Fundusze Europejskie dla Mazowsza | Rzeczpospolita Polska | Dofinansowane przez Unię Europejską | Mazowsze. serce Polski

AUTOMADE

Firma AUTOMADE Sp. z o.o. realizuje projekt pn. „Prace B+R nad systemem opartym na LLM/GPT, umożliwiającym wykorzystanie języka naturalnego do generowania odpowiedzi na pytania wymagające dostępu do danych poufnych organizacji”.

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 6 562 771,75 PLN, a wartość przyznanego dofinansowania wynosi 4 999 999,99 PLN

W ramach realizowanego projektu przeprowadzone zostaną prace badawczo-rozwojowe obejmujące badania przemysłowe i eksperymentalne prace rozwojowe, których celem jest opracowanie architektury, metodyki wdrożeniowej i wykonanie prototypu systemu, który umożliwi wdrażanie rozwiązań opartych na Large Language Models (LLM) / Generative Pretrained Transformers (GPT) w modelu On Premise z możliwością wykorzystania w działaniu systemów i danych dostępnych tylko wewnątrz organizacji, bez upubliczniania poufnych danych i systemów wewnętrznych na zewnątrz (do chmury).

Opracowany zostanie sposób współpracy czatbota z istniejącymi systemami w organizacji klienta, również tymi nieposiadającymi API, tak, aby czatbot potrafił samodzielnie dobrać z wcześniej określonej puli takie funkcje systemów, które najlepiej realizują cel użytkownika. Jednocześnie opracowany zostanie sposób podsumowywania danych pozyskanych z zewnętrznych systemów tak, aby w możliwie zwięzły sposób przekazać potrzebne informacje z powrotem do użytkownika. Wyeliminowany zostanie problem halucynacji w odpowiedziach modeli AI zw. z brakiem wiedzy ogólnych modeli chmurowych na temat specyficznych danych organizacji. W ramach projektu wykorzystane zostaną m.in. właściwości modeli LLM związanych z wnioskowaniem z tekstu oraz zastosowany zostanie prompt engeneering ukierunkowany na ekstrakcję parametrów konektorów z przebiegu konwersacji czata, integrację modułu RPA z czatem oraz prompt engineering ukierunkowany na wybór danych z kontekstu i utworzenie podsumowania (sformułowanie precyzyjnej odpowiedzi dla użytkownika).

Sytuacja rynkowa w dziedzinie modeli LLM/GPT

Pojawienie się tzw. sztucznej inteligencji (AI) opartej na dużych modelach językowych (LLM) a w szczególności modele oparte na koncepcji GPT, wywołało w ostatnim czasie ogromne zainteresowanie. Duża uniwersalność tych modeli (np. ChatGPT firmy OpenAI) wynika z tego, że agregują one ogromną ilość informacji tekstowej pozyskanej z ogólnodostępnych źródeł (strony internetowe, wikipedia etc.). Dziedzina LLM/GPT rozwija się w ostatnim czasie bardzo dynamicznie a powiązane z nią technologie znajdują coraz więcej zastosowań.

Jednocześnie występuje szereg istotnych problemów i ograniczeń w zastosowaniu ww. technologii do zastosowań w konkretnych niszach biznesowych i organizacjach. Zastosowanie rozwiązań chmurowych opartych o LLM wewnątrz organizacji napotyka dwie główne przeszkody (niezależne od branży i specyfiki organizacji):

1. Brak dostępu samego modelu do danych biznesowych organizacji,

2. Zagrożenie niekontrolowanym wypływem poufnych danych.

Po pierwsze istniejące i publicznie dostępne modele, agregują informacje z publicznie dostępnych źródeł i są w stanie odpowiadać na ogólne pytania, ale nie mają żadnych informacji o konkretnych faktach, które dotyczą danej organizacji a nie są upublicznione, gdyż stanowią tajemnice przedsiębiorstwa i/lub są obwarowane konkretnymi ograniczeniami, co do sposobu dystrybucji tych informacji i ich wykorzystania, np. wynikających z RODO. W związku z tym systemy te nie potrafią wiarygodnie odpowiedzieć na pytania dot. takich rodzajów informacji (np. informacje o klientach, sprzedaży, wewnętrznym know-how). Dodatkowy problem związanym z udostępnianiem wewnętrznych danych organizacji wynika z tego, że dane te są zazwyczaj przechowywane w różnych miejscach (bazach danych), w wewnętrznych systemach IT organizacji. Ich zastosowanie w modelach LLM wymagałoby ręcznego wprowadzania tych danych w postaci promptów przez użytkownika, co niweczy zyski z automatyzacji. Druga kwestia to bezpieczeństwo danych. Ponieważ modele LLM de facto nie są świadome o co są pytane i zawsze próbują dać jakąś odpowiedź generując tekst, to mają skłonność do tzw. halucynacji czyli generowania odpowiedzi nieprawdziwych, ale brzmiących wiarygodnie. Wykorzystywanie takich odpowiedzi przez nieświadomych użytkowników może generować duże ryzyko biznesowe. Ponadto nieświadomy użytkownik może przekazać w treści zapytania lub podpowiedzi dla czata (tzw. prompt) wewnętrzne dane organizacji, które nie powinny być przekazywane do rozwiązań chmurowych.

Realizowany projekt stanowi odpowiedź na ww. wyzwania.

Innowacyjność projektu

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

- możliwość wykorzystywania dużych modeli językowych do wnioskowania i formułowania odpowiedzi tekstowych w modelu on premise, bez udostępniania danych do chmury internetowej, weryfikowane przez możliwość generowania odpowiedzi przez model bez kontaktu z zewnętrzną siecią organizacji;

- możliwość generowania przez model odpowiedzi bazujących na wewnętrznych danych, które są przechowywane w wewnętrznych systemach organizacji, które nie posiadają API weryfikowane przez możliwość podłączenia nowego wywołania systemu bez konieczności prowadzenia prac programistycznych;

- umożliwienie automatyzacji w procesach wykorzystujących dane poufne;

- umożliwienie wdrożenia automatyzacji dla systemów nie posiadających API.

Zakres badań

Cel badawczy projektu zostanie osiągnięty poprzez:

1. Opracowanie architektury systemu pozwalającej na łączenie ze sobą rozwiązań RPA i API jako konektorów do systemów biznesowych i LLM jako silnika wnioskującego i generującego odpowiedzi dla użytkownika;

2. Ewaluację i wybór rozwiązania LLM właściwego do wdrażania on premise i opracowanie autorskiego rozwiązania umożliwiającego dostosowanie wybranego modelu LLM do realizacji kluczowego wyzwania projektowego, jakim jest realizacja celu użytkownika przy użyciu danych pobieranych z zewnętrznych systemów przez konektory - wykorzystanie właściwości modeli LLM zw. z wnioskowaniem z tekstu, zastosowanie tzw. prompt engeneering-u oraz “wstrzykiwanie” do kontekstu konwersacji chata danych transakcyjnych pozyskiwanych na bieżąco z systemów IT;

3. Opracowanie autorskiego algorytmu realizującego wybór właściwych konektorów z puli wcześniej zdefiniowanych konektorów wychodząc od celu (zapytania) stawianego przez użytkownika - opracowanie systemu promptów oraz dwuetapowego algorytmu (1. scoring konektorów pod kątem ich użyteczności 2. wstrzykiwanie kandydatów do dalszego wnioskowania), użycie wektorowej bazy danych informacji o konektorach przy wykorzystaniu tzw. embedding-u wektorów tekstowych i analizie podobieństwa wektorów;

4. Opracowanie autorskiego algorytmu wywołującego konektory z właściwymi parametrami i pobierającego dane do kontekstu konwersacji czata - prompt engineering ukierunkowany na ekstrakcję parametrów konektorów z przebiegu konwersacji czata, integracja modułu RPA z czatem, opracowanie metod realizujących koncepcję tzw. long term/mid term conversation memory, rozwiązania bazujące na sumaryzacji informacji w połączeniu z embedding-iem i przesuwnym oknem czasowym;

5. Opracowanie autorskiego algorytmu podsumowującego informacje pobrane do kontekstu konwersacji w celu sformułowania precyzyjnej odpowiedzi dla użytkownika - prompt engineering ukierunkowany na wybór danych z kontekstu i utworzenie podsumowania.

Współpraca z organizacją badawczą

Wnioskodawca w ramach realizacji projektu współpracować będzie z organizacją badawczą. Zakres usługi badawczej obejmuje:

1. Ewaluacja i wybór modelu językowego typu LLM spełniającego kryteria projektu w zakresie licencyjnym (wdrożenia on premise), wydajnościowym (szybkość odpowiedzi), zapotrzebowania na zasoby i jakościowym przy założeniu wykorzystania języka polskiego. Wymaga opracowania odpowiednich testów, przeprowadzenia testów i dostarczenie rekomendacji dla wybranego modelu.

W ramach usługi badawczej Wnioskodawca otrzyma: Wyniki testów i rekomendowany model.

2. Opracowanie sposobu reprezentacji wiedzy o możliwościach funkcjonalnych agentów (konektorów) wraz z opracowaniem mechanizmu przechowywania tej wiedzy w pamięci trwałej (bazie danych) w sposób umożliwiający dobór odpowiednich agentów do zapytania użytkownika z wykorzystaniem modelu LLM i architektury RAG (retrieval augmented generation).

W ramach usługi badawczej AUTOMADE Sp. z o.o. otrzyma wyniki testów pokazują prawidłowy wybór agenta dla danego zapytania dla przykładowych definicji agentów i różnych zapytań (na tym etapie agent nie musi działać i wykonywać zadań chodzi tylko o wybór i opis zastosowanych rozwiązań i algorytmów).

Kluczowe cechy wyników prac B+R

Projektowane rozwiązanie będzie konkurencyjne dla rozwiązań istniejących przede wszystkim pod względem użyteczności, skuteczności i sposobu obsługi. Przewaga pod kątem użyteczności będzie wynikała przede wszystkim z możliwości wdrażania rozwiązania w modelu on premise. Organizacje, dla których problemem jest używanie rozwiązań chmurowych ze względu na ochronę danych czy brak możliwości wystawienia chmurowego API do istniejących systemów nie mają w tym momencie możliwości korzystania z rozwoju technologii związanych z AI, w tym LLM/GPT. Można zatem powiedzieć, że istniejąca oferta chmurowa jest dla tych organizacji bezużyteczna. Z kolei wykorzystywanie generycznych modeli LLM/GPT bez podłączenia do wiedzy zgromadzonej w istniejących systemach IT zmniejsza praktyczną skuteczność użycia takich rozwiązań do zera. Model AI, który będzie próbował odpowiadać na pytania dotyczące danych, które są w posiadaniu organizacji, ale bez dostępu do tych danych będzie generował wyłącznie bezużyteczne halucynacje. Przewaga pod kątem użyteczności nad rozwiązaniami dostępnymi na rynku będzie wynikać z zaimplementowania rozbudowanych możliwości wnioskowania, podsumowywania i generowania tekstu, charakterystycznych dla najnowszych technologii LLM/GPT. Dzięki temu użycie automatyzacji będzie prostsze - będzie sprowadzało się do nagrania zestawu makr RPA, które będą tak dobrane, aby potencjalnie pokrywały określoną przez klienta listę celów biznesowych, które mają zaspokajać. Nie będzie konieczne ani modelowanie drzewa konwersacji czy innej formy przebiegu dialogu z użytkownikiem a priori - moduł LLM samodzielnie będzie wnioskował jakich makr i w jaki sposób użyć. W związku z tym nie będzie konieczne budowanie API do istniejących systemów, co znacząco skróci czas i koszty wdrożenia. Proponowany system zwiększy szybkość i efektywność realizacji procesów biznesowych, związanych z wyszukiwaniem i podsumowaniem informacji zgromadzonych w systemach IT organizacji (procesy obsługi klienta, weryfikacji klientów i partnerów biznesowych, procesy zarządcze) przy jednoczesnym zmniejszeniu kosztów wynagrodzeń i kosztów obsługi zatrudnienia (świadczenia, onboarding).

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

- możliwość wykorzystywania dużych modeli językowych do wnioskowania i formułowania odpowiedzi tekstowych w modelu on premise, bez udostępniania danych do chmury internetowej, weryfikowane przez możliwość generowania odpowiedzi przez model bez kontaktu z zewnętrzną siecią organizacji;
- możliwość generowania przez model odpowiedzi bazujących na wewnętrznych danych, które są przechowywane w wewnętrznych systemach organizacji, które nie posiadają API weryfikowane przez możliwość podłączenia nowego wywołania systemu bez konieczności prowadzenia prac programistycznych;
- umożliwienie automatyzacji w procesach wykorzystujących dane poufne;
- umożliwienie wdrożenia automatyzacji dla systemów nie posiadających API.