

**Centre for Credible Artificial Intelligence
Warsaw University of Technology**

Rektorska 4, 5.02
00-614 Warsaw



Warszawa, 21 lutego 2026

Prof. dr hab. inż. Przemysław Biecek
Center for Credible AI
Politechnika Warszawska
Koszykowa 75/507
00-662 Warszawa

Recenzja rozprawy doktorskiej

Tytuł rozprawy: „Bridging the Generalization Gap: A Methodological Approach to Building and Benchmarking NLP Models”

Autor rozprawy: mgr inż. Michał Pietruszka

Promotorzy rozprawy:

dr hab. inż. Paweł Morawiecki, prof. IPI PAN

dr hab. Przemysław Spurek, prof. UJ

Dziedzina: Nauki Techniczne.

Dyscyplina: Informatyka Techniczna i Telekomunikacja.

Niniejsza recenzja oparta jest o otrzymaną rozprawę doktorską, dodatkowe wyjaśnienia o które prosiłem w piśmie z 31 stycznia z dnia 2026, a które uzyskałem 11 lutego 2026, oraz treść trzech artykułów stanowiących podstawę rozprawy, tj. “Sparsifying Transformer Models with Trainable Repre-

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa

Centre for Credible Artificial Intelligence Warsaw University of Technology

Rektorska 4, 5.02
00-614 Warsaw



sensation Pooling” z ACL 2022, “STable: Table Generation Framework for Encoder-Decoder Models” z EACL 2024 oraz “Can Models Help Us Create Better Models? Evaluating LLMs as Data Scientists” z EACL 2026.

W poniższej recenzji omawiam krótko rozprawę, każdy z trzech tych artykułów, ich połączenie jako podstawę do uzyskania stopnia doktora zgodnie z wymaganiami Ustawy - Prawo o szkolnictwie wyższym i nauce, oraz przedstawiam rekomendację dotyczącą przyjęcia rozprawy do dalszego postępowania, w szczególności do publicznej obrony.

Rozprawa nie ma klasycznej struktury monografii. Autor zdecydował się zgłosić jako osiągnięcie naukowe ciąg artykułów naukowych w których pełni wiodącą rolę równorzędnego pierwszego autora. W pierwszej wersji było to osiem artykułów ale po mojej prośbie o uszczegółowienie skupię się na 3 pracach w których rola pana Pietruszki była wiodąca a kontrybucje były istotne metodologicznie. Sama rozprawa ma strukturę krótkiego ogólnego omówienia procesu trenowania i weryfikacji modeli językowych, które stanowi ramę metodologiczną dla głównych wyników oraz krótkiego omówienia każdej z opublikowanych prac naukowych, które składają się na dorobek zgłoszony do uzyskania stopnia doktora.

Pierwsza ze zgłoszonych prac, „Sparsifying Transformer Models with Trainable Representation Pooling”, została opublikowana w 2022 roku na konferencji ACL jako tzw. long paper. Praca ma dwóch równorzędnych pierwszych autorów Michała Pietruszkę i Łukasza Borchmanna i była częścią rozprawy doktorskiej Łukasza Borchmanna obronionej na Politechnice Poznańskiej.

Praca prezentuje innowacyjną, w momencie publikacji, koncepcję rozrzedzania architektury transformera, umożliwiającą szybsze trenowanie i inferencję modeli językowych na dłuższym kontekście. W rozważanym zadaniu streszczania długich dokumentów (ang. summarization) długość kontekstu stanowi istotny problem; zaproponowane przez autorów artykułu podejście jest eleganckie i na tle innych pojawiających się wówczas propozycji jego

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa

Centre for Credible Artificial Intelligence Warsaw University of Technology

Rektorska 4, 5.02
00-614 Warsaw



rozwiązania bardzo skuteczne, czego dowodzą rozmaite przedstawione zestawienia.

Największe wyzwanie w ocenie tego wyniku dotyczyło określenia indywidualnego wkładu pana mgr. Pietruszki w sytuacji, gdy artykuł miał równorzędnych pierwszych autorów. Po zapoznaniu się z oświadczeniami autorów uznałem, że główną kontrybucją metodologiczną pana Pietruszki była propozycja operatora Top-K (Successive Halving Top-k operator), stanowiącego podstawę zaproponowanego mechanizmu rozrzedzania.

Wynik badawczy opisany w artykule oceniam jako oryginalny i wartościowy. Eliminacja tokenów z dalszego przetwarzania, zamiast ograniczania zakresu uwagi (ang. attention), to koncepcyjnie nowa i elegancka droga do redukcji złożoności obliczeniowej, szczególnie wartościowa po stronie dekodera. Zgodnie z moją wiedzą podejście to jest oryginalne i nie było wcześniej rozważane w literaturze. W artykule przedstawiono również rzetelne porównania z wieloma alternatywnymi metodami (Funnel, PoWER, LSH, Efficient Transformer) w spójnych warunkach eksperymentalnych, a przeprowadzone eksperymenty potwierdzają skuteczną redukcję czasu zarówno trenowania, jak i inferencji.

Pewnym ograniczeniem proponowanej metody jest jej relacja do technik wyjaśnialności i mechanicznej interpretowalności. Jest to kwestia, do której chciałbym, aby dyplomant odniósł się szerzej podczas obrony. Autorzy przedstawiają swoje rozwiązanie jako model trenowany end-to-end, jednak zastosowany mechanizm poolingu i rozrzedzania utrudnia określenie, które tokeny i w jaki sposób wpływają na wynik działania modelu. Ponadto zaproponowane rozwiązanie nie daje się wprost zastosować do względnego kodowania pozycyjnego (Relative Positional Encoding), będącego standardowym elementem współczesnych architektur dla modeli LLM. Ponieważ pojęcie benchmarkingu pojawia się już w tytule rozprawy, pogłębiona weryfikacja rozwiązania uwzględniająca możliwości jego wyjaśnienia i mechanicznej interpretacji jest szczególnie istotna i chciałbym dowiedzieć się, jakie możliwe kierunki rozwiązania tych problemów widzi autor rozprawy.

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa

Centre for Credible Artificial Intelligence Warsaw University of Technology

Rektorska 4, 5.02
00-614 Warsaw



Druga ze zgłoszonych prac "STable: Table Generation Framework for Encoder-Decoder Models" została opublikowana w roku 2024 na konferencji EACL jako tzw long paper. Praca ma ośmiu autorów z czego trzech równorzędnych pierwszych autorów, w tym mgr Michała Pietruszkę, dzisiaj już dr Łukasza Borchmanna i mgr Michał Turskiego.

Kluczowe rozwiązanie zaprezentowane w tym artykule dotyczy transformacji dokumentu bezpośrednio do formy ustrukturyzowanej tabeli. Tego typu ekstrakcja informacji do formy strukturalnej jest oryginalnym pomysłem, potencjalnie bardzo użytecznym i może ułatwić realizację rozmaitych zadań pomocniczych, takich jak np wyszukiwanie informacji lub wielokrokowe wnioskowanie. Taka struktura tworzy też interpretowalną warstwę pośrednią pomiędzy nieustrukturyzowanym dokumentem a konkretnymi zadaniami, tzw. down-stream tasks. Zbudowanie takiej uporządkowanej tabelarycznej reprezentacji jest jednak wysoce nietrywialne. W architekturze koder-dekoder do trenowania wymaga zaplanowania zarówno mechanizmu tworzenia dokumentu z tabeli jak i odtwarzania tabeli z dokumentu. Pierwszy krok rozwiązano stosując pomysłowe sposoby zamiany linearnego dokumentu na tabelę bez określonej kolejności wierszy (zastosowano tu rozważanie różnych możliwych permutacji) jak i syntezę komórek tabeli w sposób redukujący możliwość propagacji błędów.

Również i w przypadku tej pracy znacznym wyzwaniem było określenie na czym polegał indywidualny wkład autora. Po analizie złożonych oświadczeń zakładam, że w przypadku tej pracy ten wkład był istotniejszy i obejmował też zaproponowanie głównej koncepcji badań. Informacja o kontrybucjach jest jednak opisana na poziomie bardzo narzędziowym i mam nadzieję, że autor podczas obrony w bardziej precyzyjny sposób na czym polegał jego indywidualny oryginalny wkład.

Sam wynik opisany w artykule jest ciekawy, podobnie jak i zaproponowana seria rozwiązań pozwalających na wytrenowanie modelu generującego bezpośrednio tabelę. Dla wielu zadań upraszcza cały proces, eliminuje konieczność osobnych modeli dla każdego podzadania. Autorzy systematycznie

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa

Centre for Credible Artificial Intelligence Warsaw University of Technology

Rektorska 4, 5.02
00-614 Warsaw



cznie testują różne elementy proponowanego procesu, w tym typ permutacji, różne ograniczenia kolejności dekodowania, heurystyki wyboru komórek. Wykorzystanie trzech zbiorów danych o istotnym znaczeniu biznesowym (faktury, opisy składu produktów, wyciągi bankowe) pokazują realną użyteczność zaproponowanego rozwiązania.

To rozwiązanie ma kilka ograniczeń, ale o wielu z nich (jak np. nierówną skuteczność na podzadaniach czy konieczność określenia nagłówków tabeli) autorzy wprost i szczerze piszą w rozdziale Limitations. Otwartym tematem, który chciałbym by autor szerzej omówił podczas obrony, jest kwestia interpretowalności potencjalnych błędów generowania tabeli, w tym kwestia kolejności wierszy, która krótko jest poruszona w ostatnim rozdziale ale nie jest tam rozwinięta.

Trzecia ze zgłoszonych prac „Can Models Help Us Create Better Models? Evaluating LLMs as Data Scientists”, została opublikowana w 2026 roku na konferencji EACL jako tzw. long paper. Praca ma czterech autorów, z czego dwóch oznaczonych jako równorzędni pierwsi autorzy to mgr Michał Pietruszka i dr Łukasz Borchmann.

Praca dotyczy konstrukcji zbioru referencyjnego (benchmark) służącego ocenie zdolności systemów LLM do pełnienia roli automatycznego analityka danych poprzez przeprowadzanie inżynierii cech. Wygenerowane przekształcenia są opisane kodem w języku Python, który jest wykonywany w celu transformacji danych wejściowych, a jakość uzyskanych zmiennych oceniana jest przez skuteczność dodatkowego modelu zastępczego (proxy) z rodziny XGBoost.

Podobnie jak w przypadku poprzednich prac, wyzwaniem pozostaje precyzyjne określenie indywidualnego wkładu autora rozprawy doktorskiej. Na podstawie analizy złożonych oświadczeń zakładam, że w przypadku tej pracy wkład ten był wiodący i obejmował zarówno dobór zbiorów danych do testów, implementację znacznej części rozwiązania, jak i eksplorację uzyskanych wyników.

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa

Centre for Credible Artificial Intelligence Warsaw University of Technology

Rektorska 4, 5.02
00-614 Warsaw



Podobnie jak w poprzednich pracach, zarówno postawiony problem, jak i uzyskane wyniki są interesujące i dobrze wpisują się w coraz powszechniejsze oczekiwania dotyczące systemów agentowych w zadaniach wspomagania analizy danych, w tym w zadaniu automatycznej ekstrakcji cech.

Pomysł zastosowania modelu XGBoost jako modelu pomocniczego ma pewne zalety — jest to dobry model osiągający wysoką skuteczność przy domyślnych hiperparametrach — jednak rodzi też pewne zastrzeżenia w kontekście oceny inżynierii zmiennych. Techniki drzewiaste są z zasady niewrażliwe na jednowymiarowe monotonicznie transformacje zmiennych. Pomimo że autorzy rozważają skalowanie i normalizację cech, tego rodzaju przekształcenia nie mogą wpłynąć na skuteczność metod boostingowych opartych na drzewach decyzyjnych.

Przeprowadzona ewaluacja jest interesująca, jednak chciałbym, aby autor uzasadnił, w jaki sposób można zweryfikować, że silne modele LLM nie nauczyły się „na pamięć” transformacji stosowanych w przykładowych rozwiązaniach dostępnych na platformie Kaggle. Innymi słowy, czy obserwowana skuteczność wynika z rzeczywistych zdolności uogólniania modelu na nowe, dotychczas nieopublikowane dane, czy też jedynie z umiejętności odtworzenia przekształceń opisanych przez innych użytkowników Kaggle?

Druga wątpliwość dotyczy punktu odniesienia. Jednym z obecnie najsilniejszych modeli predykcyjnych dla danych tabelarycznych jest TabPFN. Interesującym pytaniem jest, czy predykcje uzyskane z jego użyciem ulegają dalszej poprawie po zastosowaniu transformacji zaproponowanych przez badaną metodę, i jak taki wynik kształtuje się w porównaniu z modelem TabPFN na danych przekształconych.

Trzecia uwaga dotyczy charakteru proponowanych transformacji. W pracy przedstawiono taksonomię ich typów, jednak bardzo korzystne byłoby zaprezentowanie kilku konkretnych przykładów — w szczególności tych prowadzących do największej poprawy wyników. Wizualizacja znalezionych przekształceń dla wybranych zmiennych pozwoliłaby znacznie lepiej zrozumieć rzeczywiste możliwości zaproponowanego rozwiązania i nadać wynikom bardziej intuicyjną interpretację.

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa

**Centre for Credible Artificial Intelligence
Warsaw University of Technology**

Rektorska 4, 5.02
00-614 Warsaw



Wszystkie trzy prace składające się na recenzowaną rozprawę są interesujące i proponują oryginalne rozwiązania postawionych problemów naukowych. Biorąc pod uwagę, że prezentowane metody zostały opublikowane w recenzowanych materiałach konferencyjnych o wysokim poziomie naukowym, stwierdzam, że rozprawa mgr. Michała Pietruszki spełnia warunki ustawowe stawiane rozprawom doktorskim i wnioskuję o dopuszczenie jej do publicznej obrony. Jednocześnie proszę autora, aby podczas obrony ustosunkował się do pytań i uwag przedstawionych w niniejszej recenzji.

Przemysław Biecek

**Politechnika
Warszawska**

ul. Koszykowa 75
00-662 Warszawa