

Firma Business Online Services Sp. z o.o. realizuje projekt pn. „Generator odpowiedzi na pisma w oparciu o organizacyjną bazę wiedzy Generative Knowledge Base (GKB)”.

Projekt jest współfinansowany przez Unię Europejską ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach programu Fundusze Europejskie dla Mazowsza 2021-2027.

Łączna wartość projektu to: 6 553 020,00 PLN, a wartość przyznanego dofinansowania wynosi 4 993 312,00 PLN

W ramach realizowanego projektu przeprowadzone zostaną prace badawczo-rozwojowe obejmujące badania przemysłowe i eksperymentalne prace rozwojowe. Celem jest opracowanie innowacyjnego rozwiązania umożliwiającego utworzenie systemu generatywnego opartego na dużych modelach językowych (LLM) i innych technologiach z zakresu przetwarzania języka naturalnego (NLP), który będzie umożliwiał wsparcie użytkownika systemu obiegu dokumentów poprzez generowanie potencjalnych odpowiedzi na pytania użytkowników oraz na pisma przychodzące do danej organizacji na podstawie wiedzy organizacyjnej zgromadzonej w centralnej bazie wiedzy.

Na wiedzę organizacyjną składać się będą wszelkie dokumenty organizacyjne istotne z punktu widzenia działania organizacji. Będą to wewnętrzne i zewnętrzne akty normatywne (zarządzenia, instrukcje, procedury, regulaminy, akty prawne), zawarte umowy i zobowiązania, zatwierdzone dokumenty wzorcowe (wzory pism, ofert, umów etc.), jak i wcześniej wytworzone dokumenty robocze, reprezentujące już wytworzone wcześniej pisma zawierające przykładowe odpowiedzi. Projektowany system będzie zoptymalizowany pod kątem generowania odpowiedzi na przychodzące do organizacji pisma i wiadomości, a nie na pojedyncze zapytania stawiane przez użytkowników siedzących przy komputerze. Ze względu na brak dostępności gotowych rozwiązań dla zdefiniowanych w projekcie konkretnych problemów przeprowadzenie prac badawczo-rozwojowych będzie również skutkować wytworzeniem w projekcie nowej wiedzy w postaci algorytmów, które będą zoptymalizowane do rozwiązywania konkretnych postawionych problemów biznesowych z jakością umożliwiającą zastosowania produkcyjne.

Sytuacja rynkowa w dziedzinie rozwiązań opartych na AI

Rosnące zainteresowanie firm i instytucji wdrażaniem rozwiązań opartych na AI jest trendem o charakterze globalnym. Znacząco rośnie zapotrzebowanie rynku na rozwiązania wykorzystujące modele wielkojęzykowe (LLM) i wstępnie wytrenowane transformatory generatywne (GPT) w architekturze Retrieval-Augmented Generation (RAG). Architektury RAG zwiększają możliwości LLM poprzez integrację zewnętrznych źródeł wiedzy w procesie generowania, co prowadzi do wyższej dokładności i zmniejszenia halucynacji. To sprawia, że rozwiązania oparte na RAG są szczególnie atrakcyjne w przypadku aplikacji wymagających precyzyjnych i odpowiednich kontekstowo odpowiedzi (Databricks) (MS Learn). Kluczowe czynniki napędzające popyt obejmują zdolność modeli RAG do ulepszania generowania treści, zwiększania automatyzacji obsługi klienta i wspierania zaawansowanych zadań wyszukiwania informacji. Branże takie jak opieka zdrowotna, prawo, edukacja i obsługa klienta coraz częściej wdrażają te rozwiązania, aby wykorzystać możliwości. Na przykład w opiece zdrowotnej modele RAG pomagają w diagnostyce medycznej, zapewniając dostęp do najnowszych badań i wytycznych klinicznych, natomiast w dziedzinie prawa usprawniają proces przeglądania dokumentów i badań prawnych (MarketsandMarkets) (Analytics Vidhya). Przewiduje się, że rynek rozwiązań LLM i RAG będzie rósł w miarę jak coraz więcej organizacji będzie starało się wdrażać technologie sztucznej inteligencji, które mogą zarządzać ogromnymi ilościami danych i zapewniać praktyczne spostrzeżenia. Wzrost ten jest wspierany przez ciągły rozwój sztucznej inteligencji, ulepszone mechanizmy wyszukiwania i integrację danych multimodalnych, co jeszcze bardziej zwiększa skuteczność tych modeli (MS Learn) (Analytics Vidhya).

Innowacyjność projektu

W przypadku opracowania rozwiązania będącego przedmiotem projektu, zakłada się, że podejście do wykorzystania modeli LLM/GPT w architekturze RAG zostanie diametralnie zrewolucjonizowane, z uwagi na kluczowe cechy i funkcjonalności innowacji produktowej będącej przedmiotem projektu, dotychczas nie występujące na rynku.

Innowacja ta dotyczy w szczególności następujących obszarów:

- a) podział dokumentów na fragmenty semantycznie spójne i generacja zapytań,
- b) zarządzanie dokumentami, ochrona i retencja danych,
- c) wyszukiwanie wielowymiarowe i ranking dokumentów,
- d) wyszukiwanie agentowe i rozszerzanie zapytań,
- e) autoewaluacja jakości generowanej odpowiedzi,
- f) architektura “Multi-layer RAG” (autorska koncepcja nazwana roboczo “Multi-layer RAG”).

Założeniem Wnioskodawcy jest, że podstawowa funkcja systemu będzie działała w następujący sposób:

- Użytkownik przedstawia pismo/wiadomość, do której organizacja musi się ustosunkować i wygenerować odpowiedź;
- System na podstawie treści wiadomości formułuje zapytanie do bazy wiedzy w celu zlokalizowania odpowiednich dokumentów, które zawierają treści relewantne w stosunku do treści wiadomości lub zapytania;
- System w odpowiedni sposób wybiera fragmenty tekstów z bazy wiedzy, takie, które zawierają informacje, które powinny być uwzględnione w odpowiedzi na wiadomość przedstawioną na wejściu tworząc tzw. kontekst informacyjny odpowiedzi (np. odpowiednie paragrafy z umów, aktów normatywnych itp.);
- Na podstawie utworzonego kontekstu system generuje odpowiedź na wiadomość przedstawioną na wejściu. Dodatkowo system prezentuje użytkownikowi utworzony kontekst informacyjny wraz ze wskazaniem źródeł odpowiednich relewantnych fragmentów tekstów w celu umożliwienia weryfikacji sugestii systemowych przez użytkownika;

Dodatkową funkcją systemu jest możliwość eksploracji zgromadzonej bazy wiedzy samodzielnie przez użytkownika. Będzie się to odbywać na dwa sposoby:

Po pierwsze zadawanie przez użytkownika pytań w języku naturalnym. Sposób generowania odpowiedzi na pytanie zasadniczo będzie analogiczny do opisanego sposobu generowania odpowiedzi na wiadomość przychodzącą. Spodziewane są niewielkie różnice w sposobie działania opracowanych algorytmów w takiej sytuacji, ale ogólny schemat pozostanie podobny do opisanego wcześniej.

Drugim sposobem będzie eksploracja bazy wiedzy samodzielnie przez użytkownika na zasadzie “drill down” poprzez nawigowanie w strukturze bazy wiedzy pomiędzy tematami, słowami kluczowymi, “tagami” i powiązanych z nimi odpowiednimi fragmentami tekstu, z jednoczesnym wskazaniem źródłowego dokumentu i miejsca wystąpienia danego tekstu w dokumencie. Zarówno struktura tematów, tagów, jak i słów kluczowych będzie generowana automatycznie przez odpowiednie algorytmy przetwarzania języka naturalnego (NLP).

Zakres badań

1. Budowa zbioru danych obejmującego przykładowe bazy wiedzy oparte na repozytoriach danych organizacyjnych pozyskanych od kontrahentów, z którymi współpracuje spółka. Zakłada się zbudowanie przynajmniej 2 repozytoriów składających się średnio z 200 dokumentów organizacyjnych;
2. Budowa zbioru treningowego typu "ground truth", który będzie służył do podstawowej weryfikacji działania opracowanych autorskich algorytmów.
3. Ustalenie metryk weryfikacji jakości kluczowych algorytmów systemu.
4. Ustalenie poziomu bazowego metryk przy użyciu standardowej architektury RAG.
5. Opracowanie autorskich innowacyjnych algorytmów dla kluczowych elementów systemu i badanie wpływu tych algorytmów na wskaźniki jakościowe.
6. Iteracyjne rozwijanie przedstawionych algorytmów i cykliczne testowanie wpływu na przyjęte wskaźniki jakościowe.
7. Budowa i optymalizacja prototypu systemu pod kątem wdrożeń pilotażowych i produkcyjnych.

Współpraca z organizacją badawczą

Spółka w ramach realizacji projektu wraz z organizacją badawczą planuje:

1. Opracowanie algorytmu auto-ewaluacji dla agenta sterującego generowaniem odpowiedzi z użyciem kluczowych parametrów Context Relevance, Answer Faithfulness i Answer Relevance w oparciu o Natural Language Inference (ocenę semantycznych zależności pomiędzy elementami odpowiedzi, kontekstu i zapytania innymi niż tylko odległość wektorowa), w celu wykrywania i filtrowania halucynacji modelu.

Business Online Services Sp. z o.o. wspólnie z organizacją badawczą będzie uczestniczył w testowaniu oprogramowania wytworzonego podczas realizacji projektu.

Kluczowe cechy wyników prac B+R

Kluczową cechą opracowanego produktu będzie zastosowanie inteligentnego agenta opartego na algorytmie autoewaluacji, który maksymalizuje prawdopodobieństwo uzyskania wiarygodnej odpowiedzi i minimalizuje ryzyko halucynacji. Ponadto agent będzie zorientowany na rozwiązanie konkretnego, powtarzalnego problemu biznesowego, jakim jest wygenerowanie potencjalnej odpowiedzi na pismo z wykorzystaniem faktów zgromadzonych w bazie wiedzy, czyli będzie ukierunkowany na cel. Dzięki temu zostanie dostarczone gotowe rozwiązanie „pod klucz”, które będzie zdolne do szybkiego generowania oszczędności bez konieczności uruchamiania projektów programistycznych mających na celu dostosowanie generycznych rozwiązań typu RAG do potrzeb biznesowych organizacji.

Rozwiązanie Generative Knowledge Base (GKB) będzie instalowane w modelu “on premise”, czyli w infrastrukturze klienta. Możliwość instalowania rozwiązania w infrastrukturze klienta pozwoli na wykorzystanie danych organizacji również takich, które nie mogą być przekazane do infrastruktury chmurowej. Będzie to jedna z przewag oferowanego rozwiązania nad konkurencją szczególnie dla branż, w których ochrona danych ma szczególne znaczenie (np. służba zdrowia).

Ponadto zniknie problem zmienności i nieprzewidywalności opłat za wykorzystanie chmurowych rozwiązań typu LLM/GPT. Ponieważ rozwiązania te bazują na bardzo technicznych aspektach przy wycenie, takich jak ilość przetworzonych “tokenów”, bardzo trudne jest oszacowanie kosztów użytkowania tych rozwiązań. Model sprzedaży oparty będzie na jednorazowej opłacie za wdrożenie systemu oraz udzieleniu rocznej licencji.

Oferowane rozwiązanie będzie zoptymalizowane pod kątem rozwiązywania konkretnego problemu, który polega na wsparciu użytkownika w tworzeniu odpowiedzi na przychodzące do organizacji pisma i wiadomości. Zwrot z inwestycji może być w takim przypadku liczony proporcjonalnie do ilości przychodzących wiadomości do organizacji.

Oferta opierać się będzie na 3 kluczowych filarach:

1. Możliwość wdrażania systemu w modelu on-premise, co daje możliwość wykorzystania w systemie poufnych danych organizacji oraz pełną kontrolę nad kosztami wykorzystania.
2. Szybkie dostarczenie systemu gotowego “pod klucz” dzięki zastosowaniu architektury Multi-Layer RAG, która umożliwia szybkie zasilenie systemu wiedzą potrzebną do działania w danej organizacji.
3. Dostarczenie rozwiązania konkretnego problemu biznesowego (wsparcie w procesie generowania odpowiedzi na pisma), które pozwala generować policzalne oszczędności i wykazywać zwrot z inwestycji w system.

Podsumowaniem opisu innowacji produktowej są jej kluczowe cechy i funkcjonalności tj.:

- **Obsługa generowania odpowiedzi na pisma przychodzące a nie na pojedyncze zapytania użytkownika;**
- **Możliwość wdrażania on-premise i wykorzystania danych poufnych organizacji;**
- **Szybkie wdrożenie i zasilenie systemu wiedzą branżową i organizacyjną;**
- **Jasny model kosztowy umożliwiający łatwą kontrolę kosztów użytkowania;**
- **Realizacja “out of the box” całościowego procesu biznesowego polegającego na automatyzacji i wsparciu odpowiedzi na pisma przychodzące;**
- **Optymalizacja wewnętrznych algorytmów pod kątem wysokiej jakości realizowanych zadań przez wykorzystanie dedykowanych agentów zorientowanych na założony cel biznesowy.**