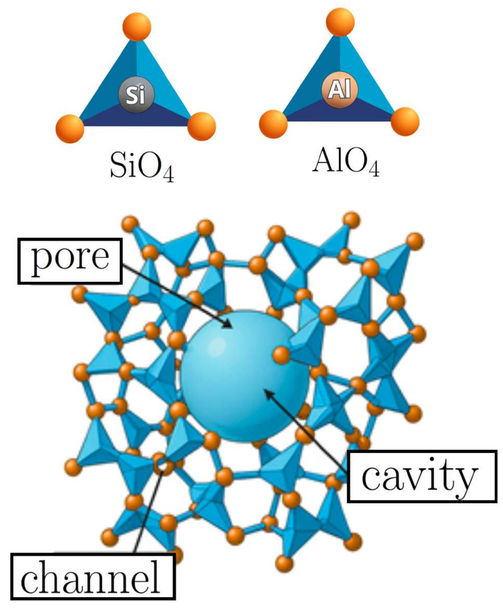
INSTYTUT INŻYNIERII CHEMICZNEJ Polskiej Akademii Nauk

IITiS INSTYTUT INFORMATYKI TEORETYCZNEJ I STOSOWANEJ POLSKIEJ AKADEMII NAUK

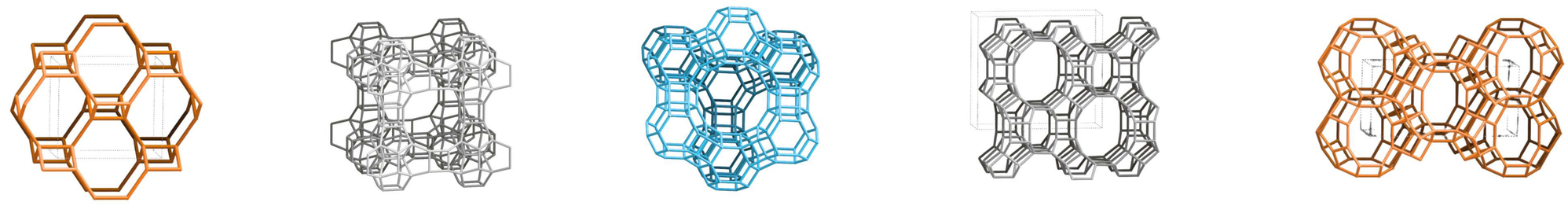
Tworzenie bazy danych zeolitów z użyciem LLM

mgr Marzena Halama[ϕ], dr hab. inż. Joanna Domańska[ϕ], dr hab. inż. Anna Gancarczyk[ψ], dr inż. Mirzokhid Abdirakhimov[ψ]
[ϕ]Instytut Informatyki Teoretycznej i Stosowanej Polskiej Akademii Nauk, [ψ] Instytut Inżynierii Chemicznej Polskiej Akademii Nauk.

Zeolity są mikroporowatymi materiałami krystalicznymi zbudowanymi z szkieletów glinokrzemianowych. Są one szeroko stosowane m.in. jako sita, wymieniacze jonowe i katalizatory w szerokim zakresie zastosowań przemysłowych i środowiskowych. Zeolity syntetyczne są preferowane w stosunku do naturalnych, ponieważ oferują większą kontrolę ich właściwości, czystość i wydajność. Mogą być produkowane różnymi metodami, w tym hydrotermalną lub metodą fuzji. Przy wyborze zeolitu do danego procesu uwzględnia się rozmiar porów, stosunek Si/Al, rodzaj kationów (Na^+, K^+, Ca^{2+}), a także stabilność termiczną i chemiczną. Właściwości zeolitu zależą od źródła krzemionki (krzemian sodu, żel lub zol krzemionkowy) i glinu (glinian sodu, boehmit (pseudo-boehmit), wodorotlenek glinu, izopropanolan glinu), a także od parametrów syntezy (temperatury i czasu krystalizacji, suszenia i starzenia). Ze względu na dużą różnorodność struktur zeolitowych i parametrów ich syntezy, a także ogromną liczbę potencjalnych zastosowań, konieczne staje się uporządkowane gromadzenie i analiza informacji o zeolitach. Poniżej przedstawione zostały przykładowe struktury szkieletowe zeolitów:



ABW BEC FAU MOR RTH



Dlatego pierwszym etapem prac było zaprojektowanie zautomatyzowanego systemu pozyskiwania i strukturyzacji danych z wykorzystaniem narzędzi sztucznej inteligencji, w szczególności dużych modeli językowych (LLM). Rozwiązanie to stanowi fundament pod przyszłe opracowanie *systemu rekomendacyjnego*, którego zadaniem będzie wspieranie doboru odpowiednich zeolitów do określonych zastosowań. Zaprojektowany system wspierał proces kompleksowego zarządzania danymi dotyczącymi zeolitów. W szczególności umożliwiał:

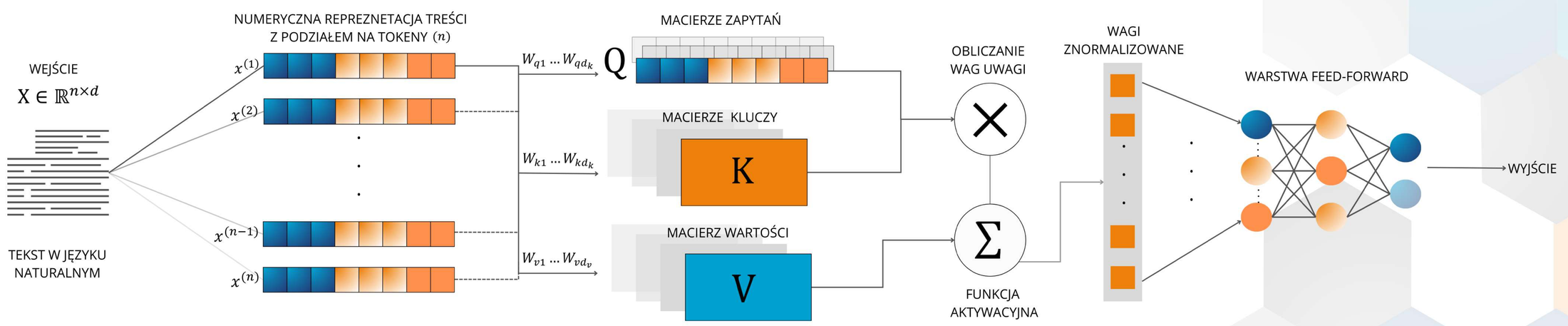
- **Identyfikację i ekstrakcję informacji materiałowych z dokumentów o charakterze tekstowym**
- **Ujednolicenie parametrów eksperymentalnych (np. jednostek, nazw związków, typów procedur)**
- **Konwersję danych do postaci strukturalnej, zgodnej z założeniami relacyjnej bazy danych**
- **Automatyczne budowanie i aktualizowanie bazy danych zawierającej charakterystyki zeolitów, parametry syntezy, właściwości i zastosowania**

LLM posiadają zdolność semantycznej interpretacji języka naturalnego, co umożliwia im automatyczne rozpoznawanie i ekstrakcję informacji naukowej z tekstów źródłowych, takich jak publikacje, raporty badawcze czy dokumentacja eksperymentalna. Dla danej sekwencji wejściowej o n tokenach, zakodowanej jako macierz osadzeń $\mathbf{X} \in \mathbb{R}^{n\times d}$, gdzie d to wymiar przestrzeni embeddingowej, definiujemy macierze zapytań, kluczy i wartości jako:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_v,$$

gdzie $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d\times d_k}$ oraz $\mathbf{W}_v \in \mathbb{R}^{d\times d_v}$, to macierze wag (parametry modelu), a d_k to wymiar przestrzeni zapytań i kluczy. Następnie mechanizm uwagi obliczany jest z wykorzystaniem poniższego wzoru:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}.$$



W procesie budowy bazy danych wykorzystano bogate repozytorium International Zeolite Association – Structure Commission. Dane zostały uporządkowane w 59 podfolderach odpowiadających trzyliterowym kodom topologii zeolitów (np. FAU). Każdy folder zawierał jeden lub więcej plików PDF. Następnie model **llama3.3** uruchomiony w Ollama, został użyty do ekstrakcji jedenastu kluczowych cech syntezy: nazwy zeolitu, wzoru empirycznego, źródeł krzemu, glinu i kationów, czynnika rozpuszczającego, parametrów starzenia (temperatura i czas), temperatury i czasu krystalizacji oraz temperatury suszenia. Dane były zapisywane w formacie JSON, odzwierciedlając strukturę folderów źródłowych, a następnie scalane do zbiorczego pliku *csv*. W efekcie powstała baza, w której zidentyfikowano 62 unikalne źródła krzemionki i 51 źródeł glinu. Najczęściej stosowane były roztwory zolu krzemionkowego, tetraetoksysilanu oraz alkoholanów glinu i glinianu sodu.