

PORÓWNANIE SKUTECZNOŚCI KART PRACY GENEROWANYCH PRZEZ AI

Iga Łukiewska*, Julia Muszalska**, Kamil Grzechowski, Marta Burzańska*

*Wydział Matematyki i Informatyki UMK

Fryderyka Chopina 12/18, 87-100 Toruń

** Zespół Szkół w Radzynie Chełmińskim

ul. Ziemowita Maślanki 1, 87-220 Radzyń Chełmiński

iga1812@mat.umk.pl, juliamuszalska14@gmail.com,

kamil.grzechowski@gmail.com, quintria@mat.umk.pl

Abstract. *This study focuses on selecting the best, free AI tool for teachers to use, when preparing reading comprehension worksheets for primary education. A controlled study was conducted with approximately 50 students aged 10-13. The research compared AI-generated materials with a human control sample through standardized tests. The results shows that AI-Content generated by Gemini, Copilot and ChatGPT 5.5 can match traditional quality but still requires professional adjustments*

1. Wstęp

W ostatnich latach duże modele językowe (LLM) stały się nieodłącznym narzędziem pracy ekspertów reprezentujących liczne dyscypliny naukowe. W kontekście szkolnictwa podstawowego nauczyciele nieustannie mierzą się z wysokim poziomem stresu, nadmiarem obowiązków oraz wypaleniem zawodowym. Jak wykazano w wielu badaniach ([1,2,3,4,5]) materiały przygotowane przez modele AI – choć starsze od analizowanych w niniejszej pracy - są w stanie osiągać wyniki porównywalne z zasobami przygotowanymi przez studentów, mimo że charakteryzują się większą schematycznością. Sugeruje to, że również nauczyciele mogliby ograniczyć czas oraz energię na przygotowanie materiałów dydaktycznych, delegując część pracy do asystenta AI. Istotne wyzwanie stanowi jednak specyfika języka polskiego, gdyż modele te są trenowane głównie na danych anglojęzycznych. W naszym badaniu podjęto próbę sprawdzenia, jak różne modele AI radzą sobie z przygotowaniem narzędzi do pracy z tekstem kultury oraz w jakim stopniu wspierają one uczniów w interpretacji i zrozumieniu czytanych treści. W pracy Silor i in. (2025) [6] wykazano, że interaktywne narzędzia oparte o technologię AI mają pozytywny wpływ na czytanie ze zrozumieniem wśród słabszych uczniów,

lecz mogą oddziaływać negatywnie na uczniów bardziej zaawansowanych. Chen (2024) w swojej pracy [7] zauważył, że Microsoft Copilot i ChatGPT są bardzo skuteczne w poszerzaniu słownictwa, wykonywaniu ćwiczeń przygotowujących do czytania, rozpoznawaniu głównych myśli, sporządzaniu streszczeń oraz generowaniu pytań, a także sprawdzają się w spersonalizowanym nauczaniu, rozwijaniu krytycznego myślenia i aktywnym uczeniu się, wspierając jednocześnie proces nauczania i uczenia się. Natomiast w badaniach Emis i in. (2024) [8] dwa narzędzia AI: Gemini i Microsoft Copilot, generowały bardziej zrozumiałe odpowiedzi niż inne narzędzia, gdy były wykorzystywane do uzyskiwania odpowiedzi.

Powyższe badania zdają się sugerować, że AI stanowi istotną pomoc dydaktyczną, o ile jest stosowana w odpowiednich grupach. Wciąż jednak pozostaje otwarte pytanie: które ze współczesnych, ogólnodostępnych narzędzi jest najbardziej efektywne w polskich realiach szkolnych oraz w jakim stopniu może ono realnie odciążyć nauczyciela? Ten problem był już adresowany w literaturze naukowej ([9,10,11]), jednakże w dobie ciągłego rozwoju narzędzi, pozostaje stale aktualny. W miarę rozwoju narzędzi, rozwijają się też kompetencje nauczycieli w zakresie generowania materiałów dydaktycznych, ale i co raz więcej uczniów ma kontakt z treściami generowanymi i zwiększa swoje wyczerpanie na ten typ materiałów.

Mają na uwadze powyższy kontekst, w niniejszym badaniu podjęto próbę porównania trzech popularnych modeli językowych pod kątem ich przydatności w przygotowaniu materiałów do analizy tekstu literackiego. Przyjęta procedura badawcza miała na celu nie tylko ocenę jakości generowanych zadań, ale także weryfikację ich poprawności językowej oraz praktycznej użyteczności w grupie wiekowej 10–13 lat.

2. Metodologia badania

W celu rzetelnej oceny przydatności modeli językowych w procesie dydaktycznym, zaprojektowano badanie komparatywne (porównawcze). Poniżej przedstawiono szczegółowy opis grupy badawczej, wykorzystanych narzędzi oraz procedury eksperymentalnej.

Grupa badawcza i materiał źródłowy W badaniu wzięło ponad 60 uczniów w wieku 10–13 lat, uczęszczających do klas 4, 6 i 7 szkoły podstawowej, oraz czterech wykwalifikowanych nauczycieli języka polskiego. Materiał do pracy z tekstem został zróżnicowany pod względem poziomu trudności i tematyki:

- **Klasa 4:** „Legenda o Królu Popielu” (tekst o charakterze baśniowym),
- **Klasa 6:** „Mit o królu Midasie” (literatura klasyczna/mitologia),
- **Klasa 7:** „Trzej Elektrycyerze” ze zbioru „Bajki Robotów” S. Lema (tekst o wyższym stopniu abstrakcji).

Narzędzia i promptowanie

Do wygenerowania materiałów dydaktycznych wykorzystano trzy darmowe modele LLM: ChatGPT 5.5, Gemini 3.1 Pro oraz Claude Sonnet 4.6. W celu zapewnienia porównywalności wyników, do wszystkich modeli skierowano identyczny prompt:

„Na bazie przedstawionego dokumentu wygeneruj zestaw zadań tak, by ułatwić dziecku w klasie (4, 6, 7) zrozumienie przeczytanej lektury. Zestaw zadań powinien zająć 15 minut. Twoim celem jest rozwinięcie umiejętności czytania ze zrozumieniem oraz zapewnienie, że dziecko zrozumiało przeczytaną treść. Odpowiedź zwróć w formie tekstu, tak by mogła być przekopiowana”.

Jako grupę kontrolną wykorzystano kartę pracy przygotowaną ręcznie przez nauczyciela (bez znajomości treści testu końcowego), co pozwoliło na zestawienie wyników AI z tradycyjnym warsztatem pedagogicznym.

Procedura badawcza

Badanie zostało przeprowadzone w dwóch etapach:

1. **Etap uczniowski:** Uczniowie zostali podzieleni na cztery grupy. Każda otrzymała ten sam tekst źródłowy, ale inną kartę pracy (czerwona - Gemini, niebieska - Cloud, żółta – Chat GPT, zielona - nauczyciel). Uczniowie wiedzieli, że wśród kart, znajdując się ćwiczenia wygenerowane przez AI, jak i ćwiczenia stworzone przez nauczyciela. Na zapoznanie się z tekstem i rozwiązanie zadań przewidziano 15 minut. Bezpośrednio po tym etapie uczniowie wypełniali ujednolicony, autorski test wiedzy oraz krótką ankietę oceniającą ich subiektywne odczucia co do przydatności materiałów.
2. **Etap ekspercki (Ślepa próba):** Wszystkie cztery zestawy zadań zostały przedstawione czterem polonistom. Nauczyciele nie znali autorstwa poszczególnych kart pracy. Ich zadaniem była ocena wartości merytorycznej, poprawności językowej oraz użyteczności dydaktycznej każdego materiału.

3. Wyniki badań

Opinie uczniów

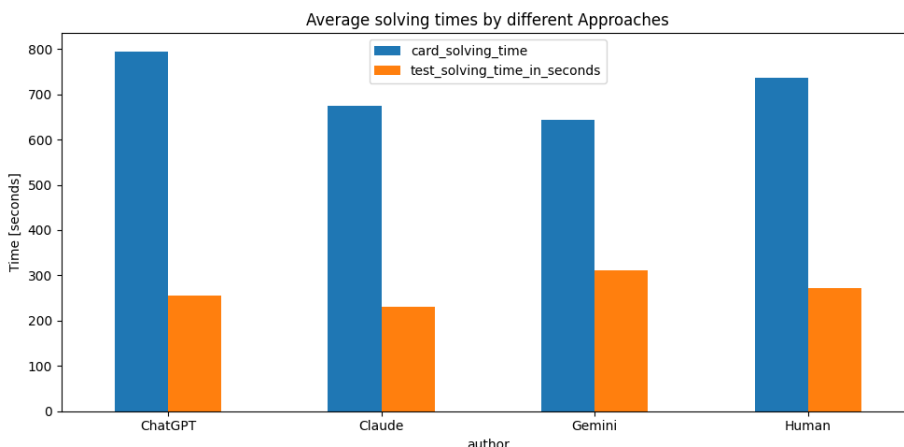
Interesującym i nieplanowanym wcześniej aspektem badania była spontaniczna reakcja uczniów na błędy językowe zawarte w materiałach generowanych przez AI (Gemini). Choć badanie zakładało anonimowość źródeł, dzieci wykazały się nadzwyczajną czujnością językową, która doprowadziła do demaskacji technologii.

W bezpośrednich uwagach kierowanych do nauczyciela, uczniowie wskazywali na konkretne błędy fleksyjne (odmianę wyrazów), argumentując: „**To musi być AI, pani**

by nigdy nie popełniła takiego błędu”. Ten nieprzewidziany element badania dowodzi, że autorytet językowy nauczyciela stanowi dla uczniów główny punkt odniesienia, a błąd językowy jest traktowany jako „cyfrowy ślad” zdradzający nie ludzkie pochodzenie tekstu.

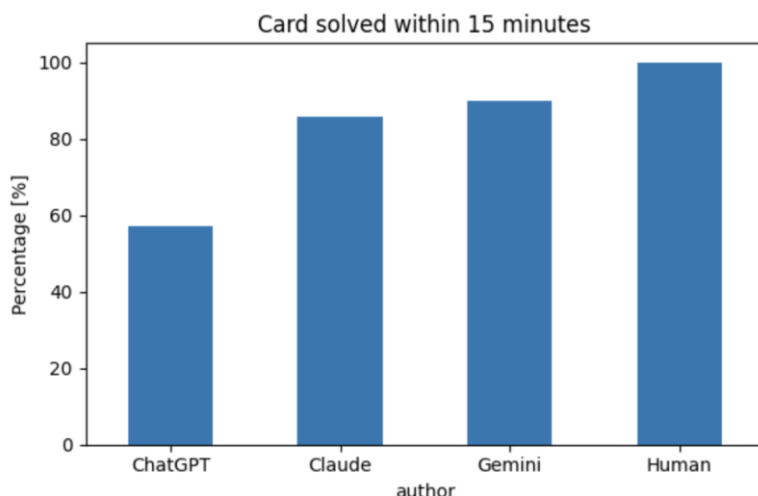
Równie intrygującym zjawiskiem, zaobserwowanym podczas badania, była odwrotna demaskacja źródeł, której dokonali niektórzy uczniowie. Część badanych błędnie zidentyfikowała kartę pracy przygotowaną przez nauczyciela jako wytwór sztucznej inteligencji. Swoją opinię argumentowali oni faktem, iż zadania te były „zbyt trudne” lub wymagały „zbyt głębokiej analizy”.

Czas poświęcony na karty pracy oraz testy



Rysunek 1. Porównanie średniego czasu rozwiązania karty pracy oraz testu wiedzy przez wybrane modele AI oraz człowieka

Pod względem tempa pracy najwyższą średnią odnotowano w grupach korzystających z modelu ChatGPT; czas wykonania zadań był w tym przypadku o 59 sekund dłuższy niż w grupie kontrolnej pracującej z materiałem autorskim nauczyciela. Jednocześnie uczniowie z grupy ChatGPT wykazywali większą sprawność w fazie sprawdzającej, potrzebując mniej czasu na rozwiązanie końcowego testu wiedzy.

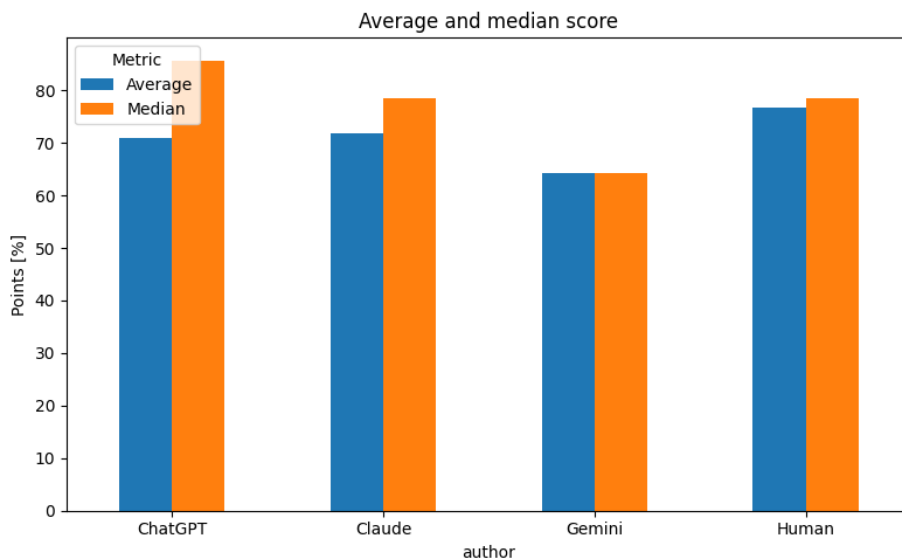


Rysunek 2. Wykres przedstawiający jaki procent uczniów, był w stanie rozwiązać kartę pracy w wyznaczonym przedziale czasowym

Analiza danych wykazała znaczące różnice w czasochłonności przygotowanych materiałów. Model Gemini wygenerował zestaw zadań o najniższym stopniu złożoności czasowej, podczas gdy materiały przygotowane przez ChatGPT okazały się najbardziej czasochłonne. W przypadku tego ostatniego modelu ponad 40% badanych nie zdołało ukończyć karty pracy w wyznaczonym limicie czasowym, co wskazuje na tendencję algorytmu do niedoszacowania trudności zadań lub ich nadmiarowości.

Wyniki jednolitego testu wiedzy

Analiza średnich wyników poprawności wykazała, że najwyższą skuteczność osiągnęła grupa kontrolna pracująca z materiałem autorskim nauczyciela (**76%**). Na kolejnych pozycjach uplasowały się modele Claude (**71,8%**) oraz ChatGPT (**71%**), natomiast najniższy wynik odnotowano w grupie Gemini (**64,3%**). Ze względu na wystąpienie wartości skrajnych (część uczniów nie podjęła próby rozwiązania testu, oddając puste karty), dokonano również analizy mediany wyników. W tym zestawieniu najwyższą efektywność wykazał ChatGPT (**85,7%**), wyprzedzając zarówno grupę kontrolną, jak i model Claude (oba wyniki na poziomie **78,5%**). Najsłabszy rezultat, zbliżony ze średnią, ponownie uzyskał model Gemini (**64,3%**).



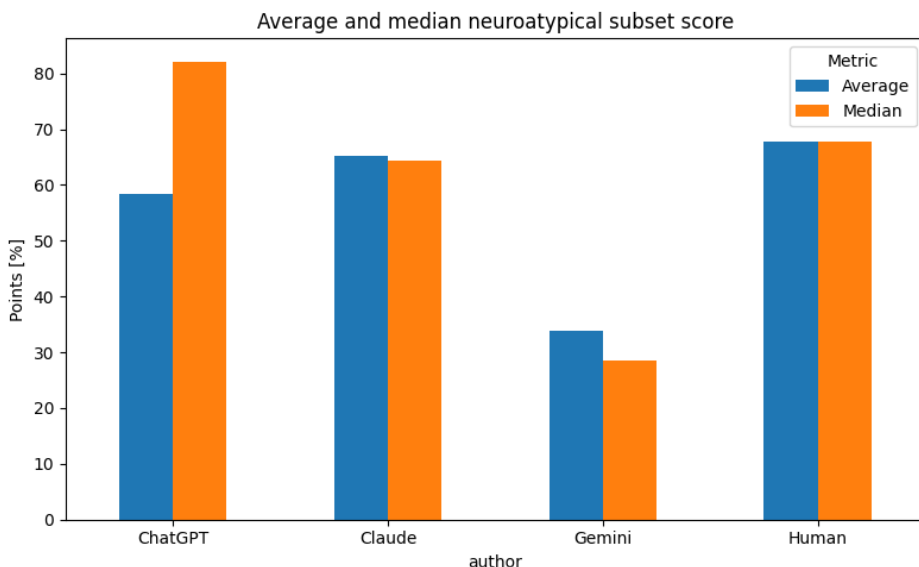
Rysunek 3. Wykres przedstawiający średnią wartość oraz medianę uzyskanych wyników z jednolitego testu wiedzy dla pełnej próbki badawczej

Wyniki jednolitego testu wiedzy w podzbiorze uczniów SPE

W celu sprawdzenia, czy struktura materiałów dydaktycznych wpływa w odmienny sposób na uczniów o specyficznych potrzebach poznawczych, wyodrębniono podpróbę dzieci neuroatypowych ($N = 18$), m.in. ze spektrum autyzmu oraz niepełnosprawnością intelektualną w stopniu lekkim). Z uwagi na małą liczebność tej grupy, analizę tę należy traktować jako przyczynkową, jednak zestawienie jej z wynikami ogólnymi pozwala na wyciągnięcie interesujących wniosków.

Co kluczowe, ogólna hierarchia efektywności poszczególnych metod okazała się niemal tożsama w obu badanych grupach. Zarówno w ujęciu ogólnym, jak i w podgrupie SPE, najwyższą średnią liczbę punktów z testu wiedzy zapewniły materiały przygotowane przez nauczyciela (Human – odpowiednio 76.67% oraz 67.86%), a tuż za nimi uplasował się model Claude.

Główna różnica polega na pogłębieniu dysproporcji i ogólnym obniżeniu pułapu punktowego u dzieci ze SPE. Szczególnie jaskrawo widać to na przykładzie modelu Gemini, który już w grupie ogólnej radził sobie najsłabiej (64.29%), natomiast w grupie SPE doprowadził do drastycznego spadku efektywności – średni wynik wyniósł tam zaledwie 33.93%, a mediana spadła do 28.57%.

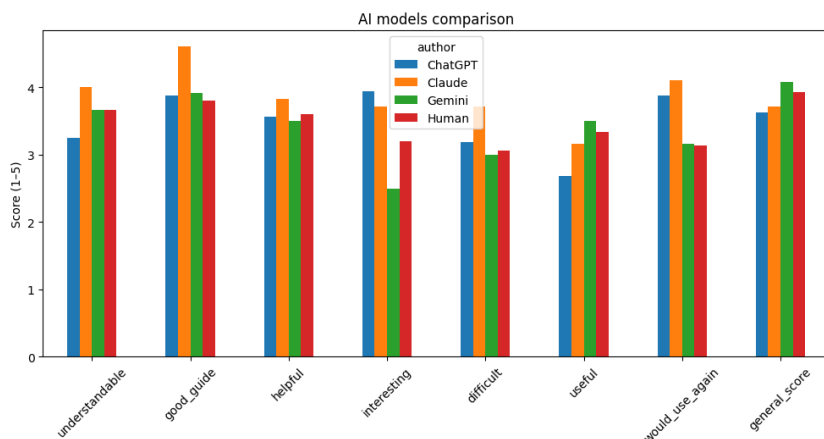


Rysunek 4. Wykres przedstawiający średnią wartość oraz medianę uzyskanych wyników z jednolitego testu wiedzy dla wyodrębnionej próbki uczniów neuroatypowych

Niezwykle ciekawie prezentują się dane dotyczące modelu ChatGPT. Choć jego średni wynik w obu grupach ustępuje wynikom nauczyciela, to wskaźnik mediany okazał się najwyższy w całym zestawieniu – zarówno dla ogółu (85.71%), jak i dla dzieci neuroatypowych (82.14%). Sugeruje to, że materiały z ChatGPT generowały polaryzację: dla dużej części uczniów (w tym SPE) były one wysoce efektywne, mimo że pojedyncze, znacznie słabsze wyniki zaniżyły ogólną średnią tego modelu.

Wyniki ankiety przeprowadzonej wśród uczniów

Pod względem zrozumiałości treści (*understandable*) oraz prowadzenia ucznia „krok za krokiem” (*good_guide*), najwyższe noty od uczniów zebrał model Claude. Co ciekawe, materiały przygotowane przez tego asystenta AI zostały uznane przez dzieci za najłatwiejsze (wartość *difficult* występuje w odwróconej skali, gdzie 1 oznacza zbyt trudne, a 5 zbyt łatwe), a uczniowie zadeklarowali również największą chęć ponownej współpracy z tym modelem (*would_use_again*).



Rysunek 5. Wyniki ankiety dotyczącej użyteczności kart pracy wypełnionych przez badanych uczniów

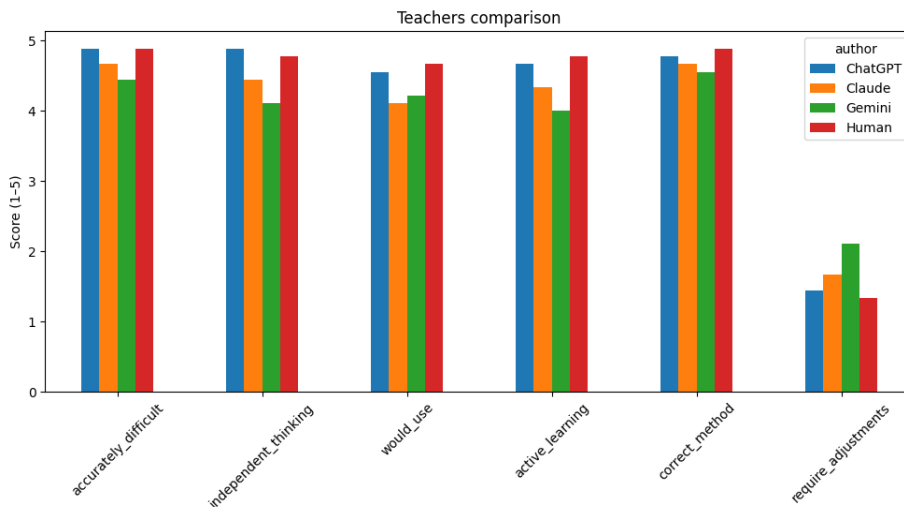
Z perspektywy uczniów bardzo interesująco wypada Gemini. Choć model ten został oceniony jako najmniej interesujący (*interesting* = 2,50) i najtrudniejszy (*difficult* = 3,00), to paradoksalnie zdobył najwyższą ocenę ogólną (*general_score* = 4,08) oraz najwyższy wskaźnik użyteczności (*useful* = 3,50), wyprzedzając pod tym względem nawet tradycyjne materiały nauczyciela (*Human*). Z kolei ChatGPT okazał się dla uczniów najbardziej angażujący i ciekawy (*interesting* = 3,94), choć w ogólnej ocenie przydatności (*useful*) wypadł najsłabiej.

Wyniki ankiety przeprowadzonej wśród nauczycieli

Perspektywa nauczycieli wnosi do badania dodatkowy, spójny kontekst oceny jakości przygotowanych materiałów dydaktycznych. W większości analizowanych kategorii najwyższe średnie oceny uzyskały materiały opracowane przez nauczyciela lub model ChatGPT. Nauczyciele ocenili materiały ChatGPT bardzo wysoko pod względem stymulowania samodzielnego myślenia uczniów (*independent_thinking* = 4,89) oraz trafności metodologicznej (*correct_method* = 4,78). Model ten uzyskał również najwyższą, ex aequo z nauczycielem, ocenę dopasowania poziomu trudności zadań do możliwości uczniów (*accurately_difficult* = 4,89).

Najwyżej ocenione pod względem chęci wykorzystania w praktyce dydaktycznej okazały się jednak materiały przygotowane przez nauczyciela (*would_use* = 4,67), nieznacznie wyprzedzając ChatGPT (4,56). Jednocześnie materiały nauczyciela wymagały najmniejszej liczby modyfikacji i poprawek (*require_adjustments* = 1,33), podczas gdy ChatGPT znalazł się na drugim miejscu (1,44). Wyniki te wskazują, że materiały generowane przez ChatGPT osiągają jakość bardzo zbliżoną do materiałów

przygotowywanych przez ekspertów, wymagając jedynie niewielkich korekt przed wykorzystaniem w praktyce szkolnej.



Rysunek 6. Wyniki ankiety dotyczącej użyteczności kart pracy wypełnionych przez badanych

Na przeciwnym biegunie ocen znalazł się model Gemini, który uzyskał najniższe średnie wyniki w zakresie wspierania samodzielnego myślenia uczniów (4,11) oraz aktywnego uczenia się (4,00). Materiały tego modelu wymagały również największej liczby modyfikacji przed wykorzystaniem w procesie dydaktycznym (require_adjustments = 2,11).

4. Podsumowanie

Przeprowadzone badanie pozwoliło na porównanie możliwości wybranych modeli językowych w kontekście generowania materiałów dydaktycznych dla uczniów szkoły podstawowej. Analizie poddano jakość przygotowywanych kart pracy, ich zgodność z założeniami dydaktycznymi oraz stopień użyteczności z perspektywy nauczyciela.

Uzyskane wyniki wskazują, że współczesne modele generatywnej sztucznej inteligencji mogą stanowić istotne wsparcie w procesie przygotowywania materiałów edukacyjnych i znacznie przyspieszyć pracę nauczyciela. Szczególnie modele Claude oraz ChatGPT charakteryzowały się wysoką jakością generowanych treści, większą spójnością materiałów oraz lepszym dostosowaniem do poziomu edukacyjnego uczniów. W przypadku modelu ChatGPT należy jednak zwrócić szczególną uwagę na kwestię

planowania czasu pracy z wygenerowanym materiałem. Tendencja tego narzędzia do tworzenia zbyt obszernej liczby zadań może sprawić, że ich ukończenie w wyznaczonym czasie okaże się dla uczniów niemożliwe. W warunkach szkolnych (np. podczas pisania sprawdzianów) może to bezpośrednio przekładać się na niższe oceny końcowe, wynikające z braku czasu na rozwiązanie całego testu, a nie z braku wiedzy. W badanych scenariuszach rozwiązania te wymagały również stosunkowo niewielkiej liczby poprawek merytorycznych i językowych, co może wskazywać na ich wysoką użyteczność w praktyce dydaktycznej. Interesującą obserwacją wynikającą z przeprowadzonej analizy jest wysoka skuteczność modelu Claude, który mimo częstego utożsamiania przede wszystkim z zastosowaniami programistycznymi i technicznymi, osiągnął wyniki porównywalne z ChatGPT — rozwiązaniem uznawanym za jednego z pionierów oraz liderów rynku generatywnej sztucznej inteligencji. W części analizowanych kategorii Claude uzyskiwał nawet lepsze rezultaty, szczególnie w subiektywnych ocenach uczniów. Może to wskazywać, że modele rozwijane pierwotnie z myślą o zastosowaniach technicznych posiadają również istotny potencjał w obszarze wspierania procesu dydaktycznego oraz tworzenia materiałów edukacyjnych.

W przypadku modelu Gemini zaobserwowano niższy poziom zgodności wygenerowanych materiałów z przyjętymi kryteriami oceny. Utworzone przez model karty pracy częściej wymagałyby dodatkowej weryfikacji oraz ingerencji nauczyciela, zarówno pod względem poprawności treści, jak i dostosowania do potrzeb uczniów. Wyniki sugerują, że aktualna wersja tego modelu może wymagać większego nadzoru eksperckiego przed wykorzystaniem wygenerowanych materiałów w procesie nauczania.

Jednocześnie przeprowadzone badanie potwierdza, że narzędzia oparte na dużych modelach językowych posiadają znaczący potencjał wspierający nauczycieli w zakresie automatyzacji przygotowywania materiałów dydaktycznych, personalizacji treści oraz ograniczania czasu potrzebnego na tworzenie ćwiczeń i kart pracy. Należy jednak podkreślić, że generowane materiały wymagają każdorazowo krytycznej oceny i nadzoru ze strony nauczyciela, szczególnie w kontekście poprawności merytorycznej oraz dostosowania do możliwości poznawczych uczniów.

Ograniczeniem przeprowadzonego badania jest dynamiczny rozwój modeli językowych oraz fakt, że uzyskane wyniki odnoszą się do konkretnych wersji narzędzi dostępnych w momencie realizacji analizy. W przyszłych badaniach warto rozszerzyć zakres porównania o większą liczbę modeli, różne poziomy edukacyjne oraz analizę wpływu wykorzystania generatywnej sztucznej inteligencji na efektywność procesu dydaktycznego w dłuższej perspektywie czasowej.

Literatura

1. Denny, P., Khosravi, H., Hellas, A., Leinonen, J., & Sarsa, S. (2023). Can we trust AI-generated educational content? Comparative analysis of human and AI-generated learning resources. arXiv. <https://arxiv.org/abs/2306.10509>.

2. Kerimbayev N., Menlibay Z., Garvanova M., Djaparova S. & Jotsov V., "A Comparative Analysis of Generative AI Models for Improving Learning Process in Higher Education," 2024 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2024, pp. 271-276, doi: 10.1109/ICAI63388.2024.1085149.
3. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55, 90–112. <https://doi.org/10.1111/bjet.13370>.
4. Sarkar, S., Sun, M., Liu, A., Tian, Z., Esbenshade, L., He, J., & Zhang, Z. (2025). Connecting feedback to choice: Understanding educator preferences in GenAI vs. human-created lesson plans in K-12 education – A comparative analysis. *arXiv preprint arXiv:2504.05449*, 2025.
5. Mavrych V, Yaqinuddin A, Bolgova O. Claude, ChatGPT, Copilot, and Gemini performance versus students in different topics of neuroscience. *Adv Physiol Educ*. 2025 Jun 1;49(2):430-437. doi: 10.1152/advan.00093.2024. Epub 2025 Jan 17. PMID: 39824512.
6. Silor, A. C., & Silor, F. S. C. (2025). Boosting reading comprehension through AI-based learning tools. *International Journal of Learning, Teaching and Educational Research*. <https://doi.org/10.26803/ijlter.24.9.4>.
7. Chen W. Intelligent tutor: Leveraging chatgpt and microsoft copilot studio to deliver a generative ai student support and feedback system within teams. *arXiv preprint arXiv:2405.13024*, 2024.
8. Ermis, S., Özal, E., Karapapak, M., Kumantaş, E., & Özal, S. A. (2025). Assessing the responses of large language models (ChatGPT-4, Claude 3, Gemini, and Microsoft Copilot) to frequently asked questions in retinopathy of prematurity: a study on readability and appropriateness. *Journal of Pediatric Ophthalmology & Strabismus*, 62(2), 84-95.
9. Więckiewicz-Modrzewska, J. ChatGPT w pracy nauczyciela. O możliwościach. *Annales Universitatis Mariae Curie-Skłodowska, sectio J–Paedagogia-Psychologia*, 39(1), 187-197.
10. Kowalczyk L.. "Możliwości wykorzystania sztucznej inteligencji w działalności dydaktycznej i naukowej." W: *Prace Naukowe Wyższej Szkoły Zarządzania i Przedsiębiorczości: Pedagogika, zarządzanie, psychologia i inżynieria zarządzania wobec wyzwań współczesności* 53.1 (2023): 9-39.
11. Rachwalska A. Lektura szkolna w dobie AI – możliwości i wyzwania w pracy z uczniem ze specjalnymi potrzebami edukacyjnymi. *Dydaktyka Polonistyczna* (2025). 20. 69-86. 10.15584/dyd.pol.20.2025.05.