



# **Sztuczna Inteligencja w badaniach historycznych**

zagrożenie, wsparcie czy krótkotrwała moda?

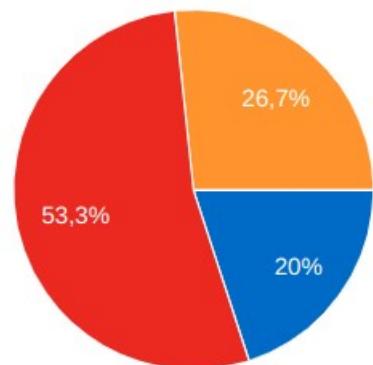
# Plan

- Omówienie ankiety
- Wprowadzenie do AI
- ChatGPT i inne
- Przetwarzanie i tworzenie tekstów naukowych
- Diagramy, wykresy, infografiki, analiza danych
- Rozpoznawanie pisma
- Masowe przetwarzanie tekstów – pozyskiwanie wiedzy
- Audio AI

# Ankieta

Jak oceniasz swoje doświadczenie z AI?

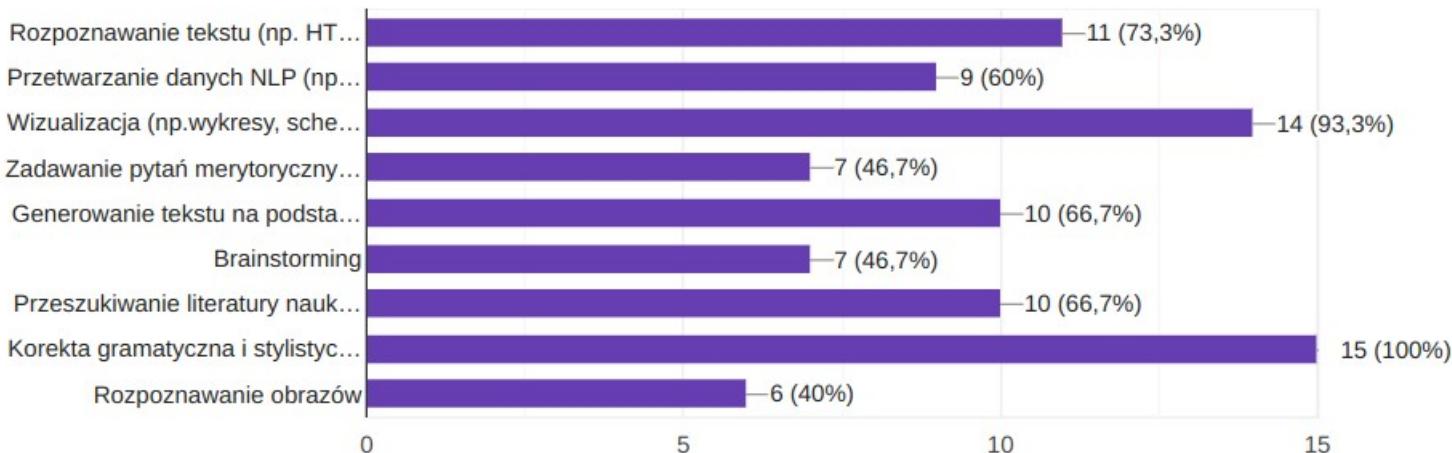
15 odpowiedzi



- Żadne
- Bardzo niewielkie (kilka prób, nic szczególnego)
- Niewielkie, ale systematyczne (regularne działania, ale rzadkie)
- Często (regularne, częste działania)
- Jestem ekspertem i nie potrzebuję szkolenia ;)

Jakie są Twoje oczekiwania względem AI?

15 odpowiedzi



# Ankieta

Główne oczekiwania uczestników wobec AI:

- analiza publikacji / przygotowywanie własnych publikacji (w tym wyszukiwanie literatury, informacji, korekta gramatyczna i stylistyczna, streszczenia, tłumaczenia)
- wizualizacja danych
- odczytywanie pisma
- przetwarzanie danych tekstowych, NLP

ale ponieważ blisko  $\frac{3}{4}$  osób deklaruje brak lub niewielkie doświadczenie z AI... zaczniemy od małego wprowadzenia.

# **1. WPROWADZENIE**

- **AI** - termin powstał w 1956 roku, opisuje systemy zdolne do interpretowania danych zewnętrznych, nauki i wykorzystywania wiedzy do wykonywania zadań
- **GenAI** - Generative AI – generatywna sztuczna inteligencja (**LLM**-y – duże modele językowe)
- **AGI** - Artificial General Intelligence – Ogólna (Prawdziwa) sztuczna inteligencja (**której jeszcze nie ma**)

# Czat a model językowy

- Czaty np.. ChatGPT czy Google Gemini nie są modelami językowymi, to interfejsy do rozmowy z modelami
- czat ‘pamięta’ rozmowę, model nie, kowersacja nie zostawa śladu w parametrach modelu, ale podczas kolejnego pytania w czacie do modelu podawana jest historia z poprzednich pytań
- model nie uczy się w trakcie rozmowy, uczenie wymaga procesu trenowania lub przynajmniej dotrenowywania (fine-tuning)
- czym więc są same modele językowe?

# LLM - duży model językowy

- to ogromne sieci neuronowe z miliardami parametrów
- określenie językowe może być mylące, obecne modele posiadają już często umiejętności multimedialne, czyli potrafią odczytywać tekst, obrazy, filmy, dźwięk
- potrafią też generować obrazy, tekst, filmy
- coraz lepiej programują...

# LLM - jak powstaje model

- trenowanie modeli bazowych (pre-trained)
  - ogromne ilości tekstu (np. do trenowania małego modelu Bielik2 zgromadzono 1.5 TB tekstu w języku polskim)
  - Bielik nie był trenowany od zera, oparto się na modelu Mistral
  - sprzęt, prąd, koszt - w przypadku modelu Bielik2 to prawdopodobnie 100-150 tys. dolarów, superkomputer Helios i 256 kart graficznych GH200 + czas (tygodnie)
- trening instrukcyjny (instruct) – sekwencje poleceń i przykładowych odpowiedzi (dla modelu Bielik stworzono 5 mln instrukcji, wspomagając się także instrukcjami generowanymi przez inny model - Mistral). Dopiero wówczas powstaje model, który potrafi wykonywać polecenia

# LLM - duży model językowy

- fine-tuning (dotrenowywanie) – ten proces nie jest już tak czasochłonny, wymaga mniej danych i umożliwia lepsze dopasowanie modelu do konkretnych zadań, jest to jednak trudne, niewłaściwie przeprowadzony fine-tuning może pogorszyć działanie modelu
- wykorzystanie (inferencja) - proces generowania odpowiedzi przez gotowy model również jest procesem kosztownym obliczeniowo, stąd dość wysokie ceny dostępu do modeli komercyjnych przez API

# LLM - tokeny

- **token** - fragment tekstu, który model przetwarza w trakcie analizy i generowania odpowiedzi
- modele ‘myślą’ tokenami
- tokenizacja zależy od języka, dla języka polskiego liczba tokenów dla tekstu o identycznej długości jak tekst angielski, będzie większa, ma to swoje **konsekwencje**
- wynika częściowo z charakterystyki języka, ale także z optymalizacji procedury tokenizacji pod kątem najpopularniejszego języka – angielskiego.

# LLM - tokeny

Przytoczone epizody z nadmorskiej podróży ostatniego Jagiellona zawierają w sobie jakby załączek i zapowiedź przyszłych jego poczynań, dotyczących Bałtyku. W kilka lat później wystąpił on otwarcie przeciw Inflantom i jednocześnie w grudniu 1556 powziął po raz pierwszy postanowienie przysposobienia wojennej floty, wprowadzane następnie ...

[Clear](#) [Show example](#)

Tokens	Characters
239	702

## Język polski

The above-mentioned episodes from the sea voyage of the last Jagiellon contain the seeds and foreshadowing of his future actions on the Baltic. A few years later he spoke openly against Inflants, and at the same time, in December 1556, he decided for the first time to prepare a navy for war, which he then implemented from 1561 in the form of a volunteer force. At the end of 1569, he took the first steps towards procuring his own warships: he instructed Antonino Angeli, a citizen of Venice, his agent and postmaster, to search for a master shipbuilder there and to

Przytoczone epizody z nadmorskiej podróży ostatniego Jagiellona w sobie jakby załączek i zapowiedź przyszłych jego poczynań, Bałtyku. W kilka lat później wystąpił on otwarcie przeciw Inflantom i jednocześnie w grudniu 1556 powziął po raz pierwszy postanowienie przysposobienia wojennej floty, wprowadzane następnie w życie od 1561 roku, w formie ochotniczego zaciągu, a na schyłku 1569 roku poczynił kolejne kroki celem zaopatrzenia się we własne okręty wojenne: oto polecił Angeli, obywatelowi Wenecji, swemu agentowi i sprawcy poczyniać tam i do Polski przyprowadzić mistrza do budowy okrętów, do którego Loredano prosił o pozwolenie wybranemu na wyjazd.

[Clear](#) [Show example](#)

Tokens	Characters
166	702

## Język angielski

The above-mentioned episodes from the sea voyage of the last Jagiellon contain the seeds and foreshadowing of his future actions on the Baltic. A few years later he spoke openly against Inflants, and at the same time, in December 1556, he decided for the first time to prepare a navy for war, which he then implemented from 1561 in the form of a volunteer force. At the end of 1569, he took the first steps towards procuring his own warships: he instructed Antonino Angeli, a citizen of Venice, his agent and postmaster, to search for a master shipbuilder there and to bring him to Poland, and asked the Doge, Peter Loredano, for permission for the chosen one to leave, for permission for the chosen one.

# LLM - tokeny konsekwencje

- skoro tekst tej samej długości w języku polskim przekłada się na więcej tokenów to:
- przetwarzanie polskich tekstów jest droższe
- wielkość kontekstu, który model może przetworzyć jednorazowo w języku polskim jest mniejsza.

# LLM - wielojęzyczność

- wiele współczesnych modeli jest wielojęzycznych, modele GPT obsługują zwykle około 100 najpopularniejszych języków, choć angielski najlepiej
- do trenowania modelu GPT-3 używano głównie tekstów w j. angielskim (93%), mimo to znał inne języki a twórcy nie rozumieli dlaczego
- słowo „najpopularniejsze” – odnosi się do dostępności tekstów w danym języku, nie liczby osób które się nim posługują, np. języki subsaharyjskie były początkowo obsługiwane bardzo słabo

# LLM - wielojęzyczność

- Modele znają łacinę (lepiej lub gorzej)

Universis et singulis venerabilibus saecularibus et cuius vis Ordinis Regularibus Praesbiteris, quovis nomine, dignitate et autoritate fungentibus et functuris inhibitetur, praecipitur et mandatur; ne, sub nullitatis absolutionis, ac excommunicationis, et a divinis suspensionis, ipso facto in currendae paenitentiis audeant, seu praesumant contra? Concilium Tridentum? et constitutione synodali?: ac iubilei praescriptum quovis praetextu et exemptione, sub ac post tempus praesentis iubilei quo suis utriusque sexus homines a quibusuis casibus communibus, aut ordinariae seu applicatae sedi reservatis absolvere, nec verbum Dei praedicare, nisi prius ab officialibus foraneis et decanis ruralibus (si ex civitatem Cracoviensem si vero in curiae Cracoviense degerint a synodalibus et aliis eis adiunctis examinatoribus) legitime approbari fuerint, et als prout in his aliis ex cancellaria dicti serenissimum emanatum et ut moris est Cracoviensem publicatis fusius continentur. In cuius rei fidem rc? Dattum Cracoviae ex cancellaria dicti serenissimum principis Cracoviensi episcopi cardinalis ducis severiensi Die ultima Junii 1634. (przykład dokumentu przysłał mi A. Karpacz)

---

Wszystkim i każdemu z osobna czcigodnym kapłanom diecezjalnym i zakonnym jakiegokolwiek zakonu, pełniącym i mającym pełnić jakąkolwiek funkcję, godność i władzę, zabrania się, nakazuje i poleca, aby pod karą nieważności rozgrzeszenia oraz ekskomuniki i suspensy a divinis, zaciąganej tym samym faktem, nie ważyli się ani nie przypuszczali wbrew Soborowi Trydenckiemu i konstytucjom synodalnym oraz przepisom jubileuszowym pod jakimkolwiek pretekstem i zwolnieniem, w czasie i po czasie obecnego jubileuszu rozgrzeszać ludzi obojga płci z jakichkolwiek przypadków wspólnych lub zarezerwowanych dla Stolicy Apostolskiej czy zwyczajnej, ani głosić słowa Bożego, jeśli wcześniej nie zostali prawnie zatwierdzeni przez urzędników zewnętrznych i dziekanów wiejskich (jeśli przebywają poza miastem Krakowem, jeśli zaś przebywają w kurii krakowskiej, przez egzaminatorów synodalnych i innych do nich przydzielonych), i inaczej, jak to szerzej zawarto w tych i innych dokumentach wydanych przez kancelarię wspomnianego najjaśniejszego i opublikowanych w Krakowie, jak to jest w zwyczaju. Na dowód czego itd. Dan w Krakowie, z kancelarii wspomnianego najjaśniejszego księcia, biskupa krakowskiego, kardynała księcia siewierskiego, dnia ostatniego czerwca 1634 roku. (przetłumaczył model: Claude 3 – Opus, marzec 2024)

# LLM - wielojęzyczność

- Modele znają staroniemiecki?

Kazimirus von gots gnaden koning czu Polen etc. grosfurste czu Littawen und erbling czu Reussen.

Grosmechtiger und erwirdiger unsir allirliebster frund. Es kompt uns in eyne sunderliche vordrislichkeit und smertcze euwir und euwirs ordens von eyme, der stete und euwirs Prewsschen volkes widerwille vom andern teyle, durch listikeit des alden fyendes erwecket, der do nicht hat geleden den stad euwir herschaft czu seeyn fredelich und bluende, sunder deme hat gereichert innwendige czutrenunge, als denen disse ding, durch die euwirn und der berurten ritterschaft und burger bothen uns vorkundet, wir haben empfangen.

---

Kazimierz, z Bożej łaski król Polski itd., wielki książę Litwy oraz dziedzic Rusi.

Wielce możliwy i czcigodny nasz umiłowany przyjacielu. Otrzymaliśmy od was, oraz waszego zakonu, informacje, które napełniają nas szczególną troską i bólem o was oraz o lud Prus, który, wbrew woli miasta, został podburzony przez przebiegłość starego wroga. Ów wróg nie mógł znieść waszej pokojowej i kwitnącej władzy, lecz sprowadził wewnętrzne nieporozumienia, jak nam to zostało zakomunikowane przez posłów waszego i wspomnianego rycerstwa oraz mieszkańców.

# LLM - wielkość

- wszystkie LLM-y są dużymi modelami, ale znacznie różnią się wielkością
- polski model Bielik 2 ma 11 miliardów parametrów (11b)
- chyba największy model open source Llama3-405 ma 405 miliardów parametrów, wielkość modeli komercyjnych jak GPT4o czy Claude nie jest dokładnie znana, ale są naprawdę duże
- modele GPT, Gemini pracują w dużych centrach komputerowych z tysiącami kart graficznych, (ciekawostka: w 2022 r. AI i kryptowaluty pochłaniały 2% światowego zużycia prądu, w 2026 ma to być 4% - tyle, ile zużywa cała Japonia)
- małe modele można uruchomić na zwykłym laptopie nawet bez GPU → mały pokaz Bielika + porównanie z Bielikiem udostępnionym w sieci

# LLM - ograniczenia (tekst)

- modele mają ograniczenia wielkości tekstu, który są w stanie przyjąć (context window) i tekstu, który są w stanie jednorazowo stworzyć (output token)
- dla modeli open-source zwykle jest to od 8 tys. do 128 tys. tokenów na wejściu, modele komercyjne przetwarzają nawet do 1.5 - 2 mln tokenów (ok. 750 tys. słów po angielsku, poniżej 0.5 mln po polsku)
- długość tekstów produkowanych przez modele jest bardziej limitowana np. GPT-4o może jednorazowo stworzyć 16 tys. tokenów

# LLM - ograniczenia (tekst)

- książka Jerzego Topolskiego „Wprowadzenie do historii” to około 96 tys. tokenów
- duży model open-source albo GPT-4o może tekst tej wielkości przetwarzać jednorazowo – mieści się z zapasem w oknie kontekstu
- model Gemini 1.5 Pro ma okno kontekstu wielkości 2 mln tokenów, więc zmieści jednorazowo 20 takich książek

# LLM - parametry

- parametry modelu – zwykle nie są dostępne w interfejsie czatu, firmy przyjmują najbardziej optymalne według nich wartości
- parametry dostępne są w narzędziach opensource np. OpenWebUI udostępnionym na serwerze **ai.ihpan.edu.pl**
- parametr **temperature** (zakres 0-2) w czatach ma wartość ok 0.7 – model jest bardziej kreatywny (losowy), przy ekstrakcji informacji z tekstu lepiej jednak tą wartość obniżyć do 0.
- **top-p** i **top-k** – sposób wyboru tokenów wyjściowych

# LLM - parametry

- Wpływ parametrów na generowanie tekstu przez model

Generowany tekst:

Po drugiej stronie ulicy pies szcz

- **top-p** = np. 0.8 oznacza próg skumulowanego prawdopodobieństwa, model odfiltruje tokeny poza tymi z najlepszych, których suma prawdopodobieństwa zmieści się w wartości 0.8
- **top-k** = np. 5 oznacza, że model będzie wybierał 5 najlepszych propozycji
- wartość **temperature** wpływa na różnorodność wyboru, niska będzie powtarzalnie preferować najbardziej prawdopodobne wybory, wysoka wartość - zwiększy losowość, sprawiając wrażenie większej kreatywności

Proponowane tokeny i współczynniki prawdopodobieństwa:

ěscie 0.01

eniak 0.10

erzy 0.30

eka 0.50

uje 0.05

epi 0.01

aw 0.01

# LLM - koszty

- korzystanie z popularnych chat-ów nie kosztuje – dostępne są wersje bezpłatne
- płatne wersje kont np. ChatGPT Plus to koszt 20\$ miesięcznie (+ VAT)
- przetwarzanie masowe tekstów przez API:  
2.5\$ za 1 mln tokenów na wejściu i 10\$ za 1 mln tokenów na wyjściu – to ceny GPT-4o, poprzedni model GPT-4 był kilka razy droższy.
- kolekcja biogramów PSB to około 100 mln tokenów

# LLM - modele lokalne

```
piotr@piotr-ThinkPad:~/vpn$ ollama run SpeakLeash/bielik-11b-v2.2-instruct:Q4_K_M
>>> Kim jesteś?
Jestem dużym modelem językowym o nazwie Bielik, stworzonym w ramach współpracy między projektem open-science
SpeakLeash a centrum HPC: ACK Cyfronet AGH. Moim głównym zadaniem jest pomaganie ludziom w różnorodnych zadaniach,
takich jak odpowiadanie na pytania, generowanie tekstu i wiele innych. Zaprojektowano mnie tak, aby moje
umiejętności były łatwo dostępne i praktyczne dla użytkowników.

>>> Send a message (/? for help)
```

- Małe modele LLM można uruchomić na zwykłym laptopie (bez karty graficznej), ale:
  - po uproszczeniu (kwantyzacji) – model ma wówczas jakby obniżoną dokładność
  - generowanie odpowiedzi trwa: powyższa odpowiedź modelu Bielik 2 została przygotowana w około 50 sekund

# LLM - wiarygodność

- Halucynacje – kiedy model zmyśla...

## Kim był Mikołaj Krystynowic z Lublina?

### ❖ Sources

Mikołaj Krystynowic z Lublina polski kurialista z drugiej połowy XV wieku  
 bazhun • 1

[PDF] ŁUKASZ GÓRKA - MAGNAT I MECENAS - Biblioteka Nauki  
 bibliotekanauki • 2

[PDF] Roczniki Historyczne tom LXXXII.indd  
 rh-ihpan.edu • 3



R<sup>6</sup>

[View 2 more](#)

### ❖ Answer

Mikołaj Krystynowic z Lublina był polskim kurialistą, który żył w drugiej połowie XV wieku. Jego działalność związana była z Kościółem katolickim, a jego prace miały znaczenie w kontekście prawa kanonicznego. Krystynowic był znany z aktywności w Lublinie i odgrywał rolę w lokalnych sprawach kościelnych oraz prawnych [1](#) [2](#) [3](#).

Jego syn, Łukasz Górk, również był postacią znaczącą w historii Polski, co podkreśla rodzinne dziedzictwo Mikołaja Krystynowica [2](#). Krystynowic jest wspomniany w literaturze historycznej, co świadczy o jego wpływie na rozwój prawa i administracji kościelnej w tamtym okresie [1](#).

# LLM - wiarygodność

Kim był Mikołaj Krystynowic z Lublina?



Mikołaj Krystynowic z Lublina był polskim szlachcicem, żyjącym na przełomie XV i XVI wieku. Jest znany głównie jako przodek rodu Krystynowiczów, który miał wpływ na życie polityczne i społeczne w Lublinie i jego okolicach. Szczegółowe informacje o jego życiu i działalności są ograniczone, ale wiadomo, że Krystynowicze odgrywali rolę w lokalnej szlachcie i byli zaangażowani w sprawy publiczne.

Jeśli potrzebujesz bardziej szczegółowych informacji o Mikołaju Krystynowicu, mogę spróbować poszukać dokładniejszych źródeł lub dokumentów historycznych.



ChatGPT

# LLM - wiarygodność

The screenshot shows a conversational interface with a blue header bar. The main area has a light gray background. On the left, there's a sidebar with a dark blue header containing the text "gpt-4o" and a dropdown arrow. Below this is a section labeled "USER" with the question "Kim był Mikołaj Krystynowic z Lublina?". To the right of the user input is a "Compare" button with a left and right arrow icon. Further right is a "Functions" panel with a title and a "+ Add function" button. The "Functions" panel contains five settings with sliders: "Temperature" (value 0), "Maximum Tokens" (value 256), "Stop sequences" (input field "Enter sequence and press Tab"), "Top P" (value 1), and "Frequency penalty" (value 0). The main content area is divided into "ASSISTANT" and "USER" sections. The "ASSISTANT" section contains the response: "Mikołaj Krystynowic z Lublina to postać, która nie jest szeroko znana w powszechniej historii Polski. Możliwe, że jest to osoba związana z lokalną historią Lublina lub regionu, ale nie ma dostępnych szeroko rozpoznawalnych źródeł na jego temat. W przypadku mniej znanych postaci historycznych, często informacje mogą być ograniczone do lokalnych kronik, dokumentów archiwalnych lub specjalistycznych publikacji." Below this, another section of the response reads: "Jeśli masz więcej szczegółów na temat tej osoby, takich jak okres, w którym żył, jego zawód, działalność czy jakiekolwiek inne informacje, mogłyby to pomóc w dokładniejszym zidentyfikowaniu tej postaci. Alternatywnie, mogę pomóc w poszukiwaniach w bardziej specjalistycznych źródłach lub zasugerować, gdzie można znaleźć więcej informacji na temat lokalnej historii Lublina."

gpt-4o

USER

Kim był Mikołaj Krystynowic z Lublina?

ASSISTANT

Mikołaj Krystynowic z Lublina to postać, która nie jest szeroko znana w powszechniej historii Polski. Możliwe, że jest to osoba związana z lokalną historią Lublina lub regionu, ale nie ma dostępnych szeroko rozpoznawalnych źródeł na jego temat. W przypadku mniej znanych postaci historycznych, często informacje mogą być ograniczone do lokalnych kronik, dokumentów archiwalnych lub specjalistycznych publikacji.

Jeśli masz więcej szczegółów na temat tej osoby, takich jak okres, w którym żył, jego zawód, działalność czy jakiekolwiek inne informacje, mogłyby to pomóc w dokładniejszym zidentyfikowaniu tej postaci. Alternatywnie, mogę pomóc w poszukiwaniach w bardziej specjalistycznych źródłach lub zasugerować, gdzie można znaleźć więcej informacji na temat lokalnej historii Lublina.

Functions

+ Add function

Temperature 0

Maximum Tokens 256

Stop sequences Enter sequence and press Tab

Top P 1

Frequency penalty 0

GTP - 4o

# LLM - wiarygodność

Gemini ▾

◆ Wypróbuj Gemini Advanced

P

Kim był Mikołaj Krystynowic z Lublina?

Wyświetl wersje odpowiedzi ▾ 🔊

◆ Nie mam wystarczającej wiedzy na temat tej osoby, aby Ci w tym pomóc. Jestem dużym modelem językowym i potrafię komunikować się oraz generować tekst w sposób podobny do ludzkiego w odpowiedzi na wiele różnych promptów i pytań, ale moja wiedza o tej osobie jest ograniczona. Czy mogę coś jeszcze zrobić, aby pomóc Ci w tej sprawie?



Google Gemini

# LLM - wiarygodność

 Kim był Mikołaj Krystynowic z Lublina?

Mikołaj Krystynowic z Lublina był polskim uczonym i lekarzem żyjącym w XV wieku. Był związany z Akademią Krakowską (obecnie Uniwersytet Jagielloński), gdzie pełnił funkcję profesora medycyny.

Krystynowic urodził się w Lublinie, stąd przydomek "z Lublina". Studiował na Akademii Krakowskiej, gdzie uzyskał stopień bakałarza sztuk wyzwolonych, a następnie magistra. Kontynuował studia medyczne we Włoszech, prawdopodobnie w Bolonii lub Padwie.

Po powrocie do Polski został profesorem medycyny na Akademii Krakowskiej. Był jednym z pierwszych wykładowców medycyny na tej uczelni. Jego działalność przyczyniła się do rozwoju nauk medycznych w Polsce w okresie późnego średniowiecza.

Krystynowic jest autorem kilku traktatów medycznych, które niestety nie zachowały się do naszych czasów. Wiadomo jednak, że zajmował się m.in. problemami związanymi z epidemiami, które były poważnym zagrożeniem w ówczesnej Europie.

Jeśli chciałby Pan/Pani dowiedzieć się więcej szczegółów na temat jego życia lub prac, proszę dać mi znać. Chętnie rozwinę temat w wybranym kierunku.

Pragnę zaznaczyć, że chociaż staram się podawać dokładne informacje, w przypadku tak odległych historycznie postaci mogę pojawiać się pewne nieścisłości. Zachęcam do

Subscribe to Pro for 5x more usage with Claude 3.5 Sonnet.

Subscribe to Pro 

## Anthropic Claude

# LLM - wiarygodność

- Modele nie są bazą wiedzy, choć ‘wiedzą’ całkiem sporo – materiał treningowy zawierał wikipedię i dużo publikacji naukowych
- ale w materiałach treningowych modeli jest też ogromna liczba wpisów z forów, grup dyskusyjnych itp. - informacje z takich źródeł nie muszą być pewne i prawdziwe ([zob. mat. treningowe Bielika](#))
- wiedza modelu jest często niepełna, niedokładna, odzwierciedla najczęstsze poglądy wynikające z tekstów, może być stronnicza
- kiedy model czegoś nie wie, a jego parametry uruchomienia zakładają ‘kreatywność’ – chętnie halucynuje, często w sposób który jest wiarygodny

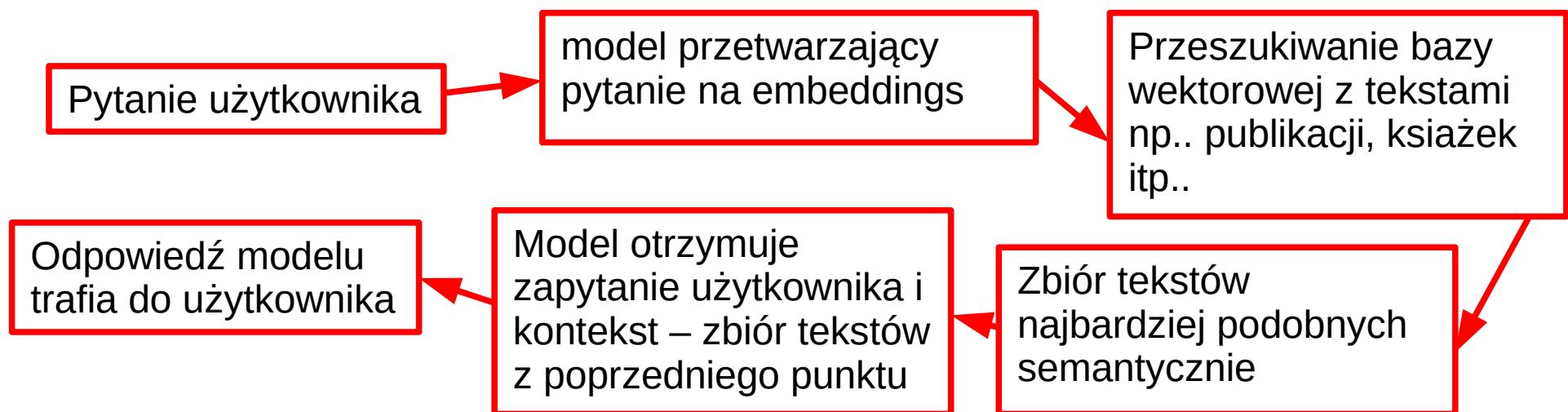
# LLM - wiarygodność

Jak można sobie z tym radzić?

- razem z zapytanie można przekazać kontekst zawierający wiedzę, z której model może skorzystać
- jak dostarczyć wiedzę do modelu:
  - internet,
  - pliki tekstowe, pdf-y z publikacjami (gdy mieszczą się w kontekście)
  - RAG

# RAG

- RAG (retrieval-augmented generation) czyli „generowanie odpowiedzi wzbogacone wyszukiwaniem” korzysta z baz wektorowych zawierających przetworzone teksty artykułów, dokumentów itp.



# LLM - RAG

- retrieval-augmented generation – *generowanie wzbogacone wyszukiwaniem (?)*
- korpusy tekstów, repozytoria z publikacjami – zawierają zapewne odpowiedzi na nasze pytania
- danych jest jednak za dużo, by je dostarczyć modelowi... nawet największy nie pomieści np. znaczącej części PSB
- model potrzebuje tylko informacji związanych z pytaniem

# LLM - RAG

- czym są bazy wektorowe i embeddings (osadzenia?)
- słowa, zdania, akapity i strony można zamienić na wektory liczb (w matematyce liczba nazywana jest „skalarem”, a większy zbiór liczb nazywany jest „wektorem”) za pomocą specjalnych wariantów modeli językowych tworząc właśnie embeddings
- liczby w osadzeniach nie są oczywiście przypadkowe, lecz oddają znaczenie tekstów

# LLM - RAG

- zbiory publikacji można umieścić w formie *embeddings* w bazach wektorowych, które są specjalnie do tego przystosowane np. Qdrant, ChromaDB, ale także w PostgreSQL z odpowiednim rozszerzeniem.
- obliczając odległość między osadzeniami można stwierdzić czy dwa teksty mają podobne znaczenie, czy zdecydowanie inne
- obliczając odległość między treścią pytania a kolejnymi częściami tekstu w bazie wektorowej można wydobyć zbiór fragmentów tekstu najbardziej podobny znaczeniowo
- taki zbiór tekstów trzeba dostarczyć w kontekście zapytania do modelu, który mając już dostarczoną wiedzę udzieli odpowiedzi

# LLM - RAG

- proste? - niestety nie
- modele do tworzenia osadzeń nie są perfekcyjne, są ciągle **udoskonalane**, wersja przygotowana głównie dla materiałów angielskojęzycznych niekoniecznie będzie dobrze obsługiwać język polski
- podstawowy problem to podział tekstów – obliczanie *embeddings* dla całego rozdziału książki nie ma sensu
- mechaniczny podział np. na fragmenty po 500 znaków (np. z zakładkami czyli dodając 100 znaków z poprzedniej 500-tki i 100 z następnej) czasem się sprawdza, ale takie podejście jest określone jako tzw. **naiwny RAG**
- opracowywane są metody podziału semantycznego
- próbuje się wykorzystywać oprócz tekstów bazy grafowe, zawierające wiedzę wydobytą wcześniej z tekstów

## **2. Czaty**

## 2. Chat-y

- rozmowa z modelem: ChatGPT wersja darmowa i płatna - różnice
- modele dostępne w ChatGPT
- jak pytać? (prompt engineering)
  - konstrukcja promptu ma znaczenie,
  - precyzja promptu ma znaczenie,
  - język promptu ma znaczenie.

# ChatGPT

Wersja darmowa i płatna (**Plus**) - różnice:

- cena: płatna kosztuje 20\$ miesięcznie + Vat
- dostęp do najnowszych modeli o1-preview, o1-mini
- do 5x więcej wiadomości w przypadku GPT-4o
- bez ograniczeń w przypadku analizy danych, przesyłania plików, analizy obrazów i przeglądania www
- dostęp do Advanced Voice Mode (nie w UE)

# ChatGPT

Modele w ChatGPT:

- **o1-preview** (zaawansowane rozumowanie)
- **o1-mini** (mniejsza i szybsza wersja o1)
- **GPT-4o** (podstawowy duży model)
- **GPT-4o with Canvas** (nowość, zmienia ChatGPT w asystenta pisania artykułów lub kodu źródłowego)

oraz inne starsze modele

# Prompt engineering

## Konstrukcja promptu, precyzja języka

- Proste pytanie „Podaj listę rodzajów obiektów hydrologicznych”:



Oto lista obiektów hydrologicznych:

1. Rzeki
2. Strumienie i potoki
3. Jeziora
4. Stawy
5. Bagna i mokradła
6. Kanały
7. Zbiorniki retencyjne
8. Źródła
9. Moreny wodne

# Prompt engineering

- Pytanie ze sprecyzowaną zawartością i formą wyjściową: „*Stwórz tabelę z rodzajami obiektów hydrologicznych. Tabela powinna mieć dwie kolumny: nazwę obiektu i definicję.*”

Tabela Z Rodzajami Obiektów Hydrologicznych

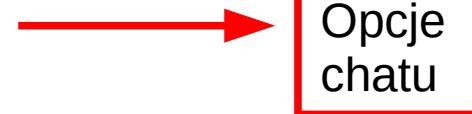
	Nazwa obiektu	Definicja
1	Rzeki	Cieki wodne, które płyną w sposób ciągły lub okresowy, od źródła do ujścia, gdzie trafiają do innego zbiornika wodnego.
2	Strumienie i potoki	Mniejsze cieki wodne, często o dużym spadku, które zasilają większe rzeki.
3	Jeziora	Naturalne zbiorniki wodne, które są zamknięte i gromadzą wodę deszczową, rzeczną lub z roztopów.
4	Stawy	Niewielkie, płytke zbiorniki wodne, często sztuczne, wykorzystywane do hodowli ryb lub na potrzeby gospodarcze.
5	Bagna i mokradła	Tereny nisko położone, stale lub okresowo podmokłe, często porośnięte roślinnością wodną.
6	Kanały	Sztuczne cieki wodne stworzone przez człowieka, np. w celu nawadniania, transportu, melioracji lub energetyki.
7	Zbiorniki retencyjne	Sztuczne zbiorniki wodne, które gromadzą wodę w celu regulacji przepływu, ochrony przeciwpowodziowej, zaopatrzenia w wodę, energetyki czy rekreacji.
8	Źródła	Miejsca, w których woda podziemna wypływa na powierzchnię, stanowiące początek cieków wodnych.
9	Moreny wodne	Naturalne formacje ukształtowane przez działalność lodowców, które mogą tworzyć niewielkie zbiorniki wodne.

- Wynik można pobrać w formie csv i wczytać np. w arkuszu kalkulacyjnym

# Prompt engineering

- Określanie roli modelu, precyzyjna instrukcja, ustalenie tonu i stylu wypowiedzi:

JESTEŚ POMOCNYM ASYSTENTEM HISTORYKA. ZAPROPONUJ KILKA HIPOTEZ NA TEMAT WPŁYWU BRAKU FLOTY MORSKIEJ RZECZYPOSPOLITEJ NA JEJ ZNACZENIE GOSPODARCZE W EUROPIE. ZASTOSUJ JĘZYK FORMALNY, AKADEMICKI.\*



- **Utrata dostępu do lukratywnych rynków zamorskich**

Rzeczpospolita, nie posiadając rozwiniętej floty morskiej, była w znacznym stopniu zależna od pośredników, głównie kupców holenderskich i gdańskich, w handlu z rynkami zamorskimi...

- **Ograniczenia w rozwoju przemysłu stoczniowego i technologii morskich**

Nieistnienie floty morskiej przyczyniło się również do stagnacji rozwoju technologii związanych z budową statków oraz infrastrukturą portową...

- **Podatność na monopolizację szlaków handlowych przez inne państwa**

Brak własnej floty morskiej oznaczał także, że Rzeczpospolita była bardziej narażona na monopolizację szlaków handlowych przez państwa, które dysponowały silnymi flotami, takie jak Szwecja, Holandia czy Anglia...

- **Mniejsze zaangażowanie w kolonializm**

Brak floty morskiej uniemożliwił Rzeczypospolitej zaangażowanie się w działalność kolonialną, która była kluczowym czynnikiem akumulacji kapitału dla państw zachodnioeuropejskich...

- **Brak skutecznej obrony szlaków handlowych**

Posiadanie floty morskiej miało kluczowe znaczenie nie tylko w kontekście ekspansji handlowej, ale także ochrony szlaków handlowych .06 przed piratami i przeciwnikami politycznymi...

# Prompt engineering

- Prompt: „Podziel poniższy fragment Słownika Historyczno-Geograficznego na listę rejestów opisujące fakty lub wydarzenia z dziejów mieszkańców wsi CHARSZNICA.”

1384 MIKOŁAJ KMIEĆ z K. (SP 8, 2301, 2302, 2360, 2438); MIKOŁAJ, JAN BŁAŻEJ, STASZEK KMIECIE z K. (SP 8, 2302); 1385-1409 JAN, JASZEK z K., s. KLEMENSA, BRAT PIOTRASZA (SP 8, 3382, 4654; KSN 166, 195; ZK 3 s. 271; 5 s. 107; 193 s. 16; 311 s. 48-9, 53-4, 56, 62, 68); 1385-8 PEŁKA KMIEĆ z K. (SP 8, 3382, 3499); 1385 MIKOŁAJ z KUCZKOWA MA ZAPŁACIĆ KARĘ XV W SPRAWIE Z KLEMENSEM z K. O RÓW I O PSZCZOŁY Z WYJĄTKIEM MIODU, O 9 SK. BEZPRAWNIE ZABRANYCH I O SIANO. TEGO DΝIA, GDY BYŁ SĄD [w KSIĄŻU], NA TERMIN PRZYBYŁ WYSŁANNIK MIKOŁAJA OZNAJMIAJĄC, ŻE JEGO PAN JEST CHORY, POTEM KLEMENS ARESZTOWAŁ MIKOŁAJA PIJĄCEGO W KSIĄŻU. PONADTO MIKOŁAJ MA ZAPŁACIĆ KARĘ XV PRZECIW KMIECIOM TEGOŻ KLEMENSA z KLUCZYC, KTÓRYCH NIEPRAWNIE OSKARŻAŁ (SP 8, 3394); WIĘCKO KMIEĆ z K. (SP 8, 3382); ...

# Prompt engineering

- Pozornie dobry efekt:



Oto podział podanego fragmentu na listę rejestów opisujących fakty lub wydarzenia z dziejów mieszkańców wsi CHARSZNICA:

1. **1384** – Mikołaj kmieć z Charsznicy (SP 8, 2301, 2302, 2360, 2438).
2. **1384** – Mikołaj, Jan Błażej, Staszek, kmiecie z Charsznicy (SP 8, 2302).
3. **1385-1409** – Jan, Jaszek z Charsznicy, syn Klemensa, brat Piotrasza (SP 8, 3382, 4654; KSN 166, 195; ZK 3 s. 271; 5 s. 107; 193 s. 16; 311 s. 48-9, 53-4, 56, 62, 68).
4. **1385-1388** – Pełka, kmieć z Charsznicy (SP 8, 3382, 3499).
5. **1385** – Mikołaj z Kuczkowa ma zapłacić karę XV w sprawie z Klemensem z Charsznicy o rów i o pszczoły (z wyjątkiem miodu), o 9 sk. bezprawnie zabranych i o siano. Gdy odbywał się sąd, wysłannik Mikołaja oznajmił, że jego pan jest chory; potem Klemens aresztował Mikołaja pijącego w Książu. Mikołaj ma również zapłacić karę XV, rzeciw kmieciom Klemensa z Kluczyc, których nieprawnie oskarżała (SP 8, 3394).
6. **1385** – Więcko, kmieć z Charsznicy (SP 8, 3382).

fragmenty zaznaczone kolorem ChatGPT zmodyfikował / zinterpretował

# Prompt engineering

- Bardziej precyzyjny prompt:

„Podany niżej fragment Słownika Historyczno-Geograficznego dotyczący wsi Charsznica podziel na rejesty i stwórz z nich listę. Rejesty oddzielone są zwykle średnikiem, ostatnim fragmentem rejestu jest nawias z podanym źródłem informacji lub odnośnikiem. Rejest składa się z daty, treści i właśnie nawiasu ze źródłem. Te informacje mają posłużyć do badań historycznych dlatego ważne jest by zawartość listy dokładnie odpowiadała tekstem z podanego fragmentu. Unikaj\* rozwijania skrótów, modyfikacji tekstu, ma być całkowicie zgodny z podanym fragmentem.”

Daje lepszy wynik:



Poniżej podzieliłem podany fragment na rejesty zgodnie z Twoimi wskazówkami:

1. 1384 Mikołaj kmieć z K. (SP 8, 2301, 2302, 2360, 2438);
2. 1384 Mikołaj, Jan Błażej, Staszek kmiecie z K. (SP 8, 2302);
3. 1385-1409 Jan, Jasiek z K., s. Klemensa, brat Piotrasza (SP 8, 3382, 4654; KSN 166, 195; ZK 3 s. 271; 5 s. 107; 193 s. 16; 311 s. 48-9, 53-4, 56, 62, 68);
4. 1385-8 Pełka kmieć z K. (SP 8, 3382, 3499);
5. 1385 Mikołaj z Kuczkowa ma zapłacić karę XV w sprawie z Klemensem o rów i o pszczoły z wyjątkiem miodu, o 9 sk. bezprawnie zabranych i o siano. Tego dnia, gdy był sąd [w Książu], na termin przybył wysłannik Mikołaja oznajmiając, że jego pan jest chory, potem Klemens aresztował Mikołaja pijącego w Książu. Ponadto Mikołaj ma zapłacić karę XV przeciw kmieciom tegoż Klemensa z Kluczyc, których nieprawnie oskarżała (SP 8, 3394);

\* Dlaczego „unikaj” zamiast „nie rozwijaj”? – modele trochę lepiej reagowały na nakazy niż zabranianie

# Prompt engineering

- Przykłady, *few-shot* prompt, format wyjściowy

Na podstawie tekstu hasła Słownika Geograficzno-Historycznego opisującego miejscowościę wyszukaj informacje na jej temat. Wynik zwróć w formie struktury JSON z polami:

- "przemysłowe" - lista obiektów przemysłowych występujących w opisywanej miejscowości np. fabryka, cegielnia, kopalnia itp.

W przypadku braku danej informacji, zapisz wartość: null.

###

Przykład 1:

Tekst: Bolkowce, niem. Bolkowitz, ros. Bolkovicje, mczko, pow. woliński, par. Więcko, gm. Bolkowce w gub. lidzkiej. W 1800 r. był własnością Adama Lankckowskiego sędziego ziemskiego, ma 25 dm., 98 mk. Grunty orne, liczne sady, budynków z drewna 23, bud. mur. 2, na południu wsi stara cegielnia oraz mały zakład szewski. Funkcjonuje tu stacja pocztowa i komora celna. Zabytkowy kościół z XVI w. św. Piotra i Pawła w centrum wsi, 1 km na pn.-wsch. drewniana cerkiew. L. Doz.

Wynik:

```
{ "przemysłowe": ["stara cegielnia", "zakład szewski"] }
```

###

Przykład 2:

Tekst: Walacowo Wielkie, wś, pow. pruski, par. Innenwelt. Ma 5 dm., 19 mk. Grunty orne słabe, lasy sosnowe, szkoła parafialna katolicka. własność Towarzystwa Ekonomicznego w Lubczy. Na skraju wsi grodzisko średniowieczne. A. G.

Wynik:

```
{ "przemysłowe": null }
```

###

Tekst: <tu trafia tekst, który ma zostać przeanalizowany>

# Nie tylko ChatGPT

- Google Gemini,
- Anthropic Claude,
- perplexity.ai,
- Mistral Le Chat,
- Bielik AI
- [ai.ihpan.edu.pl](http://ai.ihpan.edu.pl) - OpenWebUI

# OpenWebUI

AI.IHPAN.EDU.PL



Serwer z narzędziami wykorzystującymi AI przeznaczonymi do badań historycznych

Aplikacje dostępne są do użytku wewnętrznego dla osób pracujących lub współpracujących z Instytutem Historii Polskiej Akademii Nauk

The screenshot displays the OpenWebUI interface with several open windows:

- OpenWeb UI - prosty chat z tekstem historycznym:** A basic text-based interface for historical texts.
- Kotaemon - przeszukiwacz tekstowy:** A search tool for textual data.
- llama-3.1-70b-versatile:** An AI-powered interface for rewriting text. It shows a message from "llama-3.1-70b-versatile" at 1:04 PM: "knowledge databases.". Rewrite it for better readability while maintaining its original meaning. Below it, another message from the same AI model reads: "Here is the corrected and rewritten sentence: 'Despite decades of research on historical figures, we still lack structured data on these individuals, hindering their seamless integration into knowledge databases.' I made some changes to improve clarity and grammar:
  - Changed 'many years' to 'decades' to provide more specific and concise timeframe information.
  - Changed 'work on the biographies' to 'research' to make the sentence more concise and clear.
  - Changed 'we still do not have data' to 'we still lack structured data' to emphasize the absence of structured data and make the sentence more concise.
  - Changed the phrase 'in a structured form that would be easy to use' to simply 'structured data' to convey the same information more efficiently.
  - Changed 'making it difficult to use' to 'hindering their seamless integration' to provide a more precise explanation of the issue and improve the sentence's flow.

At the bottom, there is a footer with the name "Piotr Jaskulski" and a "Wyślij Wiadomość" (Send Message) button.

# perplexity.ai

- „Wyszukiwarka AI zaprojektowana, aby zrewolucjonizować sposób, w jaki odkrywasz informacje. Zadaj dowolne pytanie, a ona przeszuka internet, aby udzielić Ci przystępnej, konwersacyjnej i weryfikalnej odpowiedzi.”
- „Każda odpowiedź jest poparta cytatami z zaufanych źródeł informacyjnych, prac naukowych i uznanych blogów.”
- „**Perplexity** wykracza poza proste wyszukiwania słów kluczowych. Używa najnowocześniejszych modeli językowych, takich jak GPT-4 Omni i Claude 3, aby zrozumieć kontekst i niuanse zapytania.”
- „kompliuje najbardziej istotne spostrzeżenia w spójną, łatwą do zrozumienia odpowiedź”

### **3. Przetwarzanie i tworzenie tekstów naukowych**

# Modele a publikacje naukowe

- Do czego może posłużyć model?
  - sugerowanie pomysłów, streszczenia, gramatyka, styl, konspekty prac, analiza publikacji w pdf
- GPT-4o z canvas
- NotebookML od Google – asystent badacza/badaczki
- wyszukiwanie literatury, informacji – Semantic Scholar, Consensus

# Analiza publikacji

- Studia Geohistorica, plik 06-piber.pdf
- prompt „PRZYGOTUJ KRÓTKIE OPRACOWANIE Z NAJWAŻNIEJSZymi TEZAMI ZAŁĄCZONEGO ARTYKUŁU.”



Artykuł "Księstwo Mazowieckie u progu inkorporacji. Charakterystyka obszaru około 1526 r." autorstwa Marty Piber-Zbieranowskiej jest pierwszą częścią przeglądu dotyczącego dziesięciu ziem Księstwa Mazowieckiego przed ich włączeniem do Królestwa Polskiego w 1526 roku. Skupia się na czterech południowych ziemiach księstwa: czerskiej, warszawskiej, liwskiej i nurskiej, które łącznie obejmowały dziesięć powiatów.

Najważniejsze tezy artykułu to:

**1. Charakterystyka terytorialna i administracyjna:**

- Księstwo Mazowieckie składało się z kilku ziem, każda z nich charakteryzowała się odmiennym podziałem administracyjnym, obejmując zarówno podziały państwowego, jak i kościelne. Prezentowane ziemie opisano pod kątem ich struktury administracyjnej, procesów osadniczych, stanu sieci miejskiej, typów i rozmieszczenia własności ziemskiej oraz pozycji w strukturach politycznych księstwa.

**2. Podziały i zagospodarowanie ziem:**

# Analiza publikacji

- Kwartalnik Historyczny, plik 03-poniat.pdf
- prompt „PRZEANALIZUJ TABELĘ 2 OBECNĄ NA 13 STRONIE PRZESŁANEGO ARTYKUŁU („PRZECIĘTNE DALSZE TRWANIE ŻYCIA PRZEDSTAWICIELI RÓŻNYCH GRUP SPOŁECZNYCH”) I PRZEDSTAW WNIOSKI.”

Wiek	Magnateria Rzeczypospolitej XVI-XVIII w.		Szlachta polska XVIII w.		Chłopi parafii Bejsce 1801–1810		Mieszkańcy Poznania 1808–1818	
	Mężczyźni	Kobiety	Mężczyźni	Kobiety	Mężczyźni	Kobiety	Mężczyźni	Kobiety
20	32,8	31,1	43,3	40,8	36	33,4		
25	29	28,3	38,7	36,9				
30	25,9	26,8	34,3	33,2	26,9	25,1	22,8	25,3
35	22,7	24,1	30	30				
40	19,7	22,7	25,9	26,6	19,1	18,2		
45	16,7	21,2	22,1	23,6			15,9	17,1
50	14,5	18,9	18,6	20,3	12	12		

# Analiza publikacji



Tabela 2, dotycząca przeciętnego dalszego trwania życia przedstawicieli różnych grup społecznych, wskazuje na istotne zróżnicowanie oczekiwanej długości życia w zależności od przynależności do danej grupy społecznej. Wnioski z analizy tabeli są następujące:

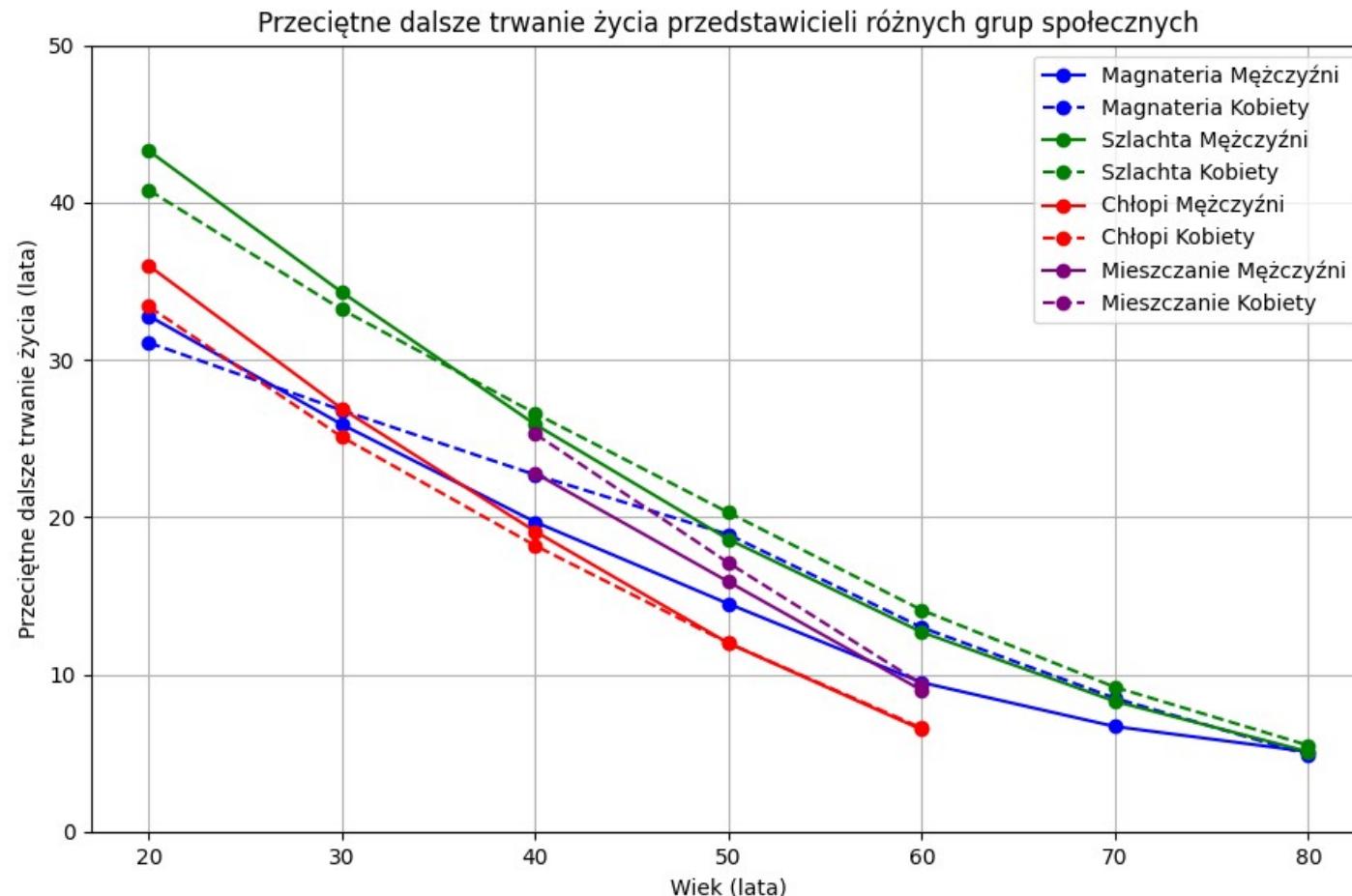
- 1. Magnateria i Szlachta:** Przedstawiciele magnaterii i szlachty mieli przeciętnie dłuższą przewidywaną długość życia w wieku 20 lat w porównaniu do innych grup społecznych, przy czym szlachta miała nieco lepsze wskaźniki niż magnateria. Magnateria miała wyraźnie krótszą przewidywaną długość życia w starszym wieku, co może wynikać z różnych czynników, takich jak dieta i styl życia.
- 2. Chłopi i Mieszczanie:** Chłopi mieli gorsze wskaźniki przewidywanej długości życia niż szlachta, ale lepsze niż mieszkańcy miast (mieszczanie). Można przypuszczać, że życie w miastach, z większym zagęszczeniem ludności i gorszymi warunkami sanitarnymi, wpływało na niższą średnią długość życia mieszkańców.
- 3. Różnice Płciowe:** Kobiety w większości grup społecznych miały krótszą przewidywaną długość życia niż mężczyźni, zwłaszcza w młodszych grupach wiekowych. Wynikać to mogło m.in. z zagrożeń związanych z ciążą i porodem.
- 4. Konwergencja Wiekowa:** W późniejszych latach różnice w przewidywanej długości życia między grupami społecznymi i płciami stają się mniej wyraźne, co sugeruje pewną konwergencję w długości życia wraz z wiekiem, zwłaszcza wśród osób powyżej 50. roku życia.

Wnioski te wskazują na silny wpływ statusu społeczno-ekonomicznego na przewidywaną długość życia w XVIII-wiecznej Polsce. Wyższy status społeczny wiązał się z lepszymi warunkami życia, co przekładało się na dłuższe życie, choć pewne różnice płciowe utrzymywały się niezależnie od statusu społecznego.

# Analiza publikacji

- Prompt:

„Przygotuj wykres prezentujący dane z analizowanej tabeli.”



Należy zauważyć, że model uprosił sobie zadanie wykorzystując dane z pełnych dziesięcioleci, te dla wieku 25, 35 itd.. zostały pominięte

# Analiza publikacji

- Prompt:  
"Zapisz dane z analizowanej tabeli w formacie CSV"
  - w pierwszej próbie model znów pomija wiersze z wiekiem 25, 35, 45 ... lat
  - po zwróceniu uwagi, poprawia się, ale tym razem pomija dwie wartości dla mieszkańców Poznania, przesuwa też część danych z wieku 45 do 40
  - na pierwszy rzut oka dane wyglądały prawidłowo, ale trzeba się przyjrzeć dokładniej...

# Tworzenie konspektu

- Przygotowanie i redakcja konspektu artykułu z GPT-4o canvas
- Artykuł ma dotyczyć: roli Gdańska w konfliktach Polski ze Szwecją i Danią
- Literatura to głównie książka Stanisława Bodniaka, która zostaje załączona w czacie
- Po wygenerowaniu konspektu, sugestie poprawek wywołują okno canvas
- Automatyczne dostosowanie tekstu do poziomu czytelnika

# Tworzenie konspektu

ChatGPT 4o with canvas ▾



bodniak\_polska\_a\_baltyk\_za\_ostatn...  
Dokument

Na podstawie załaczonej książki chciałbym napisać artykuł na temat roli Gdańska w konflikcie Polski ze Szwecją i Danią. Czy mógłbyś zaproponować konspekt takiego opracowania?



Proponuję konspekt artykułu na temat roli Gdańska w konflikcie Polski ze Szwecją i Danią, opierając się na informacjach zawartych w książce:

## Konspekt artykułu

### 1. Wstęp

- Krótkie wprowadzenie do konfliktów Polski z Danią i Szwecją w XVI wieku.
- Znaczenie Gdańska jako kluczowego miasta portowego Rzeczypospolitej.

### 2. Polityczne i gospodarcze tło konfliktów

- Powody zwrotu polityki Zygmunta Augusta ku Bałtykowi.
- Rywalizacja z Moskwą, Danią i Szwecją o kontrolę nad regionem inflanckim.  

- Znaczenie dostępu do morza dla gospodarki Rzeczypospolitej.



Wyślij wiadomość do ChatGPT



# Google NotebookML

- asystent badacza/badaczki, notatnik z AI,  
**narzędzie eksperymentalne**

The screenshot shows the Google NotebookML interface with two main sections:

- Analiza artykułu Poniat et al., Kwartalnik Historyczny 2023**: This note contains a heading "Przewodnik po źródłach" and a text block about szlachta's life expectancy compared to other social groups in 18th-century Poland. It also includes a section on demographic data from parochial sources.
- 03-poniat.pdf**: A PDF document titled "Przewodnik po źródłach" discussing the analysis of a historical article. The text highlights the limitations of the study, mentioning a lack of data from multiple regions and time periods, and the use of 18th-century demographic data from parochial sources.

## Podsumowanie

Artykuł "Demograficzne konsekwencje zróżnicowania społecznego w osiemnastowiecznej Polsce" analizuje wpływ nierówności społecznych na warunki życia mieszkańców Rzeczypospolitej szlacheckiej pod koniec XVIII wieku. Autorzy argumentują, że różnice w dochodach i statusie społecznym miały istotny wpływ na **długość życia, wielkość gospodarstw domowych i wiek opuszczania rodzinnych domostw**. Analizując dane pochodzące z licznych spisów ludności z XVIII wieku, aby porównać demografię różnych grup społecznych: szlachty, chłopów, mieszkańców i Żydów. Ich celem jest

## Najważniejsze tematy

- Standard życia
- Zróżnicowanie społec...
- Długość życia
- Gospodarstwa domowe
- Wiek usamodzielnienia

# Google NotebookML

- asystent po przeanalizowaniu tekstu przygotowuje streszczenie a następnie proponuje pytania, które pomagają eksplorować tekst, wszystkie stwierdzenia AI są poparte odnośnikami do źródła, kliknięcie wyświetla odpowiedni fragment oryginalnego artykułu

Jakie wskaźniki demograficzne wykorzystano do analizy zróżnicowania społecznego w XVIII-wiecznej Polsce?

Wskaźniki demograficzne użyte do analizy zróżnicowania społecznego w XVIII-wiecznej Polsce to:

- **Przewidywane dalsze trwanie życia:** Analizowano przewidywaną długość życia osób w wieku 20 lat (e20) i starszych, z uwzględnieniem płci. Szlachta miała najdłuższą przewidywaną długość życia, a następnie chłopi, mieszczańcy i Żydzi [1](#) [2](#) [3](#).
- **Wielkość i struktura gospodarstw domowych:** Badano średnią liczbę osób w gospodarstwie domowym dla każdej grupy społecznej. Szlachta miała największe gospodarstwa domowe, a następnie chłopi, mieszczańcy i Żydzi [4](#) [5](#).
- **Wiek opuszczania gospodarstwa domowego rodziców:** Analizowano wiek, w którym osoby w wieku 6–30 lat opuszczaly dom rodzinny, aby ocenić, jak status społeczno-ekonomiczny wpływał na moment usamodzielnienia się. Szlachta opuszczała dom rodzinny najpóźniej, a następnie Żydzi, mieszczańcy i chłopi [6](#) [7](#) [8](#).

Źródła zwracają uwagę, że te wskaźniki są ze sobą powiązane. Na przykład, **większe gospodarstwa domowe szlachty były częściowo spowodowane późniejszym wiekiem usamodzielniania się ich dzieci**. Dłuższe trwanie życia szlachty również mogło przyczyniać się do większej liczby osób mieszkających w ich gospodarstwach domowych [9](#).



Zapisz w notatce

# Google NotebookML

- NotebookML może wykonywać dla wybranych notatek działania które sugeruje w formie przycisków na dole interfejsu np. zasugerować nowe powiązane pomysły...

## 3. Zróżnicowanie wewnętrz stanu szlacheckiego:

- Tekst źródłowy zwraca uwagę na **duże zróżnicowanie majątkowe wewnętrz stanu szlacheckiego**, co miało wpływ na standard życia jego członków.
- Obok bogatych magnatów i szlachty posiadającej liczną służbę, istniała **znaczna grupa uboższej szlachty**, której warunki życia niewiele różniły się od mieszkańców czy nawet chłopstwa.
- Warto zbadać ten wątek w kontekście analizowanych wskaźników demograficznych.
- Przykładowo, porównanie długości życia, wielkości gospodarstw domowych i wieku opuszczania domu rodzinnego pomiędzy **ubogą a zamożną szlachtą** pozwoliłoby lepiej zrozumieć **skale zróżnicowania wewnętrz stanu uprzywilejowanego**.



Zapisz w notatce

# Google NotebookML

- ... lub przeprowadzić krytyczną analizę notatek.

Sugestie dotyczące poprawy treści:

- **Znaczenie doboru wskaźników:** Warto rozwinąć ten punkt, wyjaśniając dlaczego metodologia demograficzna jest bardziej miarodajna dla epoki przedprzemysłowej.
- **Ograniczenia analizy:** Można dodać punkt dotyczący potencjalnego błędu w interpretacji danych, np. późne opuszczanie domu rodzinnego mogło wynikać z innych czynników niż dobra sytuacja materialna, np. tradycji lub braku możliwości znalezienia pracy.
- **Wnioski:** Warto dodać krótkie podsumowanie, np. podkreślające wartość analizy demograficznej pomimo jej ograniczeń.

Ogólnie rzecz biorąc, notatka stanowi solidną podstawę do dalszej pracy. Wprowadzenie proponowanych zmian pozwoli na ulepszenie jej stylu i wzmacnienie argumentów.



Zapisz w notatce

Opis i wyjaśnienie

Przeprowadź analizę krytyczną

Sugeruj powiązane pomysły

Utwórz konspekt

# Google NotebookML

- podsumowanie audio

## \* Przewodnik po notatniku

Pomóż mi utworzyć



### Podsumowanie

Artykuł analizuje demograficzne konsekwencje nierówności społecznych w Polsce XVIII wieku. Autorzy badają, jak zróżnicowanie społeczno-ekonomiczne wpływało na takie czynniki, jak długość życia, wielkość gospodarstw domowych oraz wiek opuszczania domu rodzinnego. Wykorzystują dane ze spisów ludności przeprowadzonych w latach 1790–1792 przez komisje porządkowe cywilno-wojskowe. Badania wskazują, że szlachta miała najkorzystniejszą sytuację, odznaczającą się dłuższym życiem, późniejszym usamodzielnianiem się i zatrudnieniem większej liczby służących. Chłopi i mieszkańcy charakteryzowali się gorszymi wskaźnikami demograficznymi, a sytuacja ludności żydowskiej była trudna do jednoznacznej oceny ze względu na problemy z dostępnością danych.

Podsumowanie audio (i)

Generuję rozmowę...  
Może to potrwać kilka minut. Nie musisz czekać.

### Sugerowane pytania

- Jak zróżnicowanie społeczno-ekonomiczne w XVIII-wiecznej Polsce wpływało na standard życia różnych grup społecznych?
- W jaki sposób dane demograficzne z XVIII wieku ukazują relacje między stanem społecznym a długością życia, wielkością gospodarstw domowych i wiekiem opuszczania domu?
- Czy dane demograficzne z XVIII wieku potwierdzają obraz Polski jako "nieba szlachty, raju Żydów, czyśćca mieszkańców i piekła chłopów"?



Link do nagrania

# Gramatyka i styl

- jak Grammarly?
  - DeepL Write
  - QuillBot
  - ChatGPT z modelem GPT-4o:  
prompt:  
„I want you act as a proofreader. I will provide you texts and I would like you to review them for any spelling, grammar, or punctuation errors. Once you have finished reviewing the text, provide me with any necessary corrections or suggestions for improve the text.”
  - choć najnowszych modelach zapewne zadziała prostsze polecenie typu „Review article”...

# Gramatyka i styl

Text: Regardless of whether we're trying to extract knowledge from the GPT model itself, or whether we want the model to extract that knowledge from the context of the query provided, the construction and manner of the question (prompt) is important. A good prompt can significantly improve the accuracy of the model's answer, and after all, accuracy is key in tasks of automatic extraction of information from historical publications. The knowledge extracted in this way is intended to be stored in databases or knowledge graphs, and the model's output can also be used in editions of publications.

## Tekst do sprawdzenia

### Corrections and Suggestions:

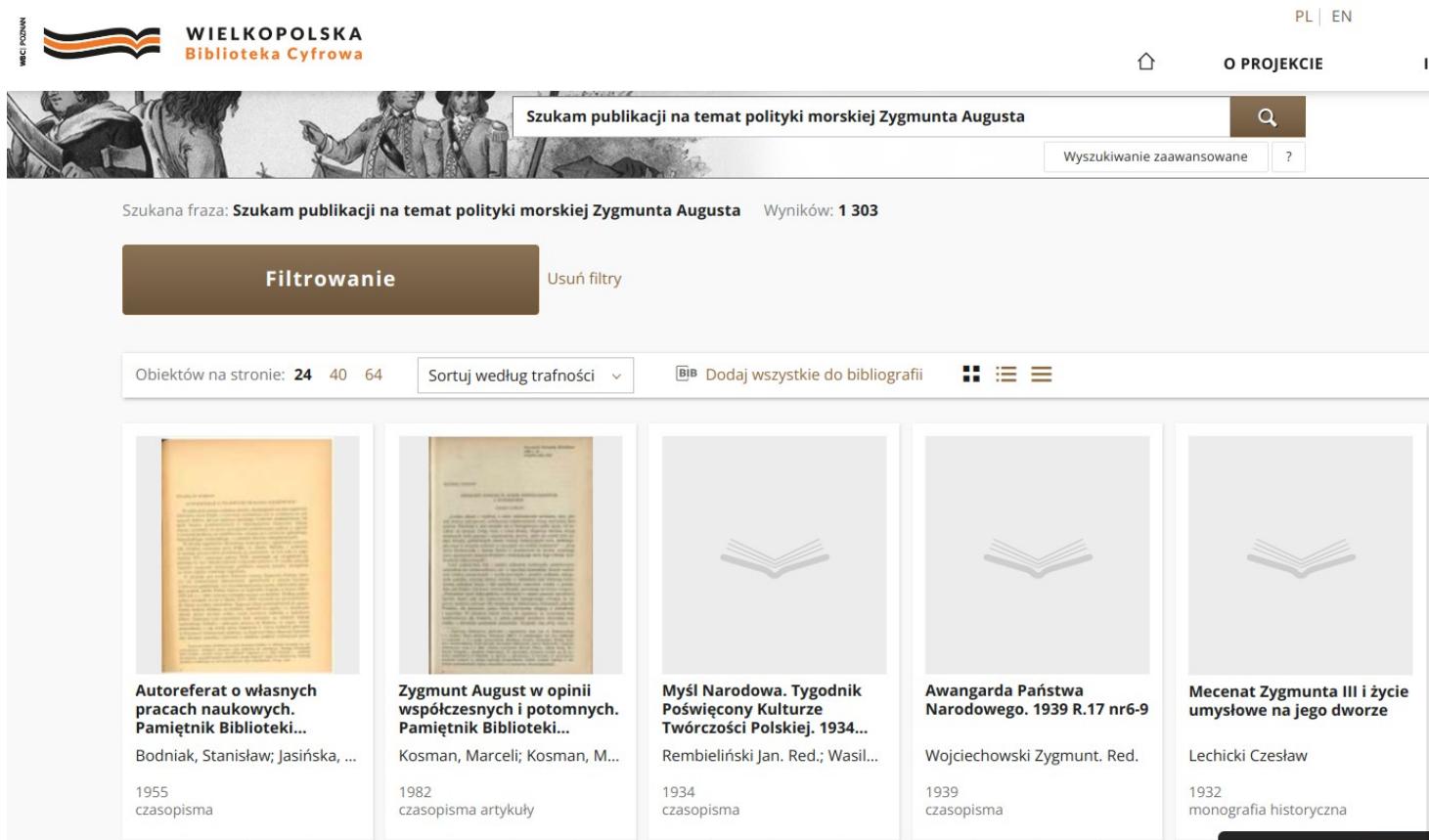
1. "we're" to "we are": Avoid contractions in formal writing.
2. "whether we want the model to extract that knowledge from the context of the query provided" to "or want the model to extract that knowledge from the context of the provided query": Simplified for clarity and flow.
3. "construction and manner of the question (prompt) is important" to "construction and phrasing of the question (prompt) are important": "Manner" is replaced with "phrasing" to improve precision, and "are" is used to ensure subject-verb agreement.
4. "A good prompt" to "A well-crafted prompt": Improved specificity.
5. "after all" separated by commas: Added a comma for correct punctuation.
6. "accuracy is key in tasks of automatic extraction of information from historical publications" to "accuracy is key in tasks involving the automatic extraction of information from historical publications": Replaced "of" with "involving" for more natural phrasing.

## Sugestie GPT



# Wyszukiwanie publikacji

- tradycyjne np.: Wielkopolska Biblioteka Cyfrowa  
[www.wbc.poznan.pl](http://www.wbc.poznan.pl)



The screenshot shows the homepage of the Wielkopolska Biblioteka Cyfrowa (WBC). The top navigation bar includes links for PL | EN, a home icon, O PROJEKCIE (About the Project), and IN. The main search bar contains the query "Szukam publikacji na temat polityki morskiej Zygmunta Augusta" and a search button. Below the search bar, it says "Wyszukiwanie zaawansowane" and "?". The search results page displays 1303 items found. A "Filtrowanie" (Filtering) button is highlighted in brown, and a "Usuń filtry" (Remove filters) button is in white. The results are shown in a grid of five items per row. Each item card includes a thumbnail image, the title, author(s), year, and type of publication.

Obiektów na stronie:	24	40	64
Sortuj według trafności	▼		
BIB	Dodaj wszystkie do bibliografii		
		☰	☰

	<b>Autoreferat o własnych pracach naukowych. Pamiętnik Biblioteki...</b> Bodniak, Stanisław; Jasińska, ... 1955 czasopisma		<b>Zygmunt August w opinii współczesnych i potomnych. Pamiętnik Biblioteki...</b> Kosman, Marcelli; Kosman, M... 1982 czasopisma artykuły		<b>Myśl Narodowa. Tygodnik Poświęcony Kulturze Twórczości Polskiej. 1934...</b> Rembieliński Jan. Red.; Wasil... 1934 czasopisma		<b>Awangarda Państwa Narodowego. 1939 R.17 nr6-9</b> Wojciechowski Zygmunt. Red. 1939 czasopisma		<b>Mecenat Zygmunta III i życie umysłowe na jego dworze</b> Lechicki Czesław 1932 monografia historyczna
---	---	---	--	--	---	---	---	---	---

# Wyszukiwanie publikacji

- [semanicscholar.org](https://semanicscholar.org)

The screenshot shows the Semantic Scholar search interface. At the top, there is a logo and a search bar containing the query "town 16th century europe". Below the search bar, a message indicates "About 12,900 results for 'town 16th century europe'".

Below this, there are several search filters: "Fields of Study", "Date Range", "Has PDF", "Author", and "Journals & Conferences".

The first result listed is an article titled "GLASS BOTTLES FROM THE 16TH CENTURY TO THE 19TH CENTURY IN THE OLD TOWN OF KLAIPĖDA: DATA FROM ARCHAEOLOGICAL EXCAVATIONS" by Indrė Šimkutė. It was published in History on 17 June 2013. The abstract states: "The article analyses material related to glass bottles excavated in Klaipėda (Memel) Old Town. Firstly, this is material from Žvejų St 4 and 6, and the area around Tomo, Didžiojo Vandens, Vežėjų and...". A "Expand" link is present. Below the abstract, there are links for "1" citation, "1" PDF, "Publisher", "Save", and "Cite".

The second result is an article titled "Communication between the Towns and the County Authority: The Free Royal Towns and Šariš County at the Beginning of the 16th Century" by I. Kádas. It was published in History on 2020. The abstract discusses the acquisition of lands by towns in Central Europe. A "Expand" link is present. Below the abstract, there are links for "8" PDF, "Publisher", "Save", and "Cite".

The third result is an article titled "The City Maps of Europe: 16th Century Town Plans from Braun & Hogenberg" by G. Braun, Franz d. Hogenberg, J. Goss, and R. McNally. It was published in History, Geography on 1992. The abstract is not visible. Below the abstract, there are links for "4" citation, "Save", and "Cite".

The fourth result is an article titled "The Town Planning History and Future Prospects of a Unique 16th Century 'idea town' in Central/Eastern Europe" by A. Kelly. It was published in History, Geography on 1999. The abstract is not visible. Below the abstract, there are links for "Save" and "Cite".

# Wyszukiwanie publikacji

- SciSpace - „to najłatwiejszy sposób na znalezienie, zrozumienie i naukę dowolnego artykułu badawczego. Do każdego przeczytanego artykułu uzyskaj proste wyjaśnienia i odpowiedzi od AI i odkryj sieć powiązanych i istotnych artykułów — wszystko w jednym miejscu”

The screenshot shows the SciSpace search interface. The top navigation bar includes a logo, a dropdown menu for 'Papers', a search bar with placeholder 'Discover Papers for your research', a magnifying glass icon, 'Pricing', and a 'Chat with PC' button.

The main search bar contains the query: 'I would like to write an article on 16th century cities in central Europe.' Below it, there are two filter buttons: 'Standard' and 'High Quality'.

The search results section is titled 'Answer from top 5 papers'. It contains three main sections with headings: 'Urban Governance and Documentation', 'Confessionalization and Urban Identity', and 'Economic and Social Changes'. Each section provides a brief summary of the topic and a 'Read More' link.

The sidebar on the left features various icons for different features: Home, Chat, Search (highlighted), Pen, Cross, List, Quotation Marks, and a lightning bolt.

The bottom right corner of the slide displays the page number: / 106.

# Wyszukiwanie publikacji

- consensus.app

The screenshot shows the homepage of **consensus.app**. On the left, there's a sidebar with the **Consensus** logo, a search bar, and a message: "Find the best science faster." Below this are four green checkmark icons: "Academic search engine", "200M+ papers", "AI-powered analysis", and "Answers always cite research". At the bottom of the sidebar are "Sign up" and "Sign in" buttons.

The main area has a search bar containing the query: "I would like to write an article on 16th century cities in central Europe." Below the search bar are three buttons: "Synthesize" (disabled), "Copilot" (selected), "Filter", "Share", and "Export".

The first result is titled "The new burgher revolution in sixteenth- and seventeenth-century partitioned Hungary". It includes a snippet: "The new burgher revolution in sixteenth-century partitioned Hungary challenges this view, revealing that towns in East Central Europe experienced significant urban development and power dynamics during the sixteenth and seventeenth centuries.", author information ("Social History | Balázs A. Szélényi et al. | 3 citations | 2009"), and "Study snapshot" and "Save", "Cite", "Share" buttons.

The second result is titled "Urbanization processes in Podolian Voivodeship (1434-1572) according to the materials of the Crown Metric". It includes a snippet: "Urbanization in Podolian Voivodeship from 1434-1572 led to the creation of 57 cities, but new cities remained mostly villages due to their later accession to the Kingdom of Poland and troubled border.", author information ("Scientific Papers of the Kamianets-Podilskyi National Ivan Ohienko University. History | Mariia Haiseniuk et al. | 0 citations | 2023"), and "Save", "Cite", "Share" buttons.

At the bottom of the page are links for "GPT", "X", and "Help".

# Problemy prawne

- czy użycie AI podczas pracy nad publikacją jest przestępstwem, tak jak plagiat?
- czy wykorzystywanie AI można wykrywać?
- polityka czasopism w kwestii wykorzystywania AI

# Prawo a AI

- według **prawników** i polskiego prawa utwory stworzone przez AI nie mają praw autorskich - maszyny nie są twórcami
- według firmy OpenAI to co stworzy użytkownik podczas pracy z AI należy do **niego** (ale co to znaczy należy?)
- modele open-source trenowane są na tekstach dostępnych publicznie - tak jest w przypadku Bielika 2 (można sprawdzić listę **źródeł**), a modele komercyjne? Co jeśli tekst wytworzony przez model będzie bardzo podobny do tekstu objętego prawami?

# Wykrywanie AI

- jest możliwe, ale nie można na tym polegać
- system antyplagiatowy JSA ma funkcję wykrywania AI, system plagiarism.pl ma komercyjną usługę określającą prawdopodobieństwo, iż dany fragment tekstu został przygotowany przez AI
- powstało wiele systemów komercyjnych i darmowych np. Smodin, Trinka, Writer. Często są bardzo nieskuteczne, np. uznają tekst napisany przez GPT jako w 100% stworzony przez człowieka

# Wykrywanie AI

- niekiedy wykrywanie AI jest zbędne, w mediach opisywano przypadki wydrukowanych publikacji zawierających sformułowania „jako model językowy nie mogę...” itp.
- niezbyt bezpieczne jest także korzystanie z modeli generujących obrazy na podstawie tekstu, takie ilustracje mogą być pozornie ładne, ale mogą też posiadać bezsensowne szczegóły, błędy w napisach itp.

# Polityka czasopism

- Kwartalnik Historii Nauki i Techniki
  - zasady wykorzystania AI dotyczą autorów, redaktorów i recenzentów
  - teksty i grafiki przygotowane z udziałem AI należy wyraźnie opisać, uzasadnić wykorzystanie materiałów stworzonych przez AI
  - redaktorzy nie wykorzystują nadesłanych tekstów w narzędziach AI (chat itp.), z wyjątkiem tych przeznaczonych do wykrywania treści AI
  - recenzenci zobowiązani są do niewykorzystywania AI w procesie recenzji, w razie podejrzeń, weryfikacją zajmuje się redakcja czasopisma

# **4. Diagramy, wykresy, infografiki, analiza danych**

# Infografiki

- Napkin AI → [link](#)

Serwis do wzbogacania tekstu o infografiki, AI czyta tekst i proponuje grafki, które można modyfikować

How to improve the accuracy of  
information extraction from  
historical publications using GPT  
models?

## Improve Prompt Design

Crafting well-structured and  
precise prompts can  
enhance the model's  
understanding and accuracy.



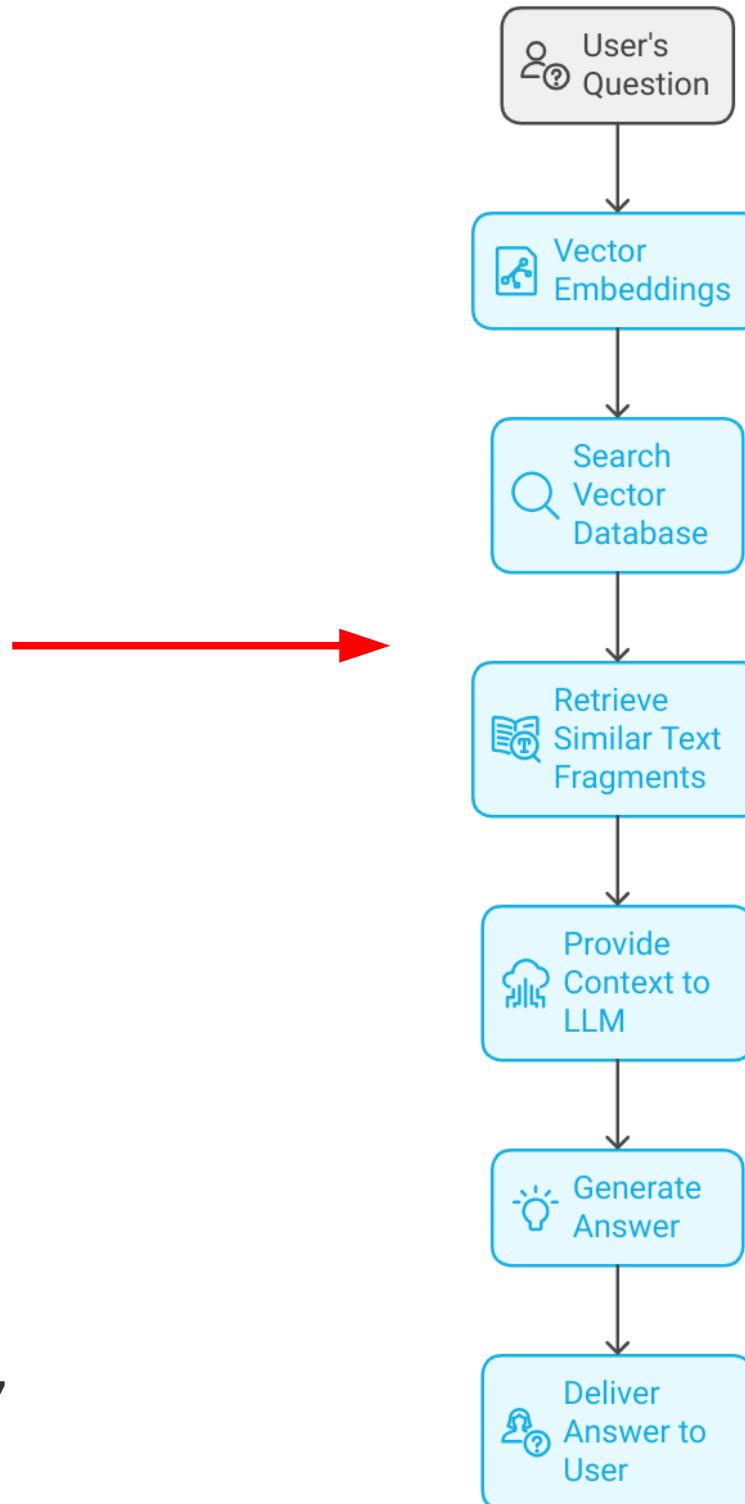
## Contextual Knowledge Extraction

Ensuring the model  
effectively extracts  
knowledge from the provided  
context can improve  
accuracy.

Przykład wizualizacji stworzonej automatycznie na podstawie  
fragmentu tekstu.

# Infografiki

**Tekst:** How does RAG (retrieval-augmented generation) work? -  
The user's question is transformed into a vector form (embeddings), the vector database with the texts of publications, books etc. is searched and a set of text fragments is selected that are most semantically similar to the text of the question. The LLM model receives the text of the question and the context on the basis of which it is to provide an answer (this set of semantically similar fragments of publications), and on this basis the model generates the answer, which is delivered to the user.



# Diagramy w ChatGPT

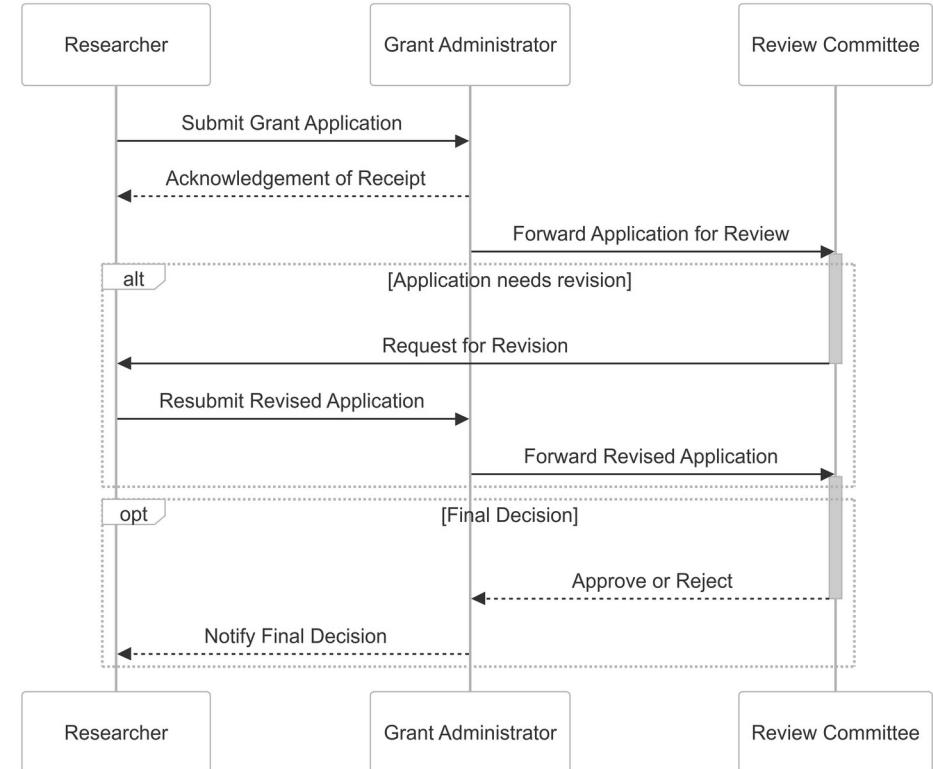
- w ChatGPT można tworzyć dopasowane instancje tzw. GPTs, które najczęściej są wyspecjalizowane w jakiejś dziedzinie np. Mermaid Charts potrafi tworzyć diagramy - na podstawie tekstów
- prompt: „*Prepare an activity sequence diagram for a researcher applying for a grant*” powoduje utworzenie skryptu ...

# Diagramy w ChatGPT

sequenceDiagram

```
participant Researcher
participant GrantAdmin as Grant Administrator
participant ReviewCommittee as Review Committee

Researcher->>GrantAdmin: Submit Grant Application
GrantAdmin-->>Researcher: Acknowledgement of Receipt
GrantAdmin->>ReviewCommittee: Forward Application for Review
activate ReviewCommittee
alt Application needs revision
    ReviewCommittee->>Researcher: Request for Revision
    deactivate ReviewCommittee
    Researcher->>GrantAdmin: Resubmit Revised Application
    GrantAdmin->>ReviewCommittee: Forward Revised Application
    activate ReviewCommittee
end
opt Final Decision
    ReviewCommittee-->>GrantAdmin: Approve or Reject
    deactivate ReviewCommittee
    GrantAdmin-->>Researcher: Notify Final Decision
end
```



- który zostanie automatycznie przetworzony przez serwis mermaidchart.com, ale można go też skopiować i przygotować diagram na stronie mermaid.live.

# Analiza danych - julius.ai

- Julius jest dialogowym systemem wspomagającym użytkownika w analizie danych

Optional: Click on columns or rows to focus on specific data

[Export](#)  [i](#)

	Wiek	Magnateria_Mężczyźni	Magnateria_Kobiety	Szlachta_Mężczyźni	Szlachta_Kobiety	Chłopi_Mężczyźni	Chłopi_Kobiety
0	20	32.8	31.1	43.3	40.8	36	33.4
1	25	29	28.3	38.7	36.9	NULL	NULL
2	30	25.9	26.8	34.3	33.2	26.9	25.1
3	35	22.7	24.1	30	30	NULL	NULL
4	40	19.7	22.7	25.9	26.6	19.1	18.2

The data includes columns for different social groups and their average life expectancy at various ages. If you have any specific analysis or visualization in mind, let me know!

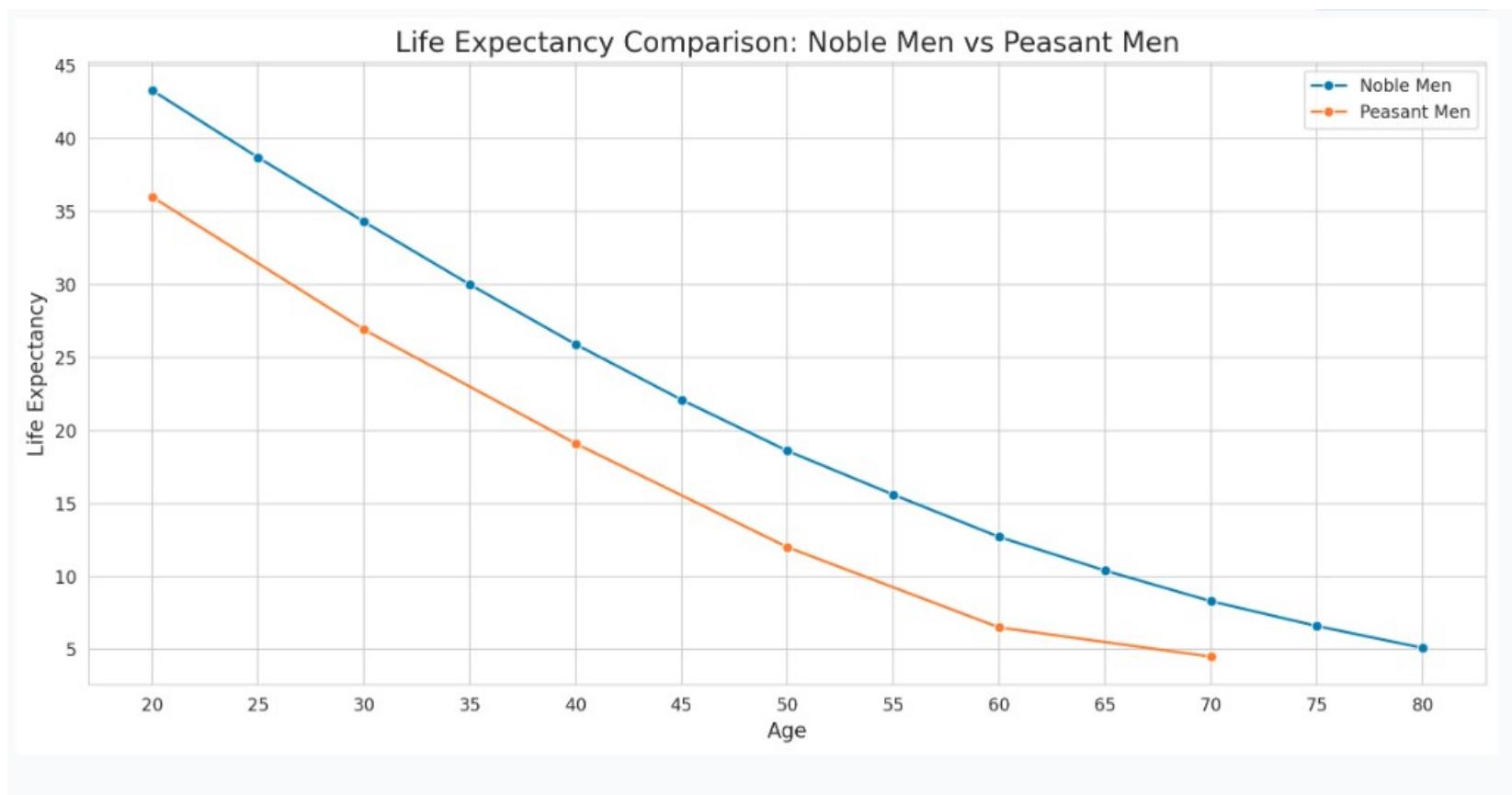
# Analiza danych - julius.ai

- aplikacja wykorzystuje duże modele językowe (GPT-4, Claude-3), podstawowy poziom dostępu (z limitami) jest bezpłatny
- w celu przygotowania analiz czy wykresów, na podstawie poleceń użytkownika, generuje w tle skrypty (python, R), po czym wyświetla wyniki, wizualizacje

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4
5 # Read the CSV file
6 df = pd.read_csv('dane.csv', encoding='utf-8')
7
8 # Select the relevant columns
9 columns_of_interest = ['Wiek', 'Szlachta_M\u0119\u017cczy\u017ani', 'Ch\u0142opi_M\u0119\u017cczy\u017an']
10 df_selected = df[columns_of_interest].dropna()
11
12 # Set up the plot style
13 plt.figure(figsize=(12, 6))
14 sns.set_style("whitegrid")
15
16 # Create the line plot
17 sns.lineplot(data=df_selected, x='Wiek', y='Szlachta_M\u0119\u017cczy\u017ani', label='Noble Men')
18 sns.lineplot(data=df_selected, x='Wiek', y='Ch\u0142opi_M\u0119\u017cczy\u017ani', label='Peasant Men')
19
20 # Customize the plot
21 plt.title('Life Expectancy Comparison: Noble Men vs Peasant Men', fontsize=16)
22 plt.xlabel('Age', fontsize=12)
23 plt.ylabel('Life Expectancy', fontsize=12)
24 plt.legend(fontsize=10)
25
26 # Show the plot
27 plt.tight_layout()
28 plt.show()
29
30 # Print the data used for the plot
31 print(df_selected)
```

# Analiza danych - julius.ai

- Przykład wykresu przygotowanego w aplikacji na podstawie danych z artykułu w Kwartalniku Historycznym



# Analiza danych - julius.ai

- analityk AI wyświetla też podstawowe obserwacje wynikające z danych i wykresu, ale...

From this analysis, we can observe:

1. The dataset includes ages at 5-year intervals, from 20 to 80 years old.
2. There are more data points for Noble Men (Szlachta\_Mężczyźni) than for Peasant Men (Chłopi\_Mężczyźni).
3. The data for Peasant Men is only available for ages 20, 30, 40, 50, 60, and 70.

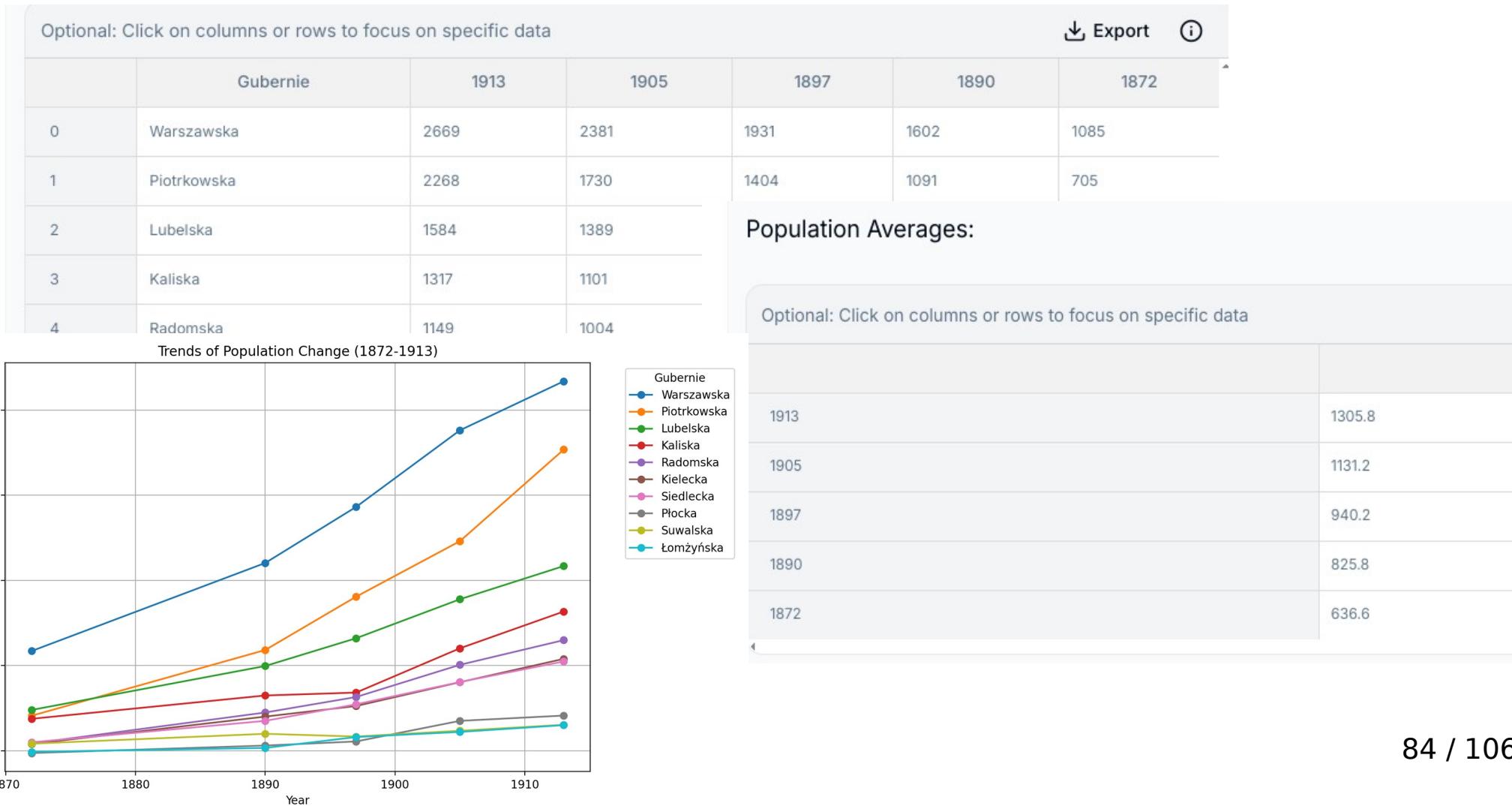
The visualization now shows all available data points, with markers indicating each point. The lines between the points for Peasant Men are interpolated, which might not accurately represent the actual trend between the missing data points.

Key observations:

1. Noble Men consistently have a higher life expectancy across all age groups compared to Peasant Men.
2. The gap in life expectancy seems to widen as age increases, particularly after age 50.
3. The rate of decrease in life expectancy appears to be steeper for Peasant Men compared to Noble Men.

# Analiza danych - julius.ai

- dane o liczbie ludności w poszczególnych guberniach Królestwa Polskiego



# AI i SPARQL

- GPT zna SPARQL
- Prompt „Heja model, przygotuj mi fajne dane” raczej nie zadziała, ale...
- prompt: „Napisz zapytanie SPARQL do bazy wiedzy WikiHum (wikibase), które zwróci wszystkie elementy będące osadami (właściwość instance of (P27) = human settlement (Q175698)) i posiadające w opisie tekst "osada historyczna".” już tak.

# AI i SPARQL

sparql

 Skopiuj kod

```
SELECT ?settlement ?settlementLabel ?description WHERE {
    # Warunek dla osad
    ?settlement wdt:P27 wd:Q175698.

    # Pobieranie opisu w języku polskim
    ?settlement schema:description ?description.
    FILTER(LANG(?description) = "pl")

    # Filtr na opis zawierający frazę "osada historyczna"
    FILTER(CONTAINS(LCASE(?description), "osada historyczna"))

    # Pobieranie etykiety w języku polskim
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "pl,en".
    }
}
```

- kod należy oczywiście obejrzeć, czasem wymaga drobnych zmian, ale w tym przypadku działa

# AI i SPARQL

- dodatkowo otrzymujemy obszerne wyjaśnienia:

## Wyjaśnienie poszczególnych części zapytania:

### 1. Selekcja zmiennych:

```
sparql Skopiuj kod
```

```
SELECT ?settlement ?settlementLabel ?description WHERE {
```

- `?settlement` : URI elementu reprezentującego osadę.
- `?settlementLabel` : Etykieta osady w preferowanym języku.
- `?description` : Opis osady w języku polskim.

### 2. Filtracja po typie obiektu:

```
sparql Skopiuj kod
```

```
?settlement wdt:P27 wd:Q175698.
```

- Ta linia wybiera wszystkie elementy, które mają właściwość `P27` (instance of) równą `Q175698` (human settlement).

# AI i SPARQL

## 3. Pobieranie i filtrowanie opisu:

```
sparql
```

 Skopiuj kod

```
?settlement schema:description ?description.  
FILTER(LANG(?description) = "pl")  
FILTER(CONTAINS(LCASE(?description), "osada historyczna"))
```

- `?settlement schema:description ?description.` : Pobiera opis danego elementu.
- `FILTER(LANG(?description) = "pl")` : Upewnia się, że opis jest w języku polskim.
- `FILTER(CONTAINS(LCASE(?description), "osada historyczna"))` : Sprawdza, czy opis zawiera frazę "osada historyczna", ignorując wielkość liter dzięki funkcji `LCase`.

## 4. Pobieranie etykiet w preferowanych językach:

```
sparql
```

 Skopiuj kod

```
SERVICE wikibase:label {  
    bd:serviceParam wikibase:language "pl,en".  
}
```

- Ta sekcja automatycznie pobiera etykiety dla wszystkich zmiennych zakończonych na `Label` w określonych językach (tu: polskim i angielskim). Jeśli etykieta nie jest dostępna w języku polskim, zostanie pobrana w języku angielskim.

# AI i SPARQL

## Uwagi dodatkowe:

- **Dostosowanie właściwości P27**: Upewnij się, że w Twojej instancji Wikibase właściwość P27 odpowiada za relację "instance of". W standardowej konfiguracji Wikidata właściwość P31 jest używana do określania typu obiektu. Jeśli w Twojej bazie danych P31 odpowiada za "instance of", zmień odpowiednio zapytanie:

```
sparql
```

 Skopiuj kod

```
?settlement wdt:P31 wd:Q175698.
```

- **Wydajność zapytania**: Jeśli baza danych jest duża, dodanie dodatkowych filtrów lub ograniczenie liczby wyników (np. poprzez użycie LIMIT) może przyspieszyć wykonanie zapytania:

```
sparql
```

 Skopiuj kod

```
LIMIT 100
```

# **5. Rozpoznawanie pisma**

# Transkribus

- Supermodele – duże modele HTR oparte na uczeniu głębokim, transformerach – wielojęzyczne, czytające jednocześnie pismo ręczne i drukowane np. Text Titan
- obsługa tabel i formularzy
- publikacja kolekcji transkrypcji - Transkribus Sites

Back Save 0 unsaved changes Test - #2 1 2 4 > In Progress

Region 1

1 f

2 acta puria providus Mathias Czwela Cinis et Brascator Calisse

3 ensis, habens pench so honestam Heinigim Stzuckzina vxore sua

4 legillima, ad infrascripta consenientem, et eid vt marito et futori

5 suo legitimo omniumodam facultate ac pleinpotencia deferente et con-

6 cedentem, nomine opusus Hednigis vxoch sue pfate ibid presentis,

7 resignamt prout putibus resignat in effectu macellung carui in

8 circulo Calissien penes area deserta macell Joannis Taudira ex

9 vna et ariam desertam macelli Nicolai Niesik, Honesto Joanni

10 Badura Lario et Cini Calissien ac eius legitimus succesforibus enn

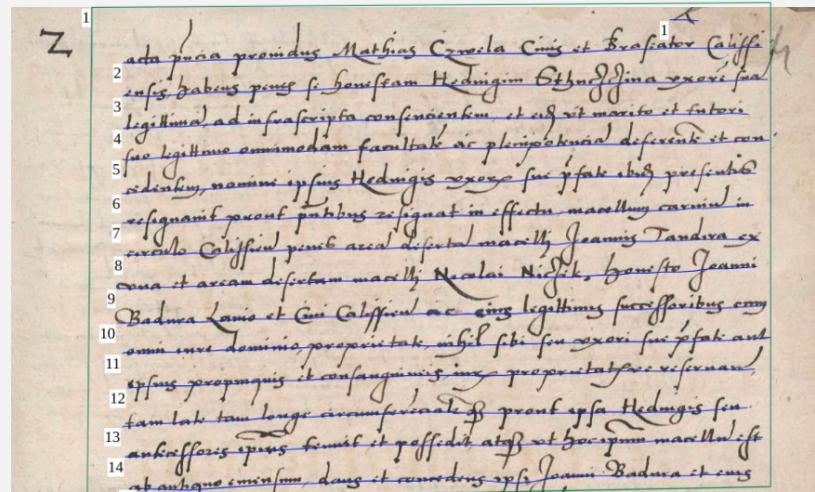
11 omn iure dominio proprietate, nihil sibi seu vxori sue pfate aut

12 epsus propinquus et consanguinus in prout ipsa Hednigis seu

13 tam late tam longe circumfereiale quod prout ipsa Hednigis seu

14 antecessoris ipius tenuit et possidit atque vt hoc ipsum macellu est

15 ib antiquo emensem, dans et concedens ipsi Joami Badura et eius



# eScriptorium

eScriptorium - narzędzie HTR open-source, dostępne na serwerach PCSS

The screenshot shows the eScriptorium interface with a document page on the left and its transcribed text on the right.

**Document Page (Left):**

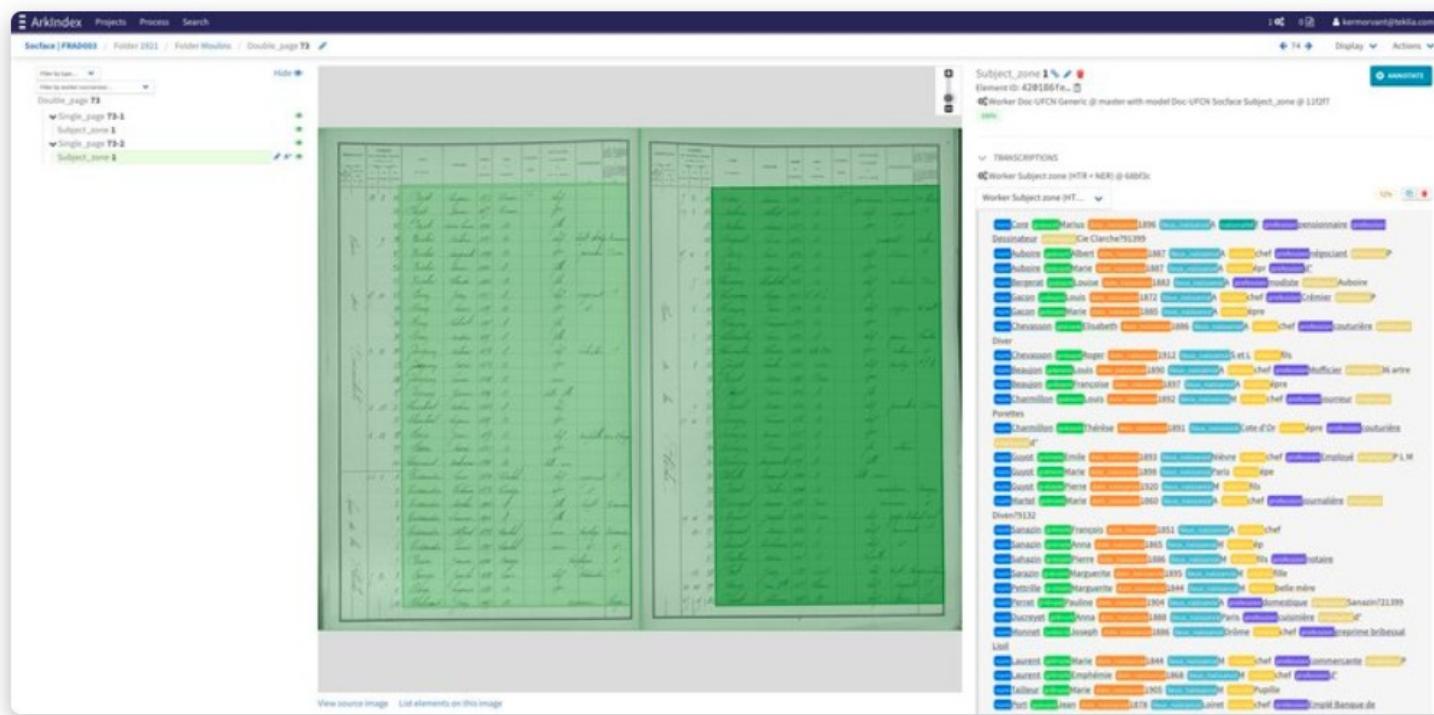
187  
rew, Szlachty, y niezliczonego ludu, ciągnące  
fię przez blisko pułroka, na Senatus Consilium po-  
przedzającym, na publikowaniu Uniwersałów,  
na Seymikach Przed-Seymowych, na Sey-  
mach? na następujących Senatus Konsiliach, na  
relacyjnych Seymikach, zważyć co to zdro-  
wia, co sortuny, co zgryzot każdego kosztu-  
ie? zważyć wiele to millionow na Seymikach  
y Seymie każdym generalnym na to potrzeba  
expensy, y w samey expendue się rzeczy?  
zważyć owe tak straszne zabiegi niepojęte y  
niewyrażone Krola, y naypieiwszych w Oy-  
czynie ludzi, dzień y noc fatyg, zważyć tak  
wielkie na Seymach dla honoru Rzplity y  
swego własnego figury w ekwipażach, Stołach,  
Pałacach, Dworach, zważyć Posłów cudzoziem-  
skich, y wszyskich Cudzoziemców y po-  
stronnych kraiow nad nami attencyą y świa-  
ta całego ciekawość, rzucić oko na ten Ma-  
jestat Pana, Senatu, y Szlacheckiego stanu, ca-  
łą Rzplą reprezentujących, na ow wybor ze  
wszystkich Wojewodztw wybrakowanego lu-  
du, zmierkować nadzieic, wota, oczekiwania  
wszystkich Wojewodztw y Provincji, co tam

**Transcribed Text (Right):**

utrzymymania Seymow.  
187  
rew, Szlachty, y niezliczonego ludu, ciągnące  
fię przez blisko pułroka, na Senatus Consilium po-  
przedzającym, na publikowaniu Uniwersałów,  
na Seymikach Przed-Seymowych, na Sey-  
mach? na następujących Senatus Konsiliach, na  
relacyjnych Seymikach, zważyć co to zdro-  
wia, co sortuny, co zgryzot każdego kosztu-  
ie? zważyć wiele to millionow na Seymikach  
y Seymie każdym generalnym na to potrzeba  
expensy, y w samey expendue się rzeczy?  
zważyć owe tak straszne zabiegi niepojęte y  
niewyrażone Krola, y naypieiwszych w Oy-  
czynie ludzi, dzień y noc fatyg, zważyć tak  
wielkie na Seymach dla honoru Rzplity y  
swego własnego figury w ekwipażach, Stołach,  
Pałacach, Dworach, zważyć Posłów cudzoziem-  
skich, y wszyskich Cudzoziemców y po-  
stronnych kraiow nad nami attencyą y świa-  
ta całego ciekawość, rzucić oko na ten Ma-  
jestat Pana, Senatu, y Szlacheckiego stanu, ca-  
łą Rzplą reprezentujących, na ow wybor ze  
wszystkich Wojewodztw wybrakowanego lu-  
du, zmierkować nadzieic, wota, oczekiwania  
wszystkich Wojewodztw y Provincji, co tam

# Teklia - ArkIndex

- **ArkIndex** - otwarta platforma przetwarzania dokumentów, szczególnie tabel i formularzy/fiszek z pismem ręcznym.  
OCR/HTR/NER



# Large-scale vision-language models

- LVLMs np. [Qwen2-VL](#), chiński model (zespół z firmy Alibaba), potrafi odczytywać tekst drukowany i rękopiśmienny, także po polsku, choć nie bez błędów i halucynacji – fine-tuning?

Qwen2-VL-7B Input

Input Picture

Dowinnością jest moja, donieść Wam iż z dyplomatu Galemberga wyrozumiałem, że życzy sobie, aby handel soli w Polszcze był

Output Text

Prawinnością jest moja, donieść Wam, iż z dyplomatu Galemberga wyrozumiałem, że życzy sobie, aby handel soli w Polszcze był

Qwen2-VL-7B Input

Input Picture

O, w Szwecji.  
Biblioteka kościelna posiada książki pochodzące z kolegiów OO. Jezuitów w Elku i Toruniu, złupione przez Karola Gustawa podczas jego najazdów na Polskę w połowie XVII w.

Bartęg, par. nowozelandzki  
BARDYJOW (Bartfa, Bartfeld), w Czechosłowacji.

Archiwum miejskie posiada liczne dokumenty (akt, kwity, listy królów, dostoyników duchownych i świeckich, ~~mact polskich~~ i t. d.), dotyczące stosunków polsko-węgierskich, zwłaszcza w drugiej poł. XV w. Por.: Papée F., Wiadomość o archiwach węgierskich i materiale ich do dziejów polskich w drugiej połowie XV w. Kraków. 1898; tegoż, studia i szkice 2 czarów Kajmierza Jagiellonczyka. Warszawa, 1907, s. 243/244; Arch. Komisji Historycznej P.A.U., T. VIII, s. 488-440.

Output Text

BARBO, w Szwecji.  
Biblioteka kościelna posiada książki pochodzące z kolegiów OO. Jezuitów w Elku i Toruniu, złupione przez Karola Gustawa podczas jego najazdów na Polskę w połowie XVII w.

BARDOJOW (Bartfa, Bartfeld), w Czechosłowacji.  
Archiwum miejskie posiada liczne dokumenty (akt, kwity, listy królów, dostoyników duchownych i świeckich i t. d.), dotyczące stosunków polsko-węgierskich, zwłaszcza w drugiej poł. XV w. Por.: Papée F., Wiadomość o archiwach węgierskich i materiale ich do dziejów polskich w drugiej połowie XV w. Kraków. 1898; tegoż, studia i szkice 2 czarów Kajmierza Jagiellonczyka. Warszawa, 1907, s. 243-244; Arch. Komisji Historycznej P.A.U., T. VIII, s. 488-440.

# **6. Masowe przetwarzanie tekstów**

# API

- LLM jako uniwersalne narzędzie NLP, przetwarzanie tekstów w ustrukturyzowaną wiedzę
- Api -dostęp do modeli językowych dla programistów
- OpenAI – wysoka jakość, kosztowny dostęp
- Google Gemini – jw., jest dostęp darmowy, ale nie w UE
- Groq Cloud – modele open source, darmowy poziom dostępu z ograniczeniami
- lokalny serwer z API – problem wydajności

# Przykłady

- PSB – ekstrakcja informacji z biogramów PSB, 27.5 tys. biogramów, 150 mb czystego tekstu przetworzone w sumie w ciągu kilku dni (kilkadziesiąt godzin pracy skryptów), wyniki niejednoznaczne, część bardzo dobra.
- Zastawy królewskie – odtwarzanie bazy danych z notatek
- SGKP – ekstrakcja informacji z haseł SGKP

# Przykłady

- SGKP:  
wieś Malejowce

```
"tom": "6",
"strona": 8,
"rodzaj": "wieś",
"warianty_nazw": [
    {
        "nieokr.": "Malijowce"
    },
    {
        "nieokr.": "Milówce"
    }
],
"powiat": "uszycki",
"gmina": "Mukarów",
"parafia": "Mukarów",
"autor": "Dr. M.",
"liczba_mieszkańców": "1100",
"właściciel": "Ignacy Orłowski",
"zabytki": [
    "cerkiew pod wez. ś. Krzyża",
    "pałac doryckiego porządku",
    "ruiny opuszczonego bazylianńskiego klasztoru"
],
"szkoły": [
    "dawna szkoła męska",
    "dawna szkoła żeńska"
],
"przemysłowe": [
    "młyny",
    "kamień ciosowy do budowy",
    "kamień litograficzny",
    "fosforyty"
]
},
```

# Problemy

- koszt: przetwarzanie PSB kosztowało kilka tysięcy złotych, **dziś najlepsze modele są trochę tańsze, coraz większa jest też dostępność dobrych modeli open-source**
- błędy techniczne – obecne modele są coraz lepsze w JSON
- limity – serwery ograniczają ilość przetwarzanych danych
- wiarygodność – 95% poprawności brzmi nieźle, ale nie chcemy przecież 5% błędów w bazie wiedzy?

# **7. Audio AI**

# Transkrypcje Audio->Text

- Model Whisper od OpenAI dostępny jest [online](#)

## Whisper Web

ML-powered speech recognition directly in your browser

The screenshot shows the Whisper Web interface. At the top, there are three input options: "From URL", "From file", and "Record". Below this is a timeline bar showing "0:00 / 7:59". In the center, a blue button says "Transcribe Audio". To the right, a circular icon with a brain-like symbol is labeled "Streszczenie audycji:". Underneath, a summary of the podcast content is displayed:

00:00 Ever feel like you're just drowning in this digital  
00:02 Like it's this tidal wave of information.  
00:04 Well, today's deep dive is going to change how  
00:07 because we're going deep, deep into digital his

Wynik w postaci pliku txt można wczytać w ChatGPT i stworzyć streszczenie:

Podcast omawia artykuł autorstwa Adama Zapały z Polskiej Akademii Nauk, który porusza temat wpływu narzędzi cyfrowych na badania historyczne. Zapała wskazuje na trzy główne zmiany: strukturację, standaryzację oraz formalizację. Strukturacja polega na organizacji danych, co może prowadzić do nowych odkryć. Standaryzacja ułatwia porównywanie dokumentów z różnych okresów i miejsc, dzięki ujednoliceniu języka. Formalizacja obejmuje wykorzystanie algorytmów i sztucznej inteligencji, takich jak rozpoznawanie tekstu ręcznego (HTR) czy przetwarzanie języka naturalnego (NLP), co ułatwia analizę dużych zbiorów danych historycznych.

Podcast porusza również tematykę użycia modeli GPT do analizy i generowania treści historycznych. Choć AI może wspierać badania, pozostaje pytanie, w jakim stopniu może „rozumieć” historię i na ile należy polegać na jej wynikach. Autorzy audycji podkreślają, że cyfrowa humanistyka nie zastępuje historyków, lecz otwiera nowe możliwości w badaniach nad przeszłością.

# Text → audio

- AI może pomagać także w odwrotnym procesie, przekształcania tekstu w nagranie audio. Można nagrać wystąpienie po angielsku z całkiem naturalną wymową lektora AI, lub przetłumaczyć tekst (również przez AI) na portugalski i nagrać nasze wystąpienie po portugalsku → [ElevenLabs](#)

**Abstract:** Large language models are tools with great potential for text processing. The present study aims to assess the reliability of the models' results in extracting structured knowledge from unstructured textual sources, particularly biographies from the Polish Biographical Dictionary (PSB). The task of the model was to extract information about the individuals such as date and place of death, birth and burial, family relationships, important people, related settlements and occupied positions. It was conducted on a sample of 250 biographies. The texts were written in Polish from the 1930s onwards and covered characters from various historical periods. The results show that the LLM is very effective in identifying basic personal data, important family relationships, and finding occupations or offices held by the characters. Weaker results were obtained when attempting to find institutions and places associated with protagonists. It suggests that LLMs can significantly assist in digitising and structuring historical biographical data and offer a promising tool for improving historical knowledge bases and speeding up the work compared to manual information extraction.

[Audio \(ang\)](#)

[Audio \(port\)](#)

# Text → audio

- Można również stworzyć klon własnego głosu i nie znając danego języka obcego wygłosić w nim referat

Modelos de linguagem grande são ferramentas com grande potencial para o processamento de textos. O presente estudo tem como objetivo avaliar a confiabilidade dos resultados dos modelos na extração de conhecimento estruturado a partir de fontes textuais não estruturadas, particularmente biografias do Dicionário biográfico polaco (PSB). A tarefa do modelo foi extrair informações sobre os indivíduos, como data e local de falecimento, nascimento e sepultamento, relações familiares, pessoas importantes, assentamentos relacionados e cargos ocupados. O estudo foi conduzido em uma amostra de 250 biografias. Os textos foram escritos em polonês a partir dos anos 1930 e cobrem personagens de vários períodos históricos. Os resultados mostram que o LLM é muito eficaz em identificar dados pessoais básicos, relações familiares importantes e encontrar ocupações ou cargos ocupados pelos personagens. Resultados mais fracos foram obtidos ao tentar encontrar instituições e locais associados aos protagonistas. Isso sugere que os LLMs podem auxiliar significativamente na digitalização e estruturação de dados biográficos históricos e oferecem uma ferramenta promissora para melhorar as bases de conhecimento histórico e acelerar o trabalho em comparação com a extração manual de informações.

[Audio \(port\)](#)

# Podsumowanie

- AI nie jest chwilową modą, raczej z nami zostanie, choć wiele aplikacji z AI w nazwie zapewne zniknie
- na pewno przetrwają najbardziej pożyteczne zastosowania AI: poprawki tekstu, tłumaczenia, streszczenia, generowanie szkiców i konspektów, ekstrakcja informacji, wyszukiwanie semantyczne, analizy obrazu – w tym HTR

# Podsumowanie

- czy obecny sposób rozwoju: większe modele + więcej materiałów treningowych dotarł do granic możliwości? Zasoby materiałów się kończą...

# Przydatne linki

- [https://github.com/pjaskulski/szkolenie\\_ai](https://github.com/pjaskulski/szkolenie_ai)