**Model procesów rozwoju aplikacji do oceny prac studenckich z wykorzystaniem modeli LLM**

Cezary Orłowski, Marek Ożarowski, Tomasz Wawer

**Abstrakt**

Artykuł opisuje procesy rozwoju aplikacji do oceny prac studenckich. Aplikacja ma na celu wspierać proces dydaktyczny w ustaleniu samodzielnego przygotowania przez studentów odpowiedzi na pytanie z wykorzystaniem ChatGPT. Wsparcie procesu dydaktycznego przez aplikację polega na badaniu za pomocą odpowiednio dobranych metryk stopnia samodzielności studentów w realizowanych przez nich pracach. Oznacza to, że wykładowca będzie mógł sprawdzić, czy dana praca wykonana przez studenta za pomocą ChatGPT została przygotowana samodzielnie, czy została przygotowana tylko przez ChatGPT.

Aby zrealizować cel pracy została ona podzielona na następujące części. We wprowadzeniu przedstawiono stan badań w zakresu budowy aplikacji do wsparcia procesów dydaktycznych. Następnie pokazano lukę badawczą w tym zakresie, aby w rozdziale drugim zaprezentować model procesów budowy aplikacji dla wsparcia ocen studentów. W tej części pracy przedstawiono założenia do budowy modelu, środowisko prowadzenia badań oraz stasowane metryki oceny.

W kolejnej części artykułu przedstawiono wyniki opracowanej aplikacji, w której ocena prac studenckich określono na podstawie wybranych metryk. Artykuł podsumowuje dyskusja nad znaczeniem zaproponowanego rozwiązania oraz wnioski. Zaprezentowano również kierunki dalszych badań autorów.

**Słowa kluczowe: Edukacja, Generatywna** Sztuczna Inteligencja, ChatGPT, Teaching and Learning

### Wprowadzenie

W 2022 roku 35% spośród wszystkich światowych biznesów korzystało z systemów opartych na AI (Artificial Intelligence). Jak pokazują trendy w statystykach, w 2025 roku już ponad 80% planuje zautomatyzować swoją pracę i połączyć AI oraz zautomatyzować swoje zadania. ChatGPT obecnie stanowi jedno z najpopularniejszych narzędzi do poszukiwania informacji. W dniu premiery, czyli 30 listopada 2022 r., ChatGPT odwiedziło (chat.openai.com) 153 000 osób. Pod koniec pierwszego tygodnia liczba ta wyniosła 15,5 miliona, a w drugim tygodniu wzrosła do 58 milionów. Od początku ChatGPT kontynuował ścieżkę hiper rozwoju aż do maja 2023 r., kiedy to osiągnął dotychczas najwyższy poziom – 1,8 miliarda wizyt miesięcznie. W tym procesie strona internetowa openai.com firmy macierzystej OpenAI stała się jedną z najczęściej odwiedzanych domen na świecie [10].

Jeśli stopień wykorzystania ChatGPT będzie dokonywał się w takim tempie, to edukacja będzie musiała dokonać dużych zmian w procesach dydaktycznych. Zmiany te są konieczne, aby edukacja mogła sprostać oczekiwaniom pracodawców. Pracodawcy chcieliby mieć możliwość zatrudniania absolwentów wyposażonych w umiejętności wykorzystania takich narzędzi, jak ChatGPT.

Edukacja, procesy dydaktyczne muszą również sprostać oczekiwaniom gospodarki. Według badań prowadzonych przez autorów artykuły, obecnie każdy student przyznaje się do korzystania z ChatGPT. Dodatkowo, biznes i gospodarka poszukują specjalistów od „prompt engineeringu” a Ośrodki akademickie muszą dostosować swoje programy kursów do tych oczekiwań.

Aby takie oczekiwania mogły być realizowane praktycznie, proces dydaktyczny musi umożliwiać wykorzystanie takich narzędzi, jak ChatGPT. Wykorzystanie możliwości, jakie daje ChatGPT w edukacji musi zatem być procesem zdefiniowanym i kontrolowany. Jednocześnie należy zmienić podejście w samym procesie dydaktycznym. Jeśli studenci mają nauczyć się posługiwać ChatGPT w sposób świadomy, to należy przebudować sam proces dydaktyczny. Oznacza to, że wykładowcy powinny również posługiwać się takimi narzędziami, jak ChatGPT. Jeśli studenci posługują się tego typu rozwiązaniami do pisania prac ewaluacyjnych, to powstaje problem. W jaki sposób sprawdzić, czy napisana praca ewaluacyjna jest pracą samodzielnie napisaną przez studenta. Innymi słowy, należy rozstrzygnąć, czy student skorzystał podczas pisania pracy z tekstów wygenerowanych przez ChatGPT w sposób uznany za plagiat.

Autorzy artykułu ustalili, że obecnie każdy student korzysta z ChatGPT (ankiety własne na zajęciach praktycznych ze studentami). Sposób wykorzystania ChatGPT jest zróżnicowany. Najczęściej aktywności z ChatGPT dokonują się na zajęciach dydaktycznych oraz służą do pisania prac w ramach zajęć, ćwiczeń i laboratoriach odbywających sią na Uniwersytecie WSB. Podobny stopień wykorzystania tego narzędzia przez studentów zauważa się w innych ośrodkach akademickich. Wydaje się, że liczba dzienna sesji z ChatGPT rośnie.

Innym przykładem wykorzystania tego narzędzia są prace ewaluacyjne. Czy zatem, student może korzystać z ChatGPT i na jakich warunkach w procesie pisania pracy ewaluacyjnej. Pracodawcy i biznes oczekują tego typu umiejętności od przyszłych absolwentów uniwersytetów. Dlatego prace ewaluacje napisane za pomocą ChatGPT powinny również zostać dopuszczone przez ośrodki akademickie. Sposób i możliwość korzystania z ChatGPT powinna zostać umożliwiona w kontrolowany sposób. Aby zapewnić kontrolę nad stopniem wykorzystania ChatGPT w pracach ewaluacyjnych pisanych przez studentów, podjęto próby zbudowania odpowiednich narzędzi.

## Analiza literatury

Villagrán, I., Hernández, R., Schuit, G., Neyem, A., Fuentes-Cimma, J., Miranda, C., oraz Varas, J. (2024) przedstawili studium przypadku, w którym wykorzystano duży model językowy (LLM) w obszarze edukacji fizjoterapeutycznej. Opisany przypadek pokazuje skuteczne wykorzystanie generatywnej sztucznej inteligencji w praktycznym szkoleniu dla szkoleń z umiejętności prowadzonych w edukacji zawodów medycznych. Proponowane rozwiązanie wykorzystuje LLM do automatycznej oceny informacji zwrotnych wprowadzanych przez instruktorów na podstawie wstępnie zdefiniowanych i opartych na literaturze kryteriów jakości i generuje praktyczne wyjaśnienia tekstowe do przeformułowania. Ponadto, jeśli instruktor tego wymaga, narzędzie obsługuje generowanie podsumowań dla dużych zestawów danych tekstowych, aby osiągnąć lepszy odbiór i zrozumienie przez studentów. Omawiane w artykule Studium przypadku opisuje, w jaki sposób te funkcje zostały zintegrowane z platformą zorientowaną na sprzężenie zwrotne. W jaki sposób ich skuteczność została oceniona w kontrolowanym otoczeniu z udokumentowanymi informacjami zwrotnymi. Prezentuje wyniki ich wdrożenia z udziałem prawdziwych użytkowników poprzez przejścia poznawcze. Wstępne wyniki wskazują, że ta innowacyjna implementacja ma duży potencjał. Poprawia naukę i wydajność w edukacji fizjoterapeutycznej i ma potencjał, aby rozszerzyć się na inne dyscypliny medyczne, w których rozwój umiejętności proceduralnych ma kluczowe znaczenie. Opisany przypadek oferuje cenne narzędzie do oceny i poprawy informacji zwrotnej w oparciu o standardy jakości dla skutecznych procesów informacji zwrotnej. [1]

Meissner, R., Pögelt, A., Ihsberner, K., Grüttmüller, M., Tornack, S., Thor, A., ... & Hardt, W. (2024, October) opisali nowy proces i aplikację napędzaną przez LLM, zwaną ItemForge. Aplikacja została dostosowana specjalnie do automatycznego generowania elementów e-oceny w matematyce. Podejście to jest w pełni dostosowane do poziomów i hierarchii celów uczenia się poznawczego opracowanych przez Andersona i Krathwohl. Aplikacja bierze pod uwagę określone koncepcje matematyczne z rozważanych kursów. Jakość wygenerowanych elementów tekstu swobodnego wraz z odpowiadającymi im odpowiedziami (przykładowe rozwiązania), a także ich adekwatność do wyznaczonego poziomu poznawczego i przedmiotu, zostały ocenione w badaniu na małą skalę. Badanie obejmowało trzy eksperymenty matematyczne, przeanalizowano łącznie 240 wygenerowanych elementów. Ustalenia pokazują, że narzędzie jest biegle w tworzeniu wysokiej jakości elementów, które są zgodne z wybranymi koncepcjami i docelowymi poziomami poznawczymi, co wskazuje na jego potencjalną przydatność do celów edukacyjnych. Zaobserwowano, że dostarczone odpowiedzi (przykładowe rozwiązania) czasami wykazywały niedokładności lub nie były całkowicie kompletne, co sygnalizuje konieczność dalszego udoskonalenia procesów narzędzia [2]

Smolić, E., Pavelić, M., Boras, B., Mekterović, I., & Jagušt, T. (2024, May) wykorzystano wielki model językowy (LLM) do ekstrakcji informacji i/lub ekstrapolacji z różnych źródeł. W edukacji informatycznej potencjalnym zastosowaniem takiej technologii jest automatyczny przegląd kodu. LLM mógłby wówczas przejąć ciężar debugowania niekompilowalnego kodu. Tego typu wsparcie zapewniane przez LLM polegałoby na wykrywaniu przeoczonych problemów optymalizacyjnych takich jak, słabe zarządzanie pamięcią w kodzie, gdzie w przeciwnym razie przechodziłby automatyczne testy. LLM mógłby wykonywać również inne zaawansowane zadania w czasie pisania kodu. Jednak obecnie LLM nie są w stanie oceniać kodu lub wyrażeń matematycznych ze 100% niezawodnością. LLM może wspierać rozpoznawanie wzorców tokenów i późniejszym generowaniem odpowiedzi probabilistycznych. Mając to na uwadze, zbadano ryzyko nieprawidłowej oceny kodu przez LLM, zarówno opisowej, jak i numerycznej. W pracy zaproponowano badania nad łagodzeniem tego ryzyka oraz zaproponowano dalsze kierunki. [3]

Joy Kulangara, K. (2024) omawia potencjał LLM w edukacji programistycznej. Powszechnie wiadomo, że duże modele językowe (LLM) mają potencjał, aby poprawić edukację programistyczną, zapewniając studentom informacje zwrotne i wskazówki. Pomimo potencjalnych korzyści, integracja LLM z edukacją stwarza wyjątkowe wyzwania, w tym ryzyko nadmiernego polegania na ich informacjach zwrotnych i niespójność jakości informacji zwrotnych. W pracy przedstawiono elastyczną platformę, która może integrować wiele LLM, zapewniając jednocześnie przestrzeń eksperymentalną do badań. Umożliwia innowacyjne podejść w celu ulepszenia edukacji programistycznej za pomocą LLM. Opracowana platforma skutecznie demonstruje wykonalność integracji LLM z edukacją programistyczną. Badanie na małą skalę oceniające ogólną użyteczność platformy otrzymało średnią ocenę 4,21 na 5,00, podczas gdy opinia zwrotna LLM otrzymała średnią ocenę użyteczności 4,28 na 5,00, co podkreśla jej skuteczność i wartość w pomaganiu studentom. Chociaż wielkość próby badawczej była niewielka, wyniki są zachęcające. [4]

Jošt, G., Taneski, V., & Karakatič, S. (2024) Artykuł bada niuanse wpływu nieformalnego wykorzystania LLM na wyniki uczenia się studentów studiów licencjackich w zakresie edukacji w zakresie rozwoju oprogramowania Zaprojektowano eksperyment z udziałem trzydziestu dwóch uczestników w ciągu dziesięciu tygodni, w którym badano nieograniczone, ale nie specjalnie zachęcane wykorzystanie LLM i ich korelację z wynikami studentów. Wyniki ujawniają znaczącą negatywną korelację między zwiększonym poleganiem na LLM w przypadku zadań wymagających intensywnego myślenia krytycznego, takich jak generowanie kodu i debugowanie, a niższymi ocenami końcowymi. Ponadto w opisanym eksperymencie obserwuje się tendencję spadkową w ocenach końcowych przy zwiększonym średnim wykorzystaniu LLM we wszystkich zadaniach. Jednak korelacja między wykorzystaniem LLM w celu poszukiwania dodatkowych wyjaśnień a ocenami końcowymi nie była tak silna, co wskazuje, że LLM mogą lepiej służyć jako uzupełniające narzędzie do nauki. Wyniki te podkreślają znaczenie łączenia studiów LLM z rozwijaniem umiejętności samodzielnego rozwiązywania problemów w nauczaniu programowania. [5]

Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., & Gu, X. (2024) Artykuł opisuje badanie i analizę wyników dla dużych modeli językowych (LLM) w projektowaniu dydaktycznym. Ma to na celu ujawnienie ich potencjalnych mocnych stron i możliwych słabości. Wpływ LLM stopniowo wzrósł w wielu dziedzinach, jednak badania eksploracyjne nad ich zastosowaniem w edukacji pozostają stosunkowo rzadkie. Wykorzystując Generative Pretrained Transformer 4, wygenerowano zbiór danych planu nauczania matematyki w szkole średniej. Na koniec oceniono wyniki LLM w projektowaniu dydaktycznym. Wyniki oceny wykazały, że plany nauczania generowane przez LLM wyróżniają się w ustalaniu celów dydaktycznych, identyfikowaniu priorytetów nauczania, organizowaniu łańcuchów problemów i zajęć dydaktycznych, artykułowaniu treści przedmiotu oraz doborze metod i strategii. Szczególnie godne pochwały wyniki odnotowano w modułach statystyki i funkcji. Istnieje jednak pole do poprawy w aspektach związanych z kulturą matematyczną i oceną interdyscyplinarną, a także w modułach geometrii i algebry. Na koniec, niniejsze badanie proponuje inicjatywy, takie jak oparte na monitach szkolenie nauczycieli wykorzystujących LLM i integrację LLM skoncentrowanych na matematyce [6]

Telesko, R., & Wilke, G. (2024) praca opisuje prototyp oparty na Pythonie, który półautomatycznie generuje egzaminy i podsumowania oparte na MCQ (pytaniach wielokrotnego wyboru). Wykorzystano tu bardzo popularne duże modele językowe (LLM). Obecne oprogramowanie prototypowe opiera się na bibliotekach Pythona langchain dla struktury LLM i streamlit dla interfejsu użytkownika (UI). Jako baza danych wektorowych używana jest ChromaDB, a jako LLM Vertex AI od Google. Oprogramowanie jest testowane z wybranymi dokumentami kursu „Uczenie maszynowe z Pythonem” (który jest częścią specjalizacji w ramach programu BIT), aby dowiedzieć się, czy jakość generowanych egzaminów i podsumowań odpowiada oczekiwaniom wykładowców i studentów. Wyniki pomogą odpowiedzieć na ważne pytania dotyczące praktycznego wykorzystania LLM w edukacji. Wkład omawianej pracy ma na celu rozwiązanie problemu w dziedzinie wsparcia egzaminacyjnego w szkolnictwie wyższym technicznym. [7]

Agostini, D., & Picasso, F. (2024) w swojej pracy omawiają przegląd literatury w obszarze wykorzystania LLM w edukacji. Wykładowcy akademiccy stają w obliczu oceny rozległych zajęć. Dzięki szybkiemu postępowi w zakresie modeli dużego języka (LLM) oraz ich rosnącej dostępności, część rozwiązania tych problemów może być dostępna w postaci wsparcia. W rzeczywistości, LLM mogą przetwarzać duże ilości tekstu, podsumowywać je i przekazywać informacje zwrotne na ich temat zgodnie z ustalonymi kryteriami. Wnioski z tej analizy można wykorzystać zarówno do przekazywania informacji zwrotnych studentom, jak i w pomocy wykładowcy w ocenie tekstu. Przy odpowiednich ramach pedagogicznych i technologicznych LLM-y mogą odciągnąć instruktorów od niektórych kwestii związanych ze zrównoważonym rozwojem, a więc od jedynego wyboru testu wielokrotnego wyboru i podobnych. [8]

Huovinen, L. (2024) w pracy opisano projekt wykorzystania LLM w edukacji. Celem projektu końcowego było zbadanie potencjału wykorzystania dużych modeli językowych (LLM) w automatyzacji i ulepszaniu procesu opracowywania programów nauczania. Projekt został zrealizowany poprzez opracowanie i wdrożenie aplikacji internetowej opartej na LLM, a następnie przeprowadzenie testów użyteczności przy użyciu podejścia poznawczego i sesji opinii użytkowników. Użytkownikom testowym przedstawiono narzędzie i zlecono ocenę jego skuteczności, przejrzystości i ogólnej użyteczności w kontekście planowania programu nauczania. Opinie użytkowników podkreśliły znaczenie jasnych wskazówek, integracji zadań w scentralizowanym narzędziu oraz potrzebę udziału człowieka w weryfikacji i udoskonalaniu generowanych treści, aby zapewnić dokładne i znaczące cele w projektowaniu programu nauczania [9]

Yuan, B., & Hu, J. (2024) Ostatnie postępy w dużych modelach językowych (LLM) w ramach sztucznej inteligencji (AI) przedstawiają obiecujące nowe drogi do usprawnienia procesów oceny kursów. W artykule opisano zastosowanie LLM w automatycznej ocenie kursów z wielu perspektyw i przeprowadzono rygorystyczne eksperymenty na 100 kursach na dużym uniwersytecie w Chinach. Wyniki wskazują, że: (1) LLM mogą być skutecznym narzędziem oceny kursów; (2) ich skuteczność zależy od odpowiedniego dostrojenia i szybkiej inżynierii; i (3) wyniki oceny generowane przez LLM wykazują znaczący poziom racjonalności i interpretowalności. [10]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mao, Chen oraz Liu (2024) poruszają w swojej pracy kwestie wykorzystania narzędzi AI jak ChatGPT. Zwracają uwagę na wykorzystanie AI w celu zwiększenia efektywności w pracy zawodowej oraz w procesie edukacji. Jednocześnie podkreślają fakt zatarcia się granicy pomiędzy plagiatem a własną pracą. Sugerują podanie autora poprzez dołączenie wpisu odniesienia podczas cytowania tekstu wygenerowanego przez AI (np. cytowanie tekstu wygenerowanego przez ChatGPT poprzez odniesienie do OpenAI). Zalecają również, aby badacze opisywali wykorzystanie przez siebie narzędzi AI w sekcji metody, jeśli użyli takich narzędzi w procesie. Nie jest niczym niezwykłym, że autorzy wymieniają ChatGPT lub inne generatory treści, jako współpracowników lub współautorów. Stało się to najnowszym wyzwaniem, przed którym stanęli naukowcy i wydawcy. Oprócz ostrzeżeń ze strony społeczności naukowej, musimy być ostrożni w kwestii niebezpieczeństw związanych z poleganiem wyłącznie na narzędziach AI do badań ze względu na zjawisko halucynacji AI. Zjawisko to odnosi się do zdolności AI do tworzenia nieistotnych lub nieprawdziwych informacji przy użyciu przekonującego języka [1].

Baidoo-anu, Owusu Ansah (2023) dzielą się badaniami zawierającymi zalecenia dotyczące tego, w jaki sposób ChatGPT można wykorzystać w celu maksymalizacji nauczania i uczenia się. Decydenci, naukowcy, edukatorzy i eksperci w dziedzinie technologii mogliby ze sobą współpracować i rozpocząć rozmowy na temat tego, w jaki sposób te rozwijające się narzędzia generatywnej sztucznej inteligencji mogłyby być bezpiecznie i konstruktywnie wykorzystywane do poprawy edukacji i wspierania nauki uczniów [2].

Prajapati, Kumar, Singh (2024) omawiają w swojej pracy zalety sztucznej inteligencji, które można dostrzec w całym szkolnictwie wyższym. Generowane są nowe wyzwania, które dają studentom więcej czasu na nowe metody uczenia się, analizę porównawczą dla miejsc pracy w konkurencyjnych przedsiębiorstwach o większej wydajności dzięki automatyzacji. Coraz częściej studenci otrzymują zindywidualizowane instrukcje edukacyjne od korepetytorów AI, które mogą uwzględniać unikalny styl uczenia się studenta. Tacy korepetytorzy mapują również szybkość uczenia się za pomocą różnych interaktywnych narzędzi. Istnieje kilka potencjalnych zalet sztucznej inteligencji dla nowych miejsca pracy z perspektywy badawczej [3].

Alier, Marc; García-Peñalvo, Francisco; Camba, Jorge D. (2024) Poruszają w swojej publikacji możliwości wykorzystania Systemów AI w edukacji. Systemy AI na przykład, może być wykorzystywana do tworzenia niestandardowych quizów, generowania podpowiedzi do esejów, a nawet oceniania esejów. Wykorzystanie AI, jako narzędzia oceny może zmniejszyć obciążenie pracą nauczycieli i pomóc uczniom w otrzymaniu szybkiej informacji zwrotnej na temat ich pracy. Włączenie AI do środowiska edukacyjnego wiąże się również z wyzwaniami związanymi z uczciwością akademicką. Wraz z dostępnością modeli AI, uczniowie mogą wykorzystywać je do nauki lub wykonywania zadań domowych, co może budzić obawy, co do autentyczności i autorstwa dostarczonej pracy. Dlatego ważne jest, aby zapewnić utrzymanie standardów akademickich i zachować oryginalność pracy studenta. Kwestia ta, podkreśla potrzebę wdrożenia etycznych praktyk w zakresie korzystania z modeli AI i zapewnienia, że technologia jest wykorzystywana do wspierania, a nie zastępowania doświadczenia edukacyjnego ucznia [4].

Baidoo-anu, Owusu Ansah, (2023) podkreślają w swojej pracy niezwykłe zdolności ChatGPT do wykonywania złożonych zadań w dziedzinie edukacji. Wykorzystanie przez uczniów i studentów tych zdolności wywołały mieszane uczucia wśród nauczycieli. Postęp w rozwiązaniach opartych na sztucznej inteligencji wydaje się rewolucjonizować istniejącą praktykę edukacyjną. W swoim artykule autorzy syntetyzuje najnowszą literaturę, aby zaoferować potencjalne korzyści i wady ChatGPT w promowaniu nauczania i uczenia się. Korzyści z ChatGPT obejmują między innymi promowanie spersonalizowanego i interaktywnego uczenia się, generowanie podpowiedzi do działań związanych z oceną kształtującą, które zapewniają bieżące informacje zwrotne w celu informowania o nauczaniu i uczeniu się [5].

Autorzy przedstawionych powyżej artykułów, jak też doświadczenia własne autorów, podkreślają wyzwania i ograniczenia tych systemów oraz przedstawiają zalecenia dotyczące przyszłych badań. W związku z powyższym, pojawiają się główne problemy utylitarne.

1. Obecnie każdy student wykorzystuje do poszukiwania wiedzy takie rozwiązanie jak ChatGPT, jednocześnie nie ma pewności, czy wiedza dostarczona przez ChatGPT została w jakikolwiek sposób zweryfikowana przez studenta.
2. Nawet jeśli student skorzysta z takiego rozwiązania jak ChatGPT i wygeneruje za jego pomocą materiał do swojej pracy, to wykładowca nie jest w stanie określić, jak dalece praca studenta jest samodzielną pracą.
3. Aby takie działania zweryfikować, wykładowca musi wykonać również dodatkowe działania z ChatGPT, co wpływa się na wydłużenie czasu procesu oceny pracy studenta przez wykładowcę.
4. Udzielenie poprawnej odpowiedzi przez studenta na zapytanie za pomocą ChatGPT, nie jest równoznaczne, że dany student zrozumiał materiał, który zawarł w pracy.

Dla tak wyodrębnionych problemów utylitarnych w artykule postawiono następujące pytania badawcze:

1. Jaki model językowy będzie w stanie optymalnie ocenić samodzielność wygenerowanej pracy studenta?
2. Jakie metryki mogą być użyteczne do oceny samodzielności prac studenckich?
3. Jeśli dostępne metryki okażą się nieskuteczne, to jakie powinny być kryteria oceny samodzielności prac studenckich?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Model procesów rozwoju aplikacji do oceny prac studenckich z wykorzystaniem modeli LLM

W tej części artykułu dla postawionych powyżej problemów utylitarnych i będących odpowiedzią na te problemy utylitarne pytań badawczych, zaprezentowano model procesu rozwoju aplikacji dla oceny prac studenckich z wykorzystaniem modeli LLM. W artykule przedstawiono założenia pokazujące: z jednej strony stan badań, a z drugiej strony, możliwości określania procesu budowy aplikacji i kryteria oceny prac studenckich. Kolejno przedstawiono środowisko badawcze w którym zaimplementowany zostanie model. Aby pokazać znaczenie samego modelu, wskazano na metryki oceny prac studenckich. Analizowane metryki są istotne zarówno dla procesu projektowania aplikacji, jak i budowy modelu dla późniejszego jego wykorzystania w ocenie prac studenckich.

### 1.1. Założenia do modelu

W celu zbadania stopnia samodzielności prac ewaluacyjnych napisanych przez studentów z wykorzystaniem ChatGPT, zaprezentowano procesy budowy aplikacji do oceny tych prac. Założeniem budowy środowiska badawczego było dopuszczenie możliwości skorzystania z ChatGPT przez studentów w czasie pisanie prac ewaluacyjnych. Sposób użycia ChatGPT został dokładnie określony przez Wykładowców. Student mógł napisać pracę korzystając ze wsparcie ChatGPT –musiał jednak kierować się pewnymi warunkami. Praca ewaluacyjna musi zostać napisana samodzielnie przez studenta. Jeśli student kopiował jakieś fragmenty, które zaproponował mu ChatGPT, musiał podać linki do źródeł w celu weryfikacji wygenerowanej informacji.

Następnie przystąpiono do realizacji kodu zapytania, który miał na celu ocenę samodzielności prac ewaluacyjnych studentów. Zapytanie korzystało z narzędzia ChatGPT. Jest to chatbot opracowany przez OpenAI, wykorzystujący model GPT (generative pre-trained transformer) i służący do generowania odpowiedzi na dane wprowadzane przez użytkownika. Model ten został opracowany na podstawie dużych zbiorów danych, tak aby mógł prowadzić rozmowę i angażować się w różne tematy, od ogólnych rozmów, po określone obszary wiedzy eksperckiej. Wybór tego narzędzia był pewną konsekwencją naszego eksperymentu. Do kompletności naszego środowiska badawczego konieczne były również kryteria oceny pracy napisanej przez studenta. Zostało to zrealizowane za pomocą metryk – algorytmów do oceny odległości, podobieństwa danych.

### Środowisko badawcze

Środowisko badawcze obejmowało trzy główne komponenty:

* Proces egzaminu zakresu systemu IoT, w którym brało udział 154 studentów. Studenci przygotowywali pracę z wykorzystaniem wiedzy własnej oraz różnych źródeł referencyjnych w tym modeli językowych
* Opracowania przez autorów artykułu aplikacji, która pozwalała na automatyczną ocenę 154 prac studenckich. Opracowanie uwzględniało konieczne kryteria oceny prac studenckich, które bazują na wybranych metrykach.
* Dobór metryk do oceny przesłanych przez studentów prac oraz tekstów referencyjnych generowanych przez modele językowe

Studenci przygotowywali prace, które następnie podlegały ocenie przez aplikację opracowaną przez autorów artykułu. Proces został przeprowadzony w okresie czerwiec, wrzesień 2024 roku.

### Metryki zastosowane do oceny prac ewaluacyjnych studentów

W procesie oceny prac napisanych przez studentów stosowano początkowo dwie metryki BLEU i ROUGE. Dobór metryki wynikał z jego zastosowań w zadaniach tłumaczenia maszynowego, w których celem jest automatyczne tłumaczenie tekstu z jednego języka na inny. Został on zaproponowany jako sposób oceny jakości tłumaczeń generowanych maszynowo poprzez porównanie ich z zestawem tłumaczeń referencyjnych dostarczonych przez tłumaczy ludzkich. Punktacja BLEU waha się od 0 do 1, przy czym wyższe wartości, wskazują na lepszą jakość tłumaczenia. Idealne tłumaczenie miałoby wynik BLEU równy 1, podczas gdy całkowicie niepoprawne tłumaczenie miałoby wynik BLEU równy 0.

Z kolei dobór metryki ROUGE był uwarunkowany jego wykorzystaniem w zadaniach streszczania tekstu, gdzie celem jest automatyczne wygenerowanie zwięzłego podsumowania dłuższego tekstu. ROUGE został zaprojektowany w celu oceny jakości podsumowań generowanych maszynowo poprzez porównanie ich z podsumowaniami referencyjnymi dostarczanymi przez ludzi. Wynik ROUGE waha się od 0 do 1, przy czym wyższe wartości wskazują na lepszą jakość podsumowania. Podobnie jak wynik BLEU, idealne podsumowanie miałoby wynik ROUGE równy 1, podczas gdy całkowicie nieprawidłowe podsumowanie miałoby wynik ROUGE równy 0.

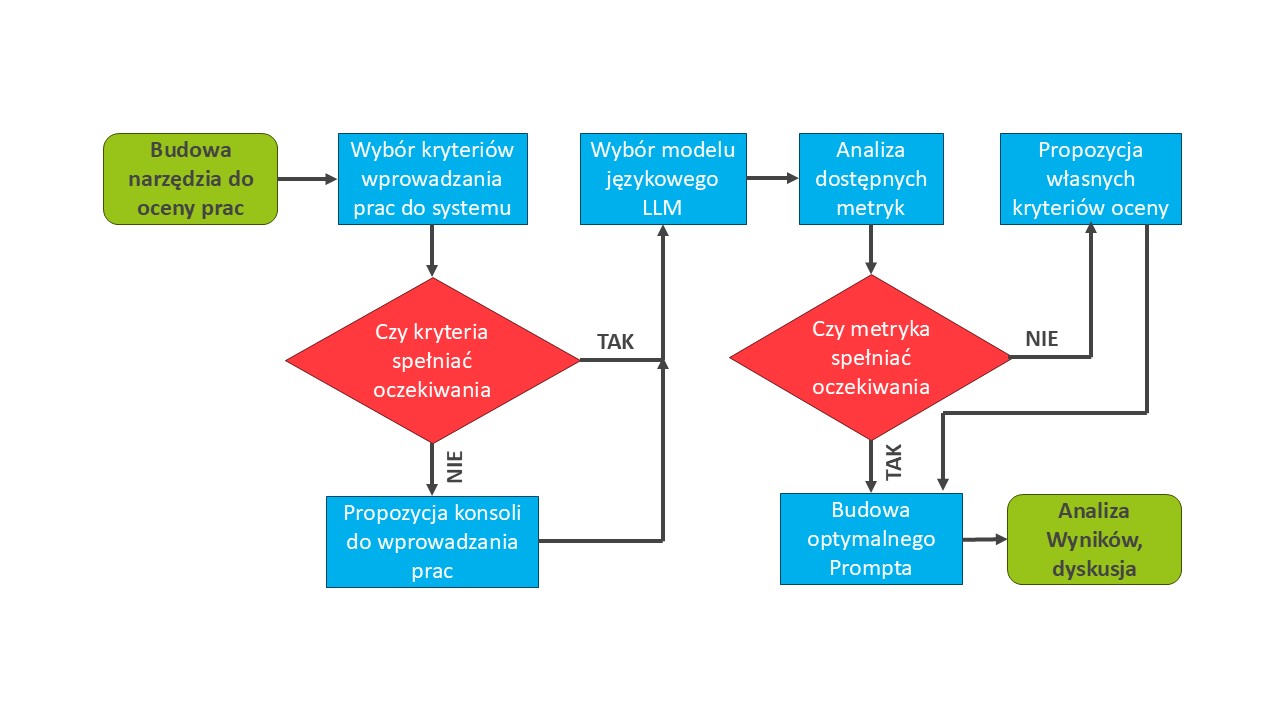
W kolejnym kroku uwzględniono także metrykę Euclidean. W uczeniu maszynowym odległość euklidesowa jest często używana jako miara podobieństwa do porównywania wektorów cech. Wektor cech reprezentuje punkt danych w wielowymiarowej przestrzeni, gdzie każdy wymiar odpowiada określonej funkcji lub atrybutowi. Obliczając odległość euklidesową między wektorami cech, możemy określić, jak bardzo są one podobne lub niepodobne.

Zastosowano także metrykę braycurtis. Metryka Bray-Curtis, znana również jako miara dystansu Bray-Curtis, jest jednym ze sposobów mierzenia różnic pomiędzy dwoma próbkami w wielowymiarowej przestrzeni cech. Jest szczególnie używana w ekologii do porównywania podobieństwa lub różnorodności próbek ekologicznych, ale ma także zastosowania w innych dziedzinach, takich jak analiza danych, biologii molekularnej, chemometrii, i innych.

Uwzględniono także metrykę Chebyshev - Metryka Chebysheva, znaną również jako odległość Chebysheva, jest miarą odległości używaną w matematyce i analizie danych do określania maksymalnej różnicy między dwoma wektorami w przestrzeni wielowymiarowej. W literaturze można ją spotkać również pod nazwą metryka L∞ lub odległość sup (supremum).

### Procesy budowy modelu

Procesy budowy modelu stanowiącego podstawę do budowy aplikacji obejmowały następujące etapy przedstawione na rys. 1.



Na rys. 1. przedstawiono procesy budowy modelu

Celem badawczym było opracowanie kryteriów w postaci „prompta do ChatGPT”, które mogłyby określić stopień samodzielności studenta w czasie pisania pracy ewaluacyjnej za pomocą ChatGPT. W tym celu, zaprojektowano i wykonano odpowiednie zapytanie do ChatGPT, które uwzględnia odpowiednie kryteria. Początkowo założono, że opracowanie takiego „zapytania do ChatGPT” powinno uwzględniać wyżej wymienione metryki, które mogłyby wyliczyć odległości pomiędzy odpowiedzią wygenerowaną przez ChatGPT a odpowiedzią wygenerowaną przez studenta – im mniejsza odległość tym teksty są podobne do siebie. Taka wiedza mogłaby wpierać w sposób automatyczny osobę oceniającą prace ewaluacyjne wygenerowane przez studentów. Metryki mogłyby wyliczyć odległości pomiędzy tekstami – referencją wygenerowaną przez ChatGPT oraz pracą napisaną przez studenta. Na obecnym stanie wybrano kilka metryk w celu sprawdzenia ich przydatności. Całość rozwiązania została przygotowana w oparciu o zebraną próbkę 154 prac ewaluacyjnych.

Do porównań odległości tekstów skorzystano z następujących metryk: euclidean, braycurtis, chebyshev, rouge\_1, rouge\_L. Następnie zaproponowano odpowiedni „prompt” do ChatGPT, który generował odpowiedź na pytanie do ChatGPT i porównywał tą odpowiedź z pracą ewaluacyjną napisaną przez studenta. Metryki miały wyliczyć zbieżność lub stopień rozbieżności obu tekstów. Ocena i weryfikacja prac studenckich polegała na podaniu wyniku, który jest interpretacją „samodzielności” studenta przy generowaniu swojej pracy ewaluacyjnej. System oceny samodzielności został tak przygotowany, aby w końcowym etapie wyliczania samodzielności studenta podać liczbę, która odpowiadałaby ocenie – 1 praca niesamodzielna, 5 praca samodzielna. Jeśli praca była napisana samodzielnie przez studenta, system ocen zwracał liczbę odpowiednią liczbę 1, 2, 3, 4 lub 5.

### Wyniki zastosowania modelu

Opracowany model został zastosowany do oceny 154 prac ewaluacyjnych studentów. Niektóre prace nie nadawały się do analizy, ponieważ nie spełniały kryteriów postawionych studentom w czasie generowanie prac ewaluacyjnych. Prace to zostały pominięte. Do analizy zakwalifikowano 128 prac ewaluacyjnych. Na tej podstawie przygotowano odpowiednie skrypty do „promptu” dla ChatGPT. Poniżej w tabeli 1. zostały zaprezentowane dane oceny dla przykładowych prac ewaluacyjnych. Dane osobowe studentów zostały ukryte, aby zapewnić ochronę danych osobowych. Z przeprowadzonych testów wygenerowano jedną tabelę dla przykładowych wyników eksperymentu. W eksperymencie, wykonano kilka podejść, jednak do artykułu została dołączona tabela z wybranymi wynikami eksperymentu. Pozostałe, pośrednie wyniki nie zostały umieszczone. Zastosowano kilka iteracji dla wygenerowania odpowiedniego „prompta”.

Uzyskane wyniki eksperymentu za pomocą kolejnych iteracji, umożliwiły wykonanie na podstawie modelu aplikację docelową. Treść jednego z „prompta” do ChatGPT zamieszczono poniżej:

*odp\_kolokwium\_prompt = """*

*jako inżynier IoT odpowiedz krotko na poniższe pytanie:*

*"""*

*check\_prompt = """ jako inżynier IoT powiedz krótko, bez szczegółów, czy to ChatGPT napisał ten tekst? : """*

*compare\_prompt = """ porównaj z tym tekstem napisanym przez ChatGPT :"""*

*ocen = "oceń w skali 1-5,bez opisu, samą liczbą, "*

*wstep = "Jako surowy nauczyciel akademicki oceń kolowium, podane poniżej, na temat: "*

*ocen\_strukture = f"{ocen}, czy struktura kolokwium jest poprawna"*

*ocen\_punkty = f"{ocen} czy punkty wyczerpują temat w odpowiedzi"*

*ocen\_zawartosc\_merytoryczna = f"{ocen} zawartość merytoryczną punktów, czy wyczerpują temat z pytania "*

*ocen\_odnosniki = f"{ocen} czy są odnośniki i linki do zewnętrznych źródeł informacji, takich jak artykuły, strony web, repozytoria github"*

*ocen\_samodzielnosc = f"{ocen}, czy kolokwium zostało napisane samodzielnie, 1 jeśli jest to kopia ChatGPT, 5 jeśli całkowicie samodzielnie."*

*parametry\_oceny = f"oceny podaj w formacie: str:[ocena struktury], pun[ocena punktow], zaw[ocena zawartości], odn[ocena odnosników], sam[ocena samodzielności]. Nie uzywaj znakow końca linii tylko przecinków"*

*def ocena\_kolokwium(pyt, odp):*

*return f"""*

*{wstep}{pyt}*

*{ocen\_strukture}*

*{ocen\_punkty}*

*{ocen\_zawartosc\_merytoryczna}*

*{ocen\_odnosniki}*

*{ocen\_samodzielnosc}*

*{parametry\_oceny}*

*Oceń poniższe kolokwium: {odp}*

*"""*

Tabela nr 1 Wyniki oceny prac studenckich na podstawie aplikacji opracowanej z wykorzystaniem modelu (przykładowe wyniki analizy dla poszczególnych metryk)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **euclidean** | **braycurtis** | **chebyshev** | **rouge\_1** | **rouge\_L** | **struktura** | **punkty** | **zawartość** | **odnośniki** | **zrozumienie** | **samodzielność** |
| **1.** | 14,3037 | 0,5075 | 3,031800032 | 0,205 | 0,114 | 4 | 4 | 4 | 2 | 4 | 3 |
| **2.** | 11,2487 | 0,4304 | 1,973399997 | 0,312 | 0,134 | 3 | 3 | 3 | 4 | 3 | 2 |
| **3.** | 10,0884 | 0,3412 | 1,7227 | 0,054 | 0,037 | 1 | 1 | 3 | 1 | 3 | 2 |
| **4.** | 6,4677 | 0,2329 | 0,917100012 | 0,161 | 0,09 | 3 | 3 | 4 | 1 | 4 | 3 |
| **5.** | 7,1435 | 0,2331 | 1,116000056 | 0,107 | 0,081 | 1 | 1 | 2 | 1 | 3 | 3 |
| **6.** | 17,9935 | 0,6843 | 3,492599964 | 0,229 | 0,102 | 1 | 3 | 3 | 1 | 3 | 3 |
| **7.** | 13,6355 | 0,5182 | 2,859400034 | 0,323 | 0,151 | 2 | 3 | 3 | 1 | 3 | 3 |
| **8.** | 19,0653 | 0,6677 | 4,021299839 | 0,203 | 0,105 | 5 | 5 | 5 | 1 | 5 | 4 |
| **9.** | 6,2819 | 0,1917 | 1,092100024 | 0,1 | 0,074 | 4 | 5 | 4 | 3 | 4 | 3 |
| **10.** | 13,5221 | 0,5087 | 2,707000017 | 0,269 | 0,113 | 3 | 4 | 4 | 1 | 4 | 2 |
| **11.** | 13,0358 | 0,4782 | 2,637000084 | 0,253 | 0,129 | 1 | 2 | 3 | 1 | 3 | 4 |
| **12.** | 15,3231 | 0,5432 | 3,3677001 | 0,221 | 0,111 | 3 | 4 | 4 | 1 | 4 | 4 |

Wyniki eksperymentu zostały porównane z ocenami prac ewaluacyjnych studentów wykonanymi „ręcznie” przez wykładowców. Następnie porównano oba wyniki dla każdej pracy ewaluacyjnej w celu dokonania analizy każdego przypadku. Ponieważ wyniki oparte na wybranych metryka nie dawały wartości oczekiwanych – oznacza to, że różniły się od wartości oczekiwanych, zastosowano kryteria oceny prac ewaluacyjnych. Kryteria oceny zostały zaproponowane przez autorów artykułu. Zaproponowane kryteria ocen cząstkowych przez autorów dotyczyły:

1. „struktury” dokumentu wykonanego przez studenta (struktura pracy ewaluacyjnej, czy struktura kolokwium jest poprawna)
2. „punkty” - czy punkty w strukturze dokumentu wyczerpują temat w odpowiedzi
3. „zawartość” merytoryczna dokumentu wykonanego przez studenta (zawartość merytoryczną punktów, czy wyczerpują temat z pytania)
4. „odnośniki” zawarte w dokumentach wykonanych przez studentów (czy praca ewaluacyjna wykonana przez studenta posiada odnośniki i linki do zewnętrznych źródeł informacji, takich jak artykuły, strony web, repozytoria GitHub)
5. „zrozumienie” – czy student odpowiedział w pracy ewaluacyjnej w sposób zrozumiały
6. „samodzielność” – czy student generujący pracę ewaluacyjną, który korzystała z ChatGPT kopiował treści, czy udzielił odpowiedzi samodzielnie.

Z przeprowadzonej analizy danych wynika, że wykorzystanie poszczególnych metryk nie daje jednoznacznych wyników pod względem przydatności zastosowanych metryk. Oznacza to, że autorzy artykułu nie są w stanie wskazać jednoznacznie, która z metryk daje oczekiwane wyniki. Lepszym rozwiązaniem okazało się zastosowanie własnych kryteriów oceny prac ewaluacyjnych wykonanych przez studentów. Wybrane wyniki zostały zagregowane a przykładowe wyniki umieszczone zostały w tabeli nr 1.

## Dyskusja

Do prowadzenie eksperymentu autorzy artykułu zdecydowali się na narzędzie ChatGPT firmy OpenAI. Wybór umotywowany został przede wszystkim popularnością tego narzędzia oraz liczbą parametrów, które mogą sprawić, że ChatGPT ma przewagę nad innymi narzędziami tego typu. ChatGPT -3 (i 3.5), z których nadal można korzystać bezpłatnie, wykorzystuje model GPT-3. Te wersje chatbota mają 175 miliardów parametrów. Z kolei najnowsza wersja ChatGPT – dostępna tylko z płatną subskrypcją – współpracuje z liczbą parametrów zawartą w GPT-4. Szacuje się, że liczba parametrów dla wersji ChatGPT-4 może osiągnąć wartość 100 bilionów. Konkretna liczba parametrów dla GPT-4 nie została oficjalnie jednak ujawniona przez OpenAI. Oznacza to, że wybór tego rozwiązania dla celów badawczych jest trafny.

Zastosowane metryki nie spełniają oczekiwań autorów artykułu. Przeprowadzona analiza danych zamieszczona w tabeli z wynikami eksperymentu nie daje pewności do wnioskowania. Dane dotyczące wybranych metryk wydają się generować losowe wyniki, trudne do interpretacji. Oczywiście temat metryk pozostaje nadal otwarty. Autorzy nie wykluczają powrotu w przyszłości do tego typu rozwiązań, jeśli pojawi się odpowiednie oprogramowanie. Wydaje się, że na chwilę obecną nie ma odpowiedniej metryki, która mogłaby spełnić oczekiwania postawione w problemach badawczych.

Autorzy artykułu zaproponowali własne kryteria do oceny samodzielności prac ewaluacyjnych wygenerowanych przez studentów. Kryteria te po iteracyjnej weryfikacji mogą stanowić materiał do rozwoju i do dalszych badań. Wydaje się, że zautomatyzowanie procesu sprawdzania poprawności wygenerowania przez studenta pracy ewaluacyjnej zostało osiągnięte. Trudność dotyczą badania samodzielności wygenerowanych prac ewaluacyjnych przez studentów. Temat ten będzie badany i rozwijany w kolejnych krokach eksperymentu.

## Wnioski

Opracowany model, procesy jego weryfikacji oraz przeprowadzona dyskusją wskazują na następujące wnioski:

1. Celem artykułu było opracowanie modelu procesowego budowy aplikacji do oceny prac studenckich
2. Zbudowano model, przeprowadzono procesy jego weryfikacji dla 154 prac studenckich i oceniono jego przydatność z zastosowaniem metryk
3. Takie podejście z jednej strony umożliwiło wstępną ocenę prac studenckich, które były weryfikowane przez prowadzącego a następnie przeprowadzenia egzaminu face to face dla weryfikacji opracowanego rozwiązania.
4. Realizacja tego typu eksperymentów pozwoli również na otwarcie drogi do zmian w obszarze zajęć dydaktycznych, które wykorzystują w sposób świadomy takich narzędzi, jak ChatGPT przez studentów. W tym obszarze nie ma zdefiniowanych reguł i kryteriów eksperymentów badawczych. Nie ma również wiedzy i doświadczenia odpowiedniego do kreowania nowych procesów w obszarze dydaktyki z użyciem ChatGPT określonym w problemach utylitarnych
5. Kierunki dalszych badań obejmują: uwzględnienie zróżnicowanych metryk do oceny prac studenckich oraz pytania wstępne dotyczące stanu wiedzy studenta przed przystąpieniem do egzaminu oraz w przypadku negatywnej oceny niedopuszczenie go do egzaminu.
6. Taka sytuacja staje się konieczna po przeprowadzeniu egzaminu ustnego, gdzie okazało się że studenci którzy otrzymali oceny niedostateczne nie rozumieją procesów danej dziedziny.
7. Przedstawiony w pracy model wymaga szczegółowego badania jego procesów, co nie miało miejsca w tym artykule. Wynikało to z traktowania tego artykułu, jako wprowadzającego do analizy możliwości wykorzystania modeli językowych do oceny prac studenckich.

## Bibliografia

[1] Villagrán, I., Hernández, R., Schuit, G., Neyem, A., Fuentes-Cimma, J., Miranda, C., ... & Varas, J. (2024). Implementing artificial intelligence in physiotherapy education: A case study on the use of large language models (LLM) to enhance feedback. IEEE Transactions on Learning Technologies.

[2] Meissner, R., Pögelt, A., Ihsberner, K., Grüttmüller, M., Tornack, S., Thor, A., ... & Hardt, W. (2024, October). LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation. In Frontiers in Education (Vol. 9, p. 1427502). Frontiers Media SA.

[3] Smolić, E., Pavelić, M., Boras, B., Mekterović, I., & Jagušt, T. (2024, May). LLM Generative AI and Students’ Exam Code Evaluation: Qualitative and Quantitative Analysis. In 2024 47th MIPRO ICT and Electronics Convention (MIPRO) (pp. 1261-1266). IEEE.

[4] Joy Kulangara, K. (2024). Designing and building a platform for teaching introductory programming supported by large language models.

[5] Jošt, G., Taneski, V., & Karakatič, S. (2024). The Impact of Large Language Models on Programming Education and Student Learning Outcomes. Applied Sciences, 14(10), 4115.

[6] Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., & Gu, X. (2024). Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design. IEEE Transactions on Learning Technologies.

[7] Telesko, R., & Wilke, G. (2024). TOWARDS A SUSTAINABLE OPERATING CONCEPT OF LLM GENERATED EXAMINATION SUPPORT IN HIGHER TECHNICAL EDUCATION. In ICERI2024 Proceedings (pp. 1064-1069). IATED.

[8] Agostini, D., & Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education. Intelligenza Artificiale, 18(1), 121-138.

[9] Huovinen, L. (2024). Assessing Usability of Large Language Models in Education.

[10] Yuan, B., & Hu, J. (2024). An Exploration of Higher Education Course Evaluation by Large Language Models. arXiv preprint arXiv:2411.02455.

~~-----------------------------------------------------------------------------------------------------------------------------~~

[11] Mao, J., Chen, B. & Liu, J.C. Generative Artificial Intelligence in Education and Its Implications for Assessment. TechTrends 68, 58–66 (2024). <https://doi.org/10.1007/s11528-023-00911-4>

[12] Baidoo-anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>

[13] Prajapati, J.B., Kumar, A., Singh, S. et al. Artificial intelligence-assisted generative pretrained transformers for applications of ChatGPT in higher education among graduates. SN Soc Sci 4, 19 (2024). <https://doi.org/10.1007/s43545-023-00818-0>

[14] Alier, Marc; García-Peñalvo, Francisco; Camba, Jorge D. (2024). "Generative Artificial Intelligence in Education: From Deceptive to Disruptive", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, issue Special issue on Generative Artificial Intelligence in Education, no. 5, pp. 5-14. <https://doi.org/10.9781/ijimai.2024.02.011>

[15] Baidoo-anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>

[16] Kisielewicz, A., (2017). *Sztuczna Inteligencja i Logika*. Warszawa, Wydawnictwo Naukowe PWN SA.

[17] Wodecki, A., (2021), *Sztuczna Inteligencja we współczesnych organizacjach*. Warszawa, Wydawnictwo Naukowe PWN SA.

[18] Flasiński, M., (2011), *Wstęp do Sztucznej Inteligencji*. Warszawa, Wydawnictwo Naukowe PWN SA.

[19] Rutkowski, L., (2012), *Metody i techniki sztucznej inteligencji*. Warszawa, Wydawnictwo Naukowe PWN SA.

[20] Carr, David F. (2023). *ChatGPT’s First Birthday is November 30: A Year in Review*, <https://www.similarweb.com/blog/insights/ai-news/chatgpt-birthday/> (dostęp: 15.11.2023).

[21] Vincent, James. (2016). *Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day*, <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist> (dostęp: 15.11.2016).

[22] Strony Parlament UE. (2023). *Akt ws. sztucznej inteligencji: pierwsze przepisy regulujące sztuczną inteligencję*, <https://www.europarl.europa.eu/topics/pl/article/20230601STO93804/akt-ws-sztucznej-inteligencji-pierwsze-przepisy-regulujace-ai> (dostęp: 13.06.2023).

[23] Vincent, James. (2016). *Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day*, <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist> (dostęp: 15.11.2016).

[24] Gates, B. (2023). *The Age of AI has begun*, <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun?WT.mc_id=20230321100000_Artificial-Intelligence_BG-EM_&WT.tsrc=BGEM> (dostęp: 21.03.2023).

[25] Vincent, J. (2022). *AI-generated answers temporarily banned on coding Q&A site Stack Overflow*, <https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers> (dostęp: 05.12.2022)

[26] OpenAI. (2022). *Aligning language models to follow instructions*, <https://openai.com/research/instruction-following> (dostęp: 27.01.2022)