dr hab. inż. Maciej Piasecki, Politechnika Wrocławska

PLLuM – polski duży model językowy wdrożony poprzez infrastrukturę badawczą CLARIN-PL

PLLuM (https://pllum.org.pl/), czyli Polish Large Language Model, to rodzina dużych generatywnych modeli językowych (ang. LLMs - Large Language Models) zbudowanych od podstaw przez polskiego konsorcjum naukowe z myślą o języku polskim i wspieraniu różnorodnych zastosowań jako część otwartej infrastruktury technologii językowych. Budowa pierwszej wersji została zrealizowana w roku 2024 jako inwestycja publiczna finansowana przez Ministerstwo Cyfryzacji przez konsorcjum sześciu jednostek naukowych, którego liderem była Politechnika Wrocławska. Dalsze prace nad rozbudową i wdrożeniem były kontynuowane w roku 2025 przez poszerzone konsorcjum HIVE, którego liderem był NASK. Idea PLLuM jako otwartej, bezpłatnej infrastruktury wspierającej przetwarzanie języka polskiego wywodzi się od wieloletniej działalności infrastruktury badawczej CLARIN-PL (https://clarin-pl.eu/).

Celem projektu PLLuM było opracowanie modelu językowego, który dobrze reprezentuje własności języka polskiego  oraz kulturowo uwarunkowane style komunikacji, jest efektywny w wielu zastosowaniach istotnych dla nauki i społeczeństwa, jak również transparentny pod względem prawnym.

W ramach wystąpienia przedstawione zostaną podstawowe cele i założenia projektu PLLuM. Przeanalizujemy, dlaczego warto budować własne duże generatywne modele językowe (GMJ, ang. LLMs), np. ze względu na ich transparentność, polską suwerenność technologiczną i zdobycie wiedzy pogłębionej praktycznie.

Przedstawione zostanie proces budowy PLLuM: analiza prawna, zebranie danych językowych, budowa zasobów językowych, trening wstępny (ang. pretraining), dostrajanie (ang. finetuning) i wychowanie (ang. alignment). Wybranym aspektom przyjrzymy się bliżej, z perspektywy zdobytych doświadczeń, szczególną uwagę poświęcając budowie zasobów językowych, dostrajaniu i wychowaniu do zadań oraz ważnemu zagadnieniu wiarygodnej ewaluacji GMJ (ang. LLM). W projekcie PLLuM, od samego początku, bardzo ważnym zagadnieniem było pilotażowe wdrożenie w postaci asystenta wiedzowego dla instytucji publicznych, czyli systemu opartego na schemacie RAG (ang. Retrieval Augmented Generation), gdzie system wyszukiwania semantycznego jest połączony z GMJ (ang. LLM) - w naszym przypadku specjalną wersją PLLuM - do generowania odpowiedzi. W trakcie wystąpienia przyjrzymy się bliżej wybranym aspektom konstrukcji systemów RAG i ich dalszemu rozwinięciu w ramach architektury agentowej.

Jednym z impulsów do budowy PLLuM była potrzeba infrastruktury CLARIN-PL do pozyskania polskiego GMJ (ang. LLM) jako rodzaju silnika do budowy narzędzi badawczych dla naukowców. W ramach wystąpienia omówimy również trwające prace nad tego typu wdrożeniami PLuM, w szczególności w ramach systemu podstawowych narzędzi językowych, AutoRAG - silnika asystentów wiedzowych, asystenta badacza - systemów agentowych wspierających autonomiczność działań oraz systemu głębokiej eksploracji danych językowych.