysokiego ryzyka

