Rozdział 8

Badanie popularności i skuteczności lokalnych LLM wśród uczniów i studentów

Zachary Czech, Tomasz Grześ, Karol Przybyszewski*

Streszczenie. W niniejszej pracy zaprezentowano wyniki badania pięciu lokalnych modeli językowych (ang. *local large language models* – LLM): Mistral-7B-Instruct-v0.3, Falcon3-7B-Instruct, Llama-3.1-8B-Instruct, GPT-J 6B i Qwen2.5-7B-Instruct. Celem było sprawdzenie, w jakim stopniu rozwiązania te, uruchamiane na lokalnym sprzęcie wyposażonym w nowoczesne karty graficzne, spełniają oczekiwania użytkowników (uczniów oraz studentów) w przypadku różnych rodzajów zadań: kreatywnych, logicznych i konwersacyjnych. Wyniki pokazują, że każdy z modeli charakteryzuje się specyficznymi mocnymi stronami (np. Falcon w logice, Llama w kreatywności, Qwen w konwersacji), co pozwala lepiej dopasowywać je do konkretnych zastosowań dydaktycznych i badawczych. Różnice w ocenie między uczniami i studentami uwidaczniają większą czułość tych ostatnich na jakość merytoryczną i stylistyczną odpowiedzi. Uzyskane obserwacje mogą stanowić podstawę do dalszych badań nad implementacją modeli językowych w różnych dziedzinach edukacji i nauki.

Słowa kluczowe: local LLM, Mistral, Falcon, Llama, ChatGPT, Qwen

Wprowadzenie

W ostatnich latach gwałtownie wzrosła popularność stosowania metod sztucznej inteligencji w odniesieniu do rozmaitych problemów życia codziennego, a jednym z najistotniejszych obszarów są duże modele językowe (ang. large language lodels, LLM) zdolne do przetwarzania i generowania języka naturalnego na wysokim poziomie. Wraz ze wzrostem dostępnej mocy obliczeniowej możliwe stało się trenowanie modeli o wielkości miliardów parametrów, co przyczyniło się do popularyzacji korzystania z nich w życiu codziennym.

Dostępność wielu modeli w chmurze umożliwiła szerokiemu gronu użytkowników korzystanie z technologii sztucznej inteligencji, jednak niesie to ze sobą problemy

* Wydział Informatyki, Politechnika Białostocka, Wiejska 45A, 15-351 Białystok,
DOI 10.24427/978-83-68673-08-1_8

związane z zapewnieniem odpowiedniego poziomu prywatności czy utratą kontroli nad danymi. W odpowiedzi na te wyzwania coraz większe zainteresowanie wzbudza możliwość uruchamiania LLM lokalnie, na własnym sprzęcie komputerowym. Oferuje ona nie tylko całkowitą prywatność, ale również daje możliwość pełnej kontroli i trenowania tych modeli na dowolnych zbiorach danych, a tym samym dostosowywania ich do własnych potrzeb.

Badania nad zastosowaniami sztucznej inteligencji w obszarach związanych z językiem naturalnym prowadzone są od dawna. Jako jedne z pierwszych pojawiły się systemy konwersacyjne (chatboty), w tym ELIZA Weizenbaum (1966) i A.L.I.C.E. Wallace (2009). Rozwój systemów konwersacyjnych pod koniec pierwszej dekady XXI w. spowodował przejście od systemów przetwarzania języka naturalnego (ang. natural language processing) Young, Hazarika, Poria i Cambria (2018) do jego rozumienia (ang. natural language understanding) i systemów opartych o głębokie uczenie (ang. deep learning), opisanych m.in. w Nivethan i Sankar (2020), Liu, He, Chen i Gao (2019).

W literaturze przedmiotu ukazały się już obszerne opracowania przeglądowe podsumowujące dotychczasowy rozwój modeli LLM oraz stojące przed nimi wyzwania, a także testujące ich techniczne możliwości, w tym m.in. prace Li, Tang, Zhao, Nie i Wen (2022), Aydin, Karaarslan, Erenay i Bacanin (2025) oraz Y. Chen, Xu, Wang, Liu i Mao (2024). Poszczególne modele stanowiły również podstawę badań realizowanych w Mehrabi, Hamou-Lhadj i Moosavi (2024), A. Chen i Tran (2024), Fasha, Hammo, Sowan, Barham i Al-Nsour (2025), Sukherman i Folajimi (2025), czy Lyu i in. (2024).

Wymienione badania ignorowały jednak niejednokrotnie preferencje użytkowników. Celem niniejszej pracy jest przedstawienie wyników badań popularności wybranych lokalnych modeli językowych wśród uczniów i studentów oraz analiza ich skuteczności w przypadku różnych typów zadań. Badania zawierają dokładną ocenę użyteczności generowanych treści pod kątem wielu kategorii. Pozyskane informacje pozwolą zweryfikować, na ile wybrane modele są w stanie zaspokoić oczekiwania użytkowników w kontekstach pracy, nauki lub życia codziennego.

8.1. Materiały i metodologia

W niniejszym rozdziale przedstawiono metodologię realizowanych badań. Szczegółowy opis uwzględnia modele poddane badaniom, środowisko przeprowadzania badań oraz obszary tematyczne użyte do testowania generatywnych modeli językowych.

8.1.1. Badane modele AI

W ramach niniejszego badania przetestowano pięć modeli językowych pochodzących z publicznego repozytorium *Hugging Face*Face (2025):

- **Mistral-7B-Instruct-v0.3**Mistral-7B-Instruct-v0.3 (2025)
  Model opracowany przez francuską firmę Mistral AI. Cechuje się wysoką wydajnością przy relatywnie niskich wymaganiach obliczeniowych, co czyni go szczególnie atrakcyjnym w kontekście uruchamiania lokalnego na komputerach użytkowników indywidualnych. Wybrany został ze względu na dobrą równowagę między jakością generowanych treści a zapotrzebowaniem na zasoby systemu, a także ze względu na dużą popularność.
- **Falcon3-7B-Instruct**Falcon3-7B-Instruct (2025)
  Model opracowany przez Technology Innovation Institute w Zjednoczonych Emiratach Arabskich. Zoptymalizowany pod kątem zadań konwersacyjnych i instruktażowych. Wyróżnia się bardzo dużym kontekstem wejściowym, co przydaje się przy dłuższych wypowiedziach lub analizach. Został wybrany ze względu na szybkość działania, dobrą jakość odpowiedzi w języku angielskim oraz wysoką kompatybilność z popularnymi frameworkami uruchamiania modeli.
- **Llama-3.1-8B-Instruct**Llama-3.1-8B-Instruct (2025)
  Model opracowany przez Meta (osiem miliardów parametrów), stanowiący kontynuację i rozwinięcie wcześniejszych wersji Llama. Jego zaletą jest duża elastyczność w dopasowywaniu do zadań konwersacyjnych, a także poprawione mechanizmy zarządzania kontekstem. Został uwzględniony w badaniu z uwagi na popularność i wsparcie społeczności w kontekście lokalnego uruchamiania.
- **GPT-J 6B**GPT-J-6b (2025)
  Model open-source stworzony przez EleutherAI, bezpośrednio bazujący na modelu GPT-3 stworzonym przez OpenAI. Ma sześć miliardów parametrów i był jednym z pierwszych dużych modeli dostępnych publicznie do uruchamiania lokalnego. Choć obecnie uważany jest za nieco przestarzały, nadal generuje bardzo przyzwoite jakościowo odpowiedzi. Został wybrany do badania jako punkt odniesienia do oceny porównawczej z nowszymi rozwiązaniami.
- **Qwen2.5-7B-Instruct**Qwen2.5-7B-Instruct (2025)
  Model opracowany przez firmę Alibaba. W testach globalnych osiąga dobre wyniki, a przy tym wspiera wielojęzyczne zapytania, co pozwala go przetestować również w języku angielskim, mimo przystosowania do języka chińskiego. Został dołączony do badania jako przykład modelu mniej znanego w Europie i Stanach Zjednoczonych.

Wybór tych konkretnych modeli wynikał zarówno z ich stosunkowo łatwej dostępności w repozytoriach *Hugging Face*, jak i ze zróżnicowanych charakterystyk (parametry, region pochodzenia, potencjalne zastosowania).

8.1.2. Sprzęt i konfiguracja środowiska

Badania zostały przeprowadzone na komputerze wyposażonym w wymienione niżej kluczowe komponenty:

1. Karty graficzne użyte do uruchamiania modeli: 2 $\times$ Palit GeForce RTX 4070 Ti SUPER JetStream OC 16GB GDDR6X.
2. Zasilacz o mocy wystarczającej do obsługi dwóch kart 4070 Ti SUPER: be quiet! Pure Power 12 M 1000W.
3. Płyta główna umożliwiająca montaż dwóch kart graficznych o dużej przepustowości: MSI Z790 GAMING PLUS WIFI.
4. Pozostałe komponenty:
    - pamięć RAM Kingston Fury Beast DDR5 16GB 5200MHz CL40,
    - procesor Intel Core i5-12400F,
    - obudowa Aerocool Shard,
    - dysk SSD Kingston NV2 500GB M.2 2280 PCI-E x4 Gen4 NVMe.

Całość pracowała pod kontrolą systemu Ubuntu 22.04 LTS. Zdalny dostęp do maszyny był możliwy przy użyciu programu AnyDesk. Wszystkie operacje (instalacja modeli, ich uruchamianie) były wykonywane w środowisku wiersza poleceń oraz środowiska PyCharm.

8.1.3. Metodologia testowa

Przed właściwymi badaniami ankietowymi przeprowadzono serię wykładów, podczas których uczestnicy mogli zapoznać się z podstawami działania LLM, ich zastosowaniami i ograniczeniami. Podczas wykładów omówiono m.in.:

- **Definicję i kluczowe cechy LLM** – wyjaśniono, czym są duże modele językowe, w jaki sposób przetwarzają dane tekstowe i jaką rolę odgrywają miliardy parametrów w rozumieniu kontekstu.
- **Architekturę Transformer** – zwrócono szczególną uwagę na *mechanizm uwagi (ang. attention)*, który pozwala modelom na identyfikację zależności między słowami w tekście.
- **Zalety uruchamiania modeli lokalnie** – w tym zachowanie prywatności (dane użytkownika nie opuszczają komputera), pełna kontrola nad modelem (możliwość dostosowania go do własnych potrzeb), niezależność od zewnętrznych usług w chmurze oraz potencjalnie niższe koszty długoterminowe.
- **Przykładowe modele open source** – uczestnikom przedstawiono takie modele, jak Mistral, Falcon, GPT-J, LLaMA, czy Qwen, omawiając ich krótką charakterystykę (liczba parametrów, obszary zastosowań).

- **Dobre praktyki przy tworzeniu promptów** – zaprezentowano, jak formułować pytania i żądania, by uzyskać jasne i zwięzłe odpowiedzi (np. podkreślanie kontekstu, określanie formatu odpowiedzi, rozbijanie złożonych zapytań na kroki).

Cały materiał szkoleniowy był przygotowany tak, aby zapewnić uczestnikom możliwie jak najbardziej praktyczną wiedzę, dzięki której podczas właściwych testów mogli samodzielnie formułować prompty oraz świadomie oceniać jakość wygenerowanych odpowiedzi.

Następnie uczestnicy otrzymali formularz obejmujący 15 pytań podzielonych na trzy kategorie:

- **pytania 1–4 (kreatywne):** pisanie esejów, wierszy, tworzenie scenerii i postaci;
- **pytania 5–11 (logiczne):** myślenie matematyczne, programowanie, instrukcje, sytuacje logiczne;
- **pytania 12–15 (konwersacyjne):** symulacja rozmowy, wspomaganie emocjonalne, porady.

Każde z pytań (prompt) zadano wszystkim pięciu modelom, a uzyskane odpowiedzi zanonimizowano w taki sposób, by głosujący (uczniowie/studenci) nie wiedzieli, który model wygenerował dany tekst. Uczestnicy wskazywali, który z tekstów uznają za najlepszy w danej kategorii (np. najbardziej kreatywny, najlogiczniejszy, najbardziej pomocny). W badaniu wykorzystano przedstawione poniżej prompty:

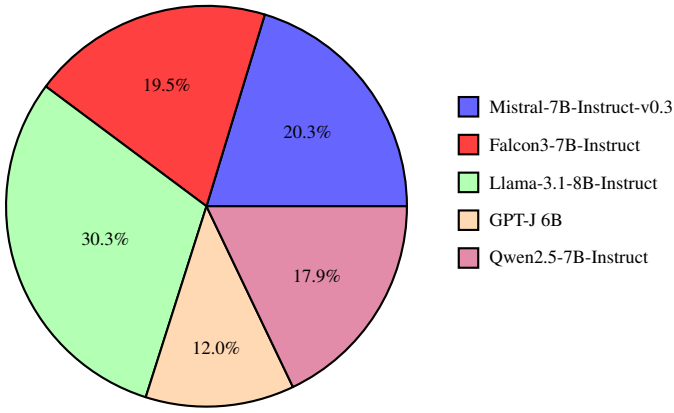
1. Write a four-line poem about hope and resilience, using vivid imagery but avoiding the word ‘hope’.
2. It’s the year 3025. All vehicles are flying cars. Invent a traffic management system for a busy city in the sky.
3. Describe a sunrise using metaphors and comparisons without any color words like red, pink, or orange.
4. Invent a new superhero. Describe their main abilities and moral code.
5. A farmer has 72 apple trees and wants 9 rows with an equal number of trees in each. How many per row, and why?
6. Explain how to find the perimeter and area of a rectangle in a concise, practical way.
7. Explain step by step how to bake a simple loaf of bread in a home oven, but keep it straightforward.
8. Name the seven continents, and briefly mention one distinctive fact about each.
9. Explain the concept of inflation in a simple way, and give one everyday example of rising prices.
10. Write a short Python function that checks if a string is a palindrome (the same forwards and backwards).
11. What is the difference between ‘who’s’ and ‘whose’? Provide two example sentences demonstrating the correct usage.

12. Pretend you're a hiring manager for a software company. I'm interviewing for a Junior Developer position. Ask me two interview questions.
13. I'm feeling a bit lonely today. Can you talk to me as a supportive friend?
14. I have an issue with my roommate leaving the kitchen messy. How can I approach this conversation without causing conflict?
15. I'm stressed about an upcoming exam. Can you give me calming words and advice on how to study effectively?

8.2. Wyniki badań

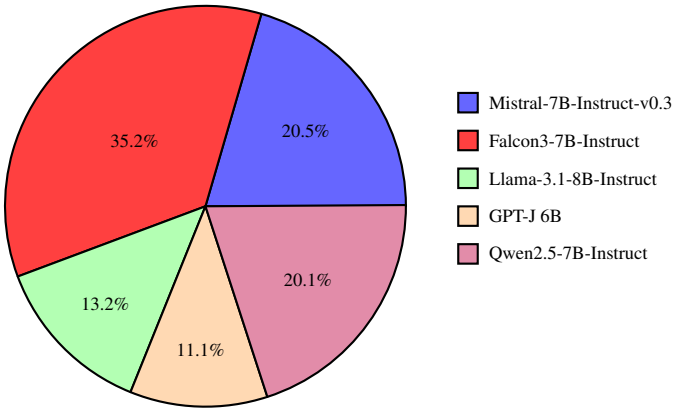
W badaniu przeprowadzono analizę odpowiedzi udzielonych przez 169 uczniów Technikum Programistycznego „Infotech" oraz 38 studentów Wydziału Informatyki Politechniki Białostockiej. Zarówno uczniowie, jak i studenci wskazywali, która z przygotowanych przez LLM odpowiedzi jest najlepsza w ich subiektywnym odczuciu.

Na Rysunkach 8.1, 8.2 i 8.3 przedstawiono rozkład odpowiedzi dotyczących preferencji w grupie badawczej uczniów w zakresie wyboru najlepszego modelu LLM. Wyniki określają odsetek uczniów, którzy uznali odpowiedzi danego modelu LLM za najlepsze w grupie pytań kreatywnych (Rysunek 8.1) logicznych (Rysunek 8.2) oraz konwersacyjnych (Rysunek 8.3).

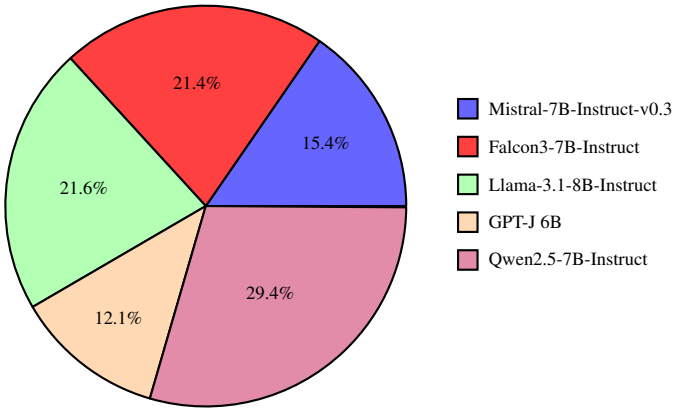


Rysunek 8.1: Rozkład odpowiedzi dotyczących preferencji uczniów dla odpowiedzi na pytania kreatywne (1–4) dla poszczególnych modeli LLM

Na Rysunkach 8.4, 8.5 i 8.6 przedstawiono rozkład odpowiedzi dotyczących preferencji w grupie badawczej studentów w zakresie wyboru najlepszego modelu LLM.



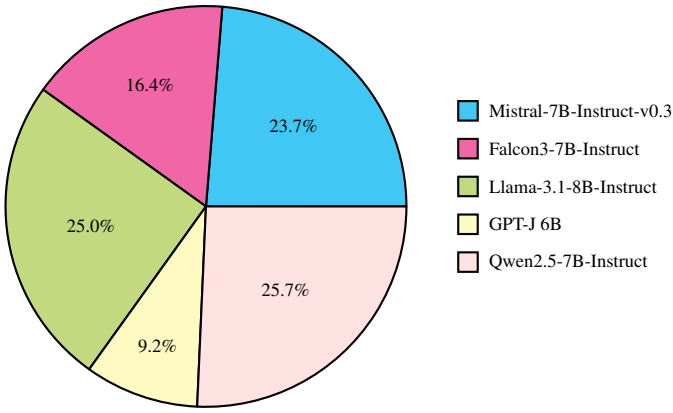
Rysunek 8.2: Rozkład odpowiedzi dotyczących preferencji uczniów dla odpowiedzi na pytania logiczne (5–11) dla poszczególnych modeli LLM



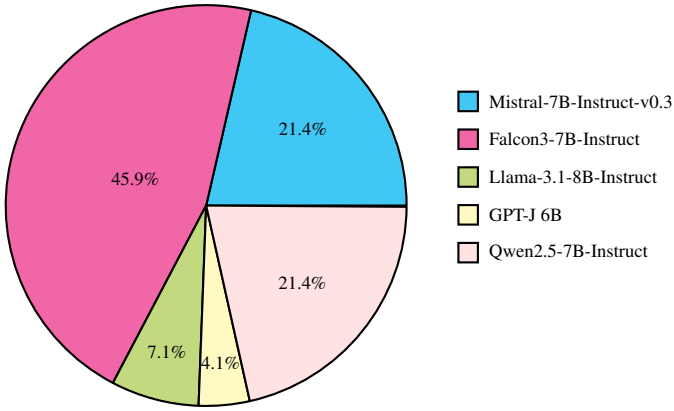
Rysunek 8.3: Rozkład odpowiedzi dotyczących preferencji uczniów dla odpowiedzi na pytania konwersacyjne (12–15) dla poszczególnych modeli LLM

Wyniki określają odsetek uczniów, którzy uznali odpowiedzi danego modelu LLM za najlepsze w grupie pytań kreatywnych (Rysunek 8.4) logicznych (Rysunek 8.5) oraz konwersacyjnych (Rysunek 8.6).

Na Rysunku 8.7 przedstawiono liczbę odpowiedzi wskazujących na model *Mistral-7B-Instruct-v0.3* jako generujący najlepsze odpowiedzi zarówno w grupie uczniów, jak i studentów.



Rysunek 8.4: Rozkład odpowiedzi dotyczących preferencji studentów dla odpowiedzi na pytania kreatywne (1–4) dla poszczególnych modeli LLM

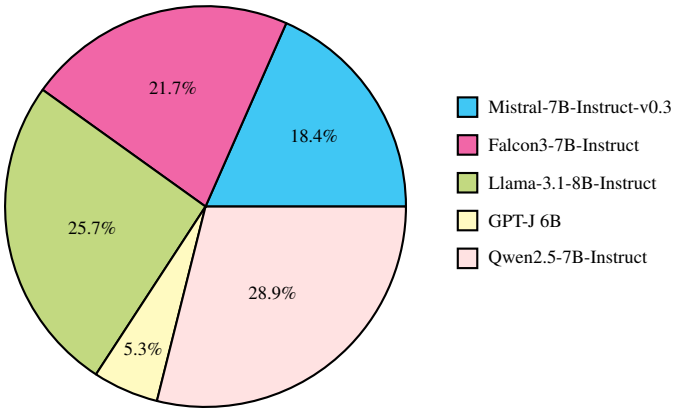


Rysunek 8.5: Rozkład odpowiedzi dotyczących preferencji studentów dla odpowiedzi na pytania logiczne (5–11) dla poszczególnych modeli LLM

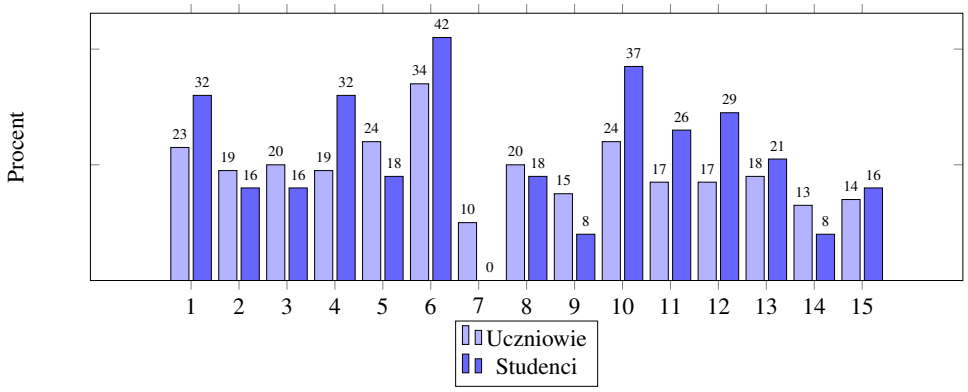
Na Rysunku 8.8 przedstawiono liczbę odpowiedzi wskazujących na model *Falcon3-7B-Instruct* jako generujący najlepsze odpowiedzi zarówno w grupie uczniów, jak i studentów.

Na Rysunku 8.9 przedstawiono liczbę odpowiedzi wskazujących na model *Lamma-3.1-8B-Instruct* jako generujący najlepsze odpowiedzi zarówno w grupie uczniów, jak i studentów.

Na Rysunku 8.10 przedstawiono liczbę odpowiedzi wskazujących na model *GPT-J 6B* jako generujący najlepsze odpowiedzi zarówno w grupie uczniów, jak i studentów.



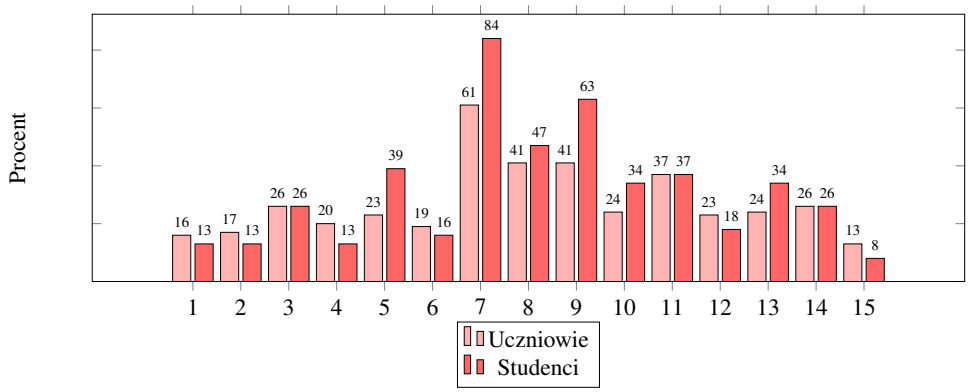
Rysunek 8.6: Rozkład odpowiedzi dotyczących preferencji studentów dla odpowiedzi na pytania konwersacyjne (12–15) dla poszczególnych modeli LLM



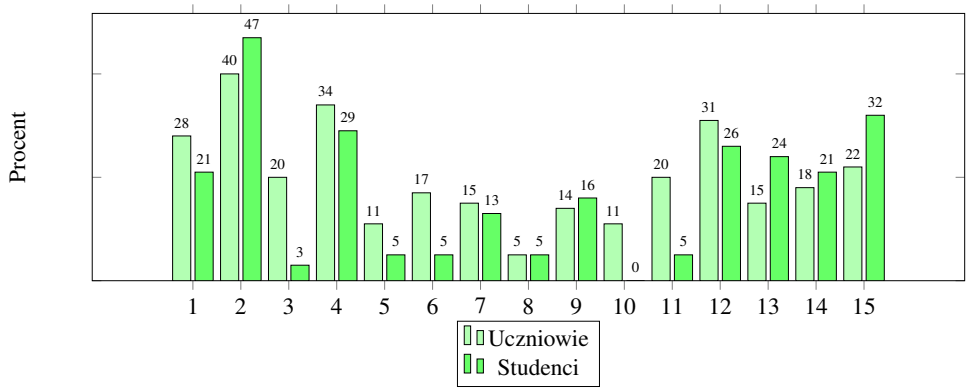
Rysunek 8.7: Odsetki uczniów i studentów wskazujących jako najlepszy model *Mistral-7B-Instruct-v0.3*

Na Rysunku 8.11 przedstawiono liczbę odpowiedzi wskazujących na model *Qwen2.5-8B-Instruct* jako generujący najlepsze odpowiedzi zarówno w grupie uczniów, jak i studentów.

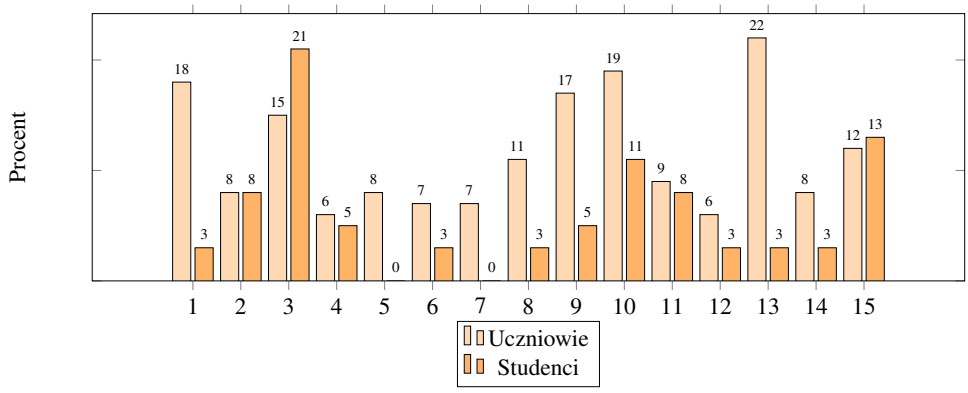
Podczas analizy otrzymanych wyników można zauważyć znaczącą różnicę w ocenie między grupami studentów i uczniów: **studenci** – dzięki większej świadomości i doświadczeniu w analizie tekstu – lepiej rozróżniali poziom odpowiedzi, co przełożyło się na większe różnice punktowe między modelami. Odpowiedzi **uczniów** cechowały się nieco bardziej wyrównaną skalą ocen. Analizę poszczególnych modeli językowych przeprowadzono w kolejnych podrozdziałach.



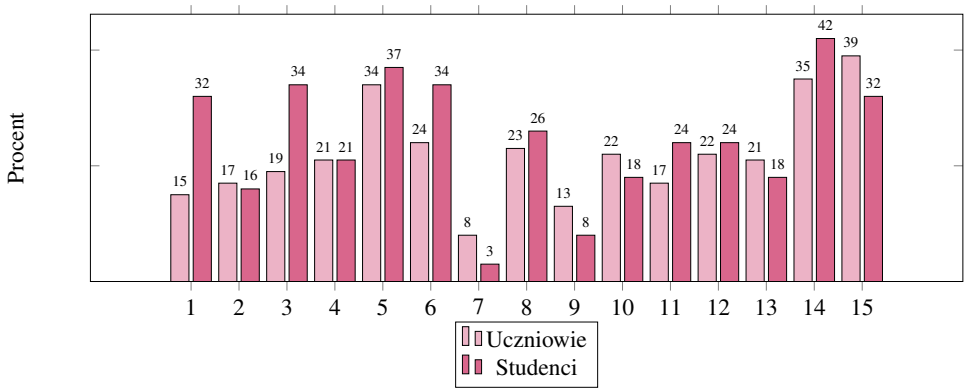
Rysunek 8.8: Odsetki uczniów i studentów wskazujących jako najlepszy model *Falcon3-7B-Instruct*



Rysunek 8.9: Odsetki uczniów i studentów wskazujących jako najlepszy model *Lamma-3.1-8B-Instruct*



Rysunek 8.10: Odsetki uczniów i studentów wskazujących jako najlepszy model *Lamma-3.1-8B-Instruct*



Rysunek 8.11: Odsetki uczniów i studentów wskazujących jako najlepszy model *Qwen2.5-8B-Instruct*

8.3. Dyskusja

Niniejszy rozdział obejmuje dyskusję na temat uzyskanych wyników badań z podziałem na poszczególne modele językowe: Mistral, Falcon3, Llama, GPT oraz Qwen. Każdy z modeli został podsumowany propozycją potencjalnego zastosowania wynikającego z otrzymanych rezultatów.

8.3.1. Mistral-7B-Instruct-v0.3

Analiza wyników wskazuje, że **Mistral-7B-Instruct-v0.3** jest dość wszechstronnym modelem, osiągającym solidne wyniki w **zadaniach logicznych** (5–11) – często plasuje się na drugim miejscu (zaraz za Falconem). Przykładowo w pytaniach szóstym (pole i obwód prostokąta) i dziesiątym (zadanie w Pythonie) osiągnął najwyższe wyniki w obu grupach. Zadania **kreatywne** (1–4) również wypadły dobrze (drugie lub trzecie miejsce), choć ustępował on Llamie i czasem Qwen (zwłaszcza wśród studentów). W kontekście **konwersacji** (12–15) Mistral radził sobie nieco gorzej (czwarte miejsce) – w promptach wymagających wsparcia emocjonalnego (numer 13) czy rozwiązywania konfliktu (numer 14) uczestnicy częściej wybierali inne modele. **Zastosowanie:** Ze względu na **uniwersalny** charakter i dobrą równowagę między logiką a kreatywnością Mistral-7B-Instruct-v0.3 świetnie nadaje się do *ogólnych zastosowań* edukacyjnych i doradczych, zwłaszcza tam, gdzie ważna jest poprawność merytoryczna, ale jednocześnie potrzebne są pewne elementy kreatywności. Nie jest natomiast najlepszym wyborem, gdy priorytetem jest bardzo naturalna rozmowa (chatbot) czy najwyższy poziom twórczej ekspresji.

8.3.2. Falcon3-7B-Instruct

Wyniki jednoznacznie pokazują, że **Falcon3-7B-Instruct dominuje w zadaniach logicznych** (5–11), osiągając pierwsze miejsce zarówno wśród uczniów, jak i studentów. W pytaniu siódmym (instrukcja pieczenia chleba) zdobył wynik 84% wśród studentów i 53% wśród uczniów, średnio natomiast osiągnął 45,9% wśród studentów i 35,2% wśród uczniów w pytaniach logicznych. Z kolei w **kreatywności** (1–4) wypada słabiej (trzecie lub czwarte miejsce), osiągając akceptowalny wynik wyłącznie w pytaniu trzecim (opis wschodu słońca). Również w **konwersacjach** (12–15) Falcon zajmuje środkową pozycję (trzecie miejsce), ustępując Qwen (pierwsze miejsce) i Llamie (drugie miejsce). **Zastosowanie: Logika** to zdecydowanie najmocniejsza strona Falcona – świetnie sprawdza się w zadaniach wymagających wyjaśnienia *krok po kroku*, poprawności rachunkowej i gramatycznej, na przykład w edukacji matematycznej czy technicznej. Nie jest natomiast idealny tam, gdzie liczy się wysoce kreatywne pisanie lub bardzo empatyczna, rozbudowana konwersacja.

8.3.3. Llama-3.1-8B-Instruct

Llama-3.1-8B-Instruct okazał się **najlepszy w zadaniach kreatywnych** (1–4), wygrywając w pytaniach pierwszym, drugim i czwartym wśród uczniów. Studenci również docenili kreatywność Llamy, choć czasem preferowali Qwen. W pytaniach **logicznych** (5–11) Llama radzi sobie zdecydowanie słabiej (czwarte miejsce), co sugeruje, że rozbudowane wyjaśnienia krok po kroku nie są jego najmocniejszą stroną. W **konwersacjach** (12–15) Llama czasem trafia na drugie miejsce, potrafi być dość empatyczny (np. w promptach numer 13 i 15), ale ostatecznie ustępuje Qwen (zwłaszcza wśród studentów). **Zastosowanie: Kreatywność** to domena Llamy – świetnie sprawdzi się przy pisaniu opowiadań, wierszy, scenariuszy czy wszelkiego rodzaju twórczych tekstów. Może też być dobrym wyborem do lekkich konwersacji, choć nie osiąga tu najwyższej oceny. Nie poleca się go jednak w zadaniach, gdzie kluczowe są rygorystyczne rozumowanie logiczne czy analizy matematyczne.

8.3.4. GPT-J 6B

W każdej z trzech kategorii (**kreatywne 1–4**, **logiczne 5–11**, **konwersacyjne 12–15**) **GPT-J 6B** zajmuje ostatnie miejsce. Zarówno uczniowie, jak i studenci rzadko wskazywali jego odpowiedzi jako najlepsze – dotyczy to wierszy (1), obliczeń (5–6), tworzenia kodu w Pythonie (10) czy przyjacielskich porad (13). Mniejsza liczba parametrów oraz starsza architektura sprawiają, że GPT-J generuje poprawne, lecz

mało porywające lub nie zawsze w pełni trafne treści. **Zastosowanie:** Obecnie **GPT-J 6B** trudno polecić do bardziej wymagających zadań. Może ewentualnie odgrywać rolę *punktu odniesienia* lub modelu rezerwowego, gdy ważne są lekka waga i szybkość startu (np. do prototypowania). Nie jest jednak konkurencyjny w porównaniu z nowocześniejszymi rozwiązaniami.

8.3.5. Qwen2.5-7B-Instruct

Qwen2.5-7B-Instruct przoduje w zadaniach konwersacyjnych (12–15) w obu grupach, co oznacza, że potrafi generować najbardziej empatyczne, przyjazne i płynne dialogi. Studenci także docenili kreatywność Qwen (czasem nawet bardziej niż Llamy), natomiast uczniowie w zadaniach 1–4 lokowali go w okolicach trzeciego lub czwartego miejsca. W **logice** (5–11) Qwen radzi sobie przyzwoicie (trzecie miejsce), ale daleko mu do Falcona. Nadal jednak wypada lepiej niż GPT-J czy czasem Llama w typowo definicyjnych promptach (np. *difference between who's and whose* w zadaniu numer 11). **Zastosowanie:** Dzięki wysokiej **jakości dialogowej** Qwen2.5-7B-Instruct nadaje się doskonale do *chatbotów* i wszelkich interaktywnych systemów wsparcia. Z uwagi na nienajgorszą kreatywność może też tworzyć teksty w lekkim, przystępnym stylu. Nie jest jednak najlepszym wyborem do bardzo złożonych wyjaśnień logicznych i matematycznych, gdzie króluje Falcon.

8.4. Wnioski z badania

- **Falcon3-7B-Instruct** bezdyskusyjnie dominuje w zadaniach logicznych, plasując się na pierwszym miejscu w obu badanych grupach. Radzi sobie jednak dużo gorzej z zadaniami kreatywnymi.
- **Llama-3.1-8B-Instruct** radzi sobie najlepiej w pytaniach wymagających kreatywności (szczególnie w oczach uczniów) oraz jest bardzo dobry w konwersacjach, ale wypada słabo w kwestiach logicznych.
- **Qwen2.5-7B-Instruct** okazuje się liderem konwersacyjnym, natomiast wśród studentów często przoduje również w kreatywnych. W logice plasuje się na średnim poziomie.
- **Mistral-7B-Instruct-v0.3** zajmuje wysokie pozycje w zadaniach logicznych (drugie miejsce) i wypada przyzwoicie w kreatywności, natomiast nieco słabiej w konwersacjach.
- **GPT-J 6B** pozostaje najniżej oceniany przez obie grupy w każdym z typów pytań (kreatywne, logiczne, konwersacyjne).

Widać zatem, że każde z badanych rozwiązań wykazuje specjalizację: Falcon preferuje logikę, Qwen konwersacje, a Llama kreatywność. Mistral to solidny „uniwersalny" model, nieco słabszy w dialogu, za to dobry w logice. GPT-J zdecydowanie odbiega jakością od reszty. Co istotne, studenci – mając z reguły większe obycie z tekstami naukowymi czy kreatywnymi – bardziej kategorycznie wskazywali różnice w jakości odpowiedzi. Uczniowie natomiast w wielu zadaniach (zwłaszcza kreatywnych) oceniali modele w sposób bardziej zbliżony do siebie. W konsekwencji okazało się, że studenci stanowią cenniejszą i bardziej miarodajną grupę badawczą – lepiej wyłapują błędy logiczne, usterki stylistyczne czy subtelności formy.

Przeprowadzone badania jednoznacznie pokazują, że żaden z analizowanych modeli nie jest idealny do wszystkich zastosowań. Każdy z nich posiada swoje unikalne mocne strony, które czynią go bardziej odpowiednim do specyficznych zadań.

Podsumowanie

Uzyskane wyniki wyraźnie wskazują na specyficzne mocne strony poszczególnych modeli. **Falcon3-7B-Instruct** zdominował zadania wymagające ścisłego rozumowania i poprawności merytorycznej, co czyni go dobrym kandydatem do zastosowań edukacyjnych w nauczaniu przedmiotów technicznych czy matematycznych. **Llama-3.1-8B-Instruct** z kolei wykazał się bardzo dobrą kreatywnością, co może być szczególnie przydatne przy pisaniu rozbudowanych narracji czy tekstów o charakterze artystycznym. **Qwen2.5-7B-Instruct** wyróżniał się najbardziej naturalnym stylem w pytaniach wymagających empatii i interakcji z użytkownikiem, co czyni go atrakcyjnym wyborem do zadań konwersacyjnych. **Mistral-7B-Instruct-v0.3** osiągał wyrównane wyniki w różnych kategoriach, ale zdaje się być nieco w tyle, gdy chodzi o zdolność tworzenia bardzo empatycznych wypowiedzi. **GPT-J 6B** – najmniej zaawansowany model – wypadł wyraźnie słabiej w każdym rodzaju zadań, co pokazało, że starsza architektura i mniejsza liczba parametrów stanowią zauważalną barierę jakości. Podobne zróżnicowanie kompetencji poszczególnych modeli odnotowano również w innych badaniach porównawczych dotyczących LLM Rangapur i Rangapur (2024) Aydin i in. (2025), gdzie poszczególne systemy wykazywały przewagi tylko w wybranych typach zadań. Różnice między ocenami uczniów i studentów również są widoczne – studenci częściej zwracali uwagę na jakość merytoryczną i styl odpowiedzi, przez co surowiej oceniali modele odstające od ideału. Młodsi uczniowie okazywali się nieco bardziej pobłażliwi dla pewnych uchybień, skupiając się raczej na ogólnym zrozumieniu odpowiedzi niż na szczegółach. Te obserwacje sugerują, że poziom doświadczenia i wykształcenia użytkowników wpływa na percepcję jakości generowanych przez LLM treści.

Wyniki przeprowadzonych testów dają solidną podstawę do zdefiniowania dalszych zadań badawczych oraz praktycznych. Poniżej przedstawiono główne obszary rozwoju:

1. **Ocena w środowisku bardziej zróżnicowanym wiekowo i edukacyjnie.** Dotychczasowe badanie obejmowało głównie uczniów szkół średnich i studentów. W potencjalnych kolejnych etapach byłaby możliwość zaproszenia do testów również osób dorosłych (np. pracujących w zawodach technicznych, humanistycznych) oraz uczniów młodszych klas, aby sprawdzić, jak modele reagują na prompty i potrzeby użytkowników o bardzo zróżnicowanych oczekiwaniach.
2. **Rozszerzenie zestawu testowanych języków.** Wszystkie modele oceniano głównie pod kątem języka angielskiego. W przyszłych badaniach planuje się weryfikację jakości i stylu generowania treści w języku polskim w porównaniu z angielskim, co pozwoli ocenić, na ile modele te nadają się do realizacji zadań w wielojęzycznym środowisku edukacyjnym.
3. **Włączenie do badań innych modeli LLM.** Testowany zestaw obejmował przede wszystkim popularne modele o zróżnicowanych cechach (np. Falcon, Llama, Qwen). Na rynku wciąż pojawiają się jednak nowe architektury i warianty dostrajania. Następne badania włączyłyby kolejne modele LLM – zwłaszcza te wyróżniające się obniżonym zapotrzebowaniem na zasoby obliczeniowe lub unikalnymi cechami (np. wyjątkowa znajomość języka polskiego). Pozwoli to przeprowadzić bardziej kompleksową ocenę i uzupełnić obecny zestaw wniosków.
4. **Badanie innych rodzajów modeli AI.** Kolejne etapy prac mogą objąć nie tylko duże modele językowe, ale również inne typy modeli generatywnych (np. modele obrazowe, multimodalne czy wyspecjalizowane modele dialogowe). Pozwoli to zbadać, w jakich obszarach edukacji i nauki inne technologie AI mogą komplementarnie wspierać proces kształcenia.

Cele i plany na kolejne etapy badań wynikają więc zarówno z praktycznych potrzeb środowisk szkolnych i akademickich, jak i z chęci dalszego rozwoju technologicznego modeli dostępnych lokalnie. Wstępne testy potwierdziły, że nawet stosunkowo niewielkie architektury (w porównaniu z modelami chmurowymi) mogą zapewniać zadowalającą jakość w konkretnie zdefiniowanych zadaniach. Rozbudowa i adaptacja modeli do ściśle określonych obszarów zastosowań mogą dodatkowo podnieść ich skuteczność, tworząc nowe możliwości wykorzystania AI w edukacji, nauce i życiu codziennym.

Adnotacja

Realizacja powyższych badań była możliwa dzięki wsparciu finansowemu dostarczonego przez Politechniczną Sieć VIA CARPATIA w ramach trzeciej edycji projektu Naukolatek. Projekt realizowany był w okresie od października 2024 do marca 2025.

Bibliografia

Aydin, O., Karaarslan, E., Erenay, F. S. i Bacanin, N. (2025). *Generative ai in academic writing: A comparison of deepseek, qwen, chatgpt, gemini, llama, mistral, and gemma.* Retrieved from https://arxiv.org/abs/2503.04765

Chen, A., i Tran, S. (2024). Supercharging document composition with generative ai: A secure, custom retrieval-augmented generation approach. W: *2024 11th ieee swiss conference on data science (sds)* 123–130.

Chen, Y., Xu, B., Wang, Q., Liu, Y. i Mao, Z. (2024, Mar.). Benchmarking large language models on controllable generation under diversified instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17808-17816. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/29734 doi: 10.1609/aaai.v38i16.29734

Face, H. (2025). *Hugging face.* https://huggingface.co/. (Accessed: 2025-05-29)

Falcon3-7B-Instruct. (2025). *Falcon3-7b-instruct.* https://huggingface.co/tiiuae/Falcon3-7B-Instruct. (Accessed: 2025-05-29)

Fasha, M., Hammo, B., Sowan, B., Barham, H. i Al-Nsour, E. (2025). Parameter efficient fine tuning llama 3.1 for answering arabic legal questions: A case study on jordanian laws. W: *2025 1st international conference on computational intelligence approaches and applications (icciaa)* 1–5.

GPT-J-6b. (2025). *Gpt-j-6b.* https://huggingface.co/EleutherAI/gpt-j-6b. (Accessed: 2025-05-29)

Li, J., Tang, T., Zhao, W. X., Nie, J.-Y. i Wen, J.-R. (2022). *Pretrained language models for text generation: A survey.* Retrieved from https://arxiv.org/abs/2201.05273

Liu, X., He, P., Chen, W. i Gao, J. (2019, July). Multi-task deep neural networks for natural language understanding. W: A. Korhonen, D. Traum i L. Màrquez (red.), *Proceedings of the 57th annual meeting of the association for computational linguistics* 4487–4496. Retrieved from https://aclanthology.org/P19-1441/ doi: 10.18653/v1/P19-1441

Llama-3.1-8B-Instruct. (2025). *Llama-3.1-8b-instruct.* https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. (Accessed: 2025-05-

29)

Lyu, C., Yan, L., Xing, R., Li, W., Samih, Y., Ji, T. i Wang, L. (2024). Large language models as code executors: An exploratory study. *arXiv preprint arXiv:2410.06667*.

Mehrabi, M., Hamou-Lhadj, A. i Moosavi, H. (2024). The effectiveness of compact fine-tuned llms in log parsing. W: *2024 ieee international conference on software maintenance and evolution (icsme)* 438–448.

Mistral-7B-Instruct-v0.3. (2025). *Mistral-7b-instruct-v0.3.* https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. (Accessed: 2025-05-29)

Nivethan, i Sankar, S. (2020). Sentiment analysis and deep learning based chatbot for user feedback. W: S. Balaji, Á. Rocha i Y.-N. Chung (red.), *Intelligent communication technologies and virtual mobile networks* 231–237.

Qwen2.5-7B-Instruct. (2025). *Qwen2.5-7b-instruct.* https://huggingface.co/Qwen/Qwen2.5-7B-Instruct. (Accessed: 2025-05-29)

Rangapur, A., i Rangapur, A. (2024). *The battle of llms: A comparative study in conversational qa tasks.* Retrieved from https://arxiv.org/abs/2405.18344

Sukherman, L., i Folajimi, Y. (2025). Enhancing mathematical reasoning in gpt-j through topic-aware prompt engineering. W: *Adjunct proceedings of the 33rd acm conference on user modeling, adaptation and personalization* 388–392.

Wallace, R. S. (2009). The anatomy of a.l.i.c.e. W: R. Epstein, G. Roberts i G. Beber (red.), *Parsing the turing test: Philosophical and methodological issues in the quest for the thinking computer* 181–210. Retrieved from https://doi.org/10.1007/978-1-4020-6710-5_13 doi: 10.1007/978-1-4020-6710-5_13

Weizenbaum, J. (1966, January). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9*(1), 36–45. Retrieved from https://doi.org/10.1145/365153.365168 doi: 10.1145/365153.365168

Young, T., Hazarika, D., Poria, S. i Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, *13*(3), 55-75. doi: 10.1109/MCI.2018.2840738