Politechnika Wrocławska

Katedra Sztucznej Inteligencji

Wrocław-Singapur, 03 lutego 2026r.

Prof. dr hab. inż. Przemysław Kazienko
Katedra Sztucznej Inteligencji
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska
Email: kazienko@pwr.edu.pl
http://kazienko.eu

RECENZJA

rozprawy doktorskiej mgr Jana Ludziejewskiego pt. „Improving Performance of Mixture of Experts Large Language Models"

Rozprawę napisano pod kierunkiem dr hab. inż. Marka Cygana na Uniwersytecie Warszawskim i złożono we wrześniu 2025 r.

Recenzję wykonano na zlecenie Rady Naukowej Dyscypliny Informatyka Uniwersytetu Warszawskiego i dotyczy ona procesu o nadanie stopnia doktora w dziedzinie nauk ścisłych i przyrodniczych, w dyscyplinie informatyka.

I. Przedmiot, problematyka i charakter rozprawy

Tematyka rozprawy mieści się w szerokiej dziedzinie uczenia maszynowego (*machine learning*), która jest główną składową sztucznej inteligencji. Konkretniej, rozprawa dotyczy problemu uczenia zasadniczego lub wstępnego czyli pretreningu dla mieszaniny modeli (*Mixture of Experts – MoE*), które w tym przypadku są podklasą wielkich modeli językowych (*Large Language Models – LLMs*). Tematyka, bez wątpienia, zawiera się w dyscyplinie informatyka, w której pracę złożono.

Mieszanina modeli MoE sama w sobie nie jest nową koncepcją i można ją odnosić do idei *Adaptive Mixture of Local Experts* zaproponowanej w 1991r. m.in. przez Geoffreya Hintona. W innych niż NLP dziedzinach bywa ona praktycznie stosowana od wielu lat, np. w systemach rekomendacyjnych YouTube. Duża część pracy w istocie dotyczy optymalizacji wielokryterialnej, czyli wyznaczenia zależności między wieloma zmiennymi, czyli hiperparametrami opisującymi architekturę modelu, parametrami samego procesu pretreningu oraz miarami końcowymi tego procesu.

Ogólnie, rozprawa ma charakter koncepcyjno-eksperymentalny, tzn. zaproponowano w niej nowe sformułowania znanych wcześniej ogólnych problemów pretreningu, opracowano nowe metody i architektury, a następnie je zwalidowano eksperymentalnie na wybranych konfiguracjach, ustalonych scenariuszach i dla wybranych struktur.

W tematyce wielkich modeli językowych (i szerzej sztucznej inteligencji) czas przeprowadzenia badań odgrywa bardzo dużą rolę, tzn. nowe rozwiązania i

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

metody opracowywane w grupach badawczych często bardzo szybko stają się zdezaktualizowane lub przynajmniej wymagają szybkiego odniesienia się do innych metod, które właśnie – niestety często równolegle – są publikowane w literaturze światowej. Oznacza to, że wszelkie osiągnięcia, tj. artykuły naukowe powinno się oceniać w kontekście momentu ich publikacji. Z tego względu należy zdecydowanie pochwalić recenzowaną monografię, a raczej zbiór prac, które są jej podstawą. Nie mam wątpliwości, że w czasie ich powstawania były – i w dużej mierze nadal są – one w głównym nurcie badań. Warto tutaj zauważyć, że podstawa dysertacji powstała na przestrzeni ostatnich dwóch lat, tj. 2024-25.

Praca formalnie jest monografią, w której Autor zadeklarował „*Oświadczam, że niniejsza rozprawa doktorska została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.*" (s.4) Z drugiej strony niemal cały jej tekst to treść pochodząca z pięciu współautorskich artykułów naukowych napisanych przez łącznie niemal 20 autorów, w zasadzie bez żadnych zmian. Rodzi to oczywiste pytanie co jest wkładem Doktoranta, który to wkład jako recenzent powinienem przede wszystkim ocenić.

Liczba autorów w każdej z tych pięciu prac wynosi 10 i więcej osób. Trzy z nich (P1, P2, P4) zostały opublikowane na najlepszych konferencjach, tj. 2xICML oraz 1xNeurIPS, zaś dwie pozostałe to preprint oraz wersja odrzucona w systemie recenzji. Doktorant opisuje swój wkład w owe prace procentowo (pkt. 1.8) i wynosi on: 23% (P1, ICML'2024), 40% (P2, ICML'2025, poster), 40% (P3, wersja nieprzyjęta na ICLR), 10% (P4, NeurIPS), 15% (arXiv, przyjęta na warsztat przy ICLR'2024[1]), przy czym pierwszym autorem Doktorant jest w pierwszych trzech pracach (P1, P2, P3). Daje to średnią ok. 26% na artykuł. Trywializując, można by powiedzieć, że zaprezentowana monografia to wartość ok. 3-4 doktoratów i treść pracy będzie mogła być wykorzystana w jeszcze 2-3 monografiach doktorskich a każdy autor będzie mógł zadeklarować, że napisał ją samodzielnie. Uwzględniając powyższe dane, uważam, że najważniejsze z punktu widzenia osiągnięcia naukowego Doktoranta są prace P1 oraz P2 i dodatkowo P3, czyli treści rozdziałów 2-4.

Procentowy wkład jest niestety mało precyzyjny, gdyż może np. obejmować przygotowanie danych, samo uruchomienie eksperymentów lub przygotowanie wykresów i rysunków, co w opisywanych przypadkach nie byłoby bardzo wartościowym osiągnięciem naukowym. Nie wiemy więc kto postawił i zdefiniował problem naukowy, kto wymyślił opisane metody i kto je opisał (w tym wzory), kto zaplanował eksperymenty, czy kto miał zasadniczy wkład w tworzenie koncepcji i samej treści. Mamy wprawdzie przeniesioną z tych prac do monografii listę elementów, które można nazwać wkładem merytorycznym dla każdego rozdziału (*contribution*), co jest dobrą i powszechną praktyką. Nie wiemy jednak jaki jest w istocie udział Doktoranta w owych wkładach. Poziom części pracy, które są bardziej spoza owych artykułów, tj. wstępny rozdział 1, podsumowanie (rozdział 7

[1] W pracy nie ma o tym informacji.

HR EXCELLENCE IN RESEARCH

Evaluated by IEP INSTITUTIONAL EVALUATION PROGRAMME www.iep-qaa.org

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

Conclusion) a zwłaszcza bibliografia są generalnie dobre, ale nie tak dobre jak wysoki poziom współautorskich rozdziałów.

Jestem wielkim zwolennikiem pracy grupowej w nauce. Duże problemy i ambitne zadania w informatyce, zwłaszcza w eksperymentalnej sztucznej inteligencji zwykle muszą być realizowane przez zespoły, także dlatego, że postęp i konkurencja światowa w tej dziedzinie jest bardzo duża. Z drugiej jednak strony stopień naukowy jest nadawany konkretnej osobie za konkretne osiągnięcie.

Warto tutaj zauważyć, że treści trzech rozdziałów (2, 3, 5) zostały pozytywnie zrecenzowane i zatwierdzone do publikacji na najlepszych, bardzo konkurencyjnych i jednocześnie największych konferencjach naukowych, tj. 2xICML oraz 1xNeurIPS, co zdecydowanie uwiarygadnia ich wysoki poziom. Pierwsza z tych prac z ICML 2024 uzyskała już ponad 100 cytowań w bazie Scholar, zaś praca P5 także z 2024 roku (wersja z arXiv) – prawie 100, co jest świetnym wynikiem biorąc pod uwagę krótki czas ich dostępności. W bazie Scopus pierwsza z tych prac była cytowana tylko raz przez zewnętrznych autorów, druga wcale, ze względu na brak indeksowania, co jest jednocześnie bardzo wymownym dowodem na duże przesunięcie nauki (zwłaszcza informatyki i sztucznej inteligencji) w kierunku otwartych repozytoriów. W efekcie, wiele mierników bibliometrycznych, np. *impact factor* czasopism tracą na znaczeniu.

II. Hipotezy i cele pracy

Cele pracy zostały dość ogólnie i mało precyzyjnie sformułowane we wstępie, w ostatnich dwóch zdaniach w pkt. *1.5 Towards Generalized Scaling*:

> "*This thesis investigates what is the most efficient Transformer variant, and how to scale it under various constraints, introducing new theoretical frameworks and validating them with large-scale experiments. We aim to provide practical guidelines for designing MoE-based LLMs that maintain efficiency at scale.*"

Niestety nie podano tutaj, co Autor rozumie przez pojęcie *LLM efficiency*, aczkolwiek z dalszej części pracy można wywnioskować, że chodzi o wartość funkcji straty lub perplexity.

Formalnie nie sformułowano hipotez, ale osobiście nie uważam, że prace doktorskie i badania naukowe muszą posiadać sformułowane hipotezy i w tym konkretnym przypadku nie brakuje mi tego w ocenianej monografii.

III. Oryginalne osiągnięcia

Praca zawiera kilka bardzo wartościowych osiągnięć, w szczególności:

1. Analiza zjawiska skalowania dla mieszaniny ekspertów MoE, w której testowano różne poziomy zmniejszania wielkość składowych ekspertów

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

(granulacja) a także innych hiperparametrów (np. liczby ekspertów czy łącznej wielkości modelu) oraz poszukiwano 'optymalnych' wielkości dla zadanej wielkości zbioru wejściowego wyrażonego przez liczbę tokenów oraz budżetu (moc obliczeniowa razy czas). Kontynuacją tego było wyznaczanie 'optymalnej' liczby ekspertów dla zadanego budżetu, Rozdział 2.

2. Rozszerzona formuła prawa skalowania i jej eksperymentalna weryfikacja (w tym także inferencja) oraz wyznaczenie 'optymalnych' wartości hiperparametrów. W szczególności dotyczy to wyznaczenia dla zadanego budżetu (mocy obliczeniowej w czasie) oraz liczby ekspertów najlepszej łącznej liczby parametrów modelu (odpowiadającej wielkości dostępnej pamięci) oraz wielkości zbioru, tj. liczby tokenów (Tab. 3.1), Rozdział 3.
3. Zaproponowanie *relative learning rate*, który jest adaptacyjnie i niezależnie przyporządkowywany do różnych składowych modelu oraz wykazanie eksperymentalne, że takie podejście przyspiesza proces uczenia, Rozdział 4.
4. Zaproponowanie nowej architektury *Mixture of Tokens - MoT*, która tworzy mieszaninę tokenów i w efekcie zmienia mechanizm przypisywania ich do ekspertów MoE. Przeprowadzone eksperymenty wykazały zwiększenie szybkości pretreningu takiego modelu względem zwykłych transformerów i typowych modeli MoE, Rozdział 5.
5. Opracowanie nowej architektury MoE-Mamba, która łączy koncepcję inną niż transformer, tj. *State Space Models* (*Mamba*) z mieszaniną ekspertów MoE. Na podstawie przeprowadzonych eksperymentów można stwierdzić, efektywność uczenia podobną z Mamba osiągnięto w kilka razy mniejszej liczbie kroków, Rozdział 6.

Łącznie, powyższa lista osiągnięć sprawia, że recenzowana monografia stanowi nowe i ważne osiągnięcie naukowe, w którym zastosowano właściwe, eksperymentalne metody badawcze. Jednocześnie mieszczą się one w głównym i aktualnym nurcie badań nad sztuczną inteligencją.

IV. Ocena treści rozprawy

Rozprawa została napisana w języku angielskim. Jego poziom jest bardzo dobry a tekst zrozumiały. Treść składa się z 7 rozdziałów oraz bibliografii.

Wszystkie zasadnicze rozdziały (2-6) zawierają eksperymentalne badania umożliwiające walidację zaproponowanych rozwiązań, w tym np. wzorów opisujących prawo skalowania. Należy tutaj zwrócić uwagę, że chociaż użyte w nich modele nie są bardzo duże (mieszczą się na jednej karcie graficznej) a dane użyte do pretrenowania także nie są poziomu modeli produkcyjnych[2], to jednak ich przeprowadzenie wymagało sporej ilości zasobów zarówno obliczeniowych jak i

[2] Dla przykładu, dla potrzeb wytrenowania modeli PLLuM zgromadziliśmy korpus ponad 160 miliardów tokenów w języku polskim.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

związanych z przygotowaniem środowiska eksperymentalnego. Wszystkie kody źródłowe umożliwiające replikację badań opisanych w monografii zostały umieszczone w repozytorium GitHub, co jest bardzo dobra praktyką, często także oczekiwaną na najlepszych konferencjach i w czasopismach naukowych.

W każdym rozdziale wyszczególniono wkład danego rozdziału (czyli źródłowego artykułu) do nauki – co jest powszechna praktyką na dobrych konferencjach i w dobrych czasopismach naukowych[3] – jednak bez bliższej informacji o udziale w tym Doktoranta (prócz ogólnego określenia procentowego w pkt. 1.8).

Rozdział 1 to wprowadzenie, zawierające także elementy dotyczące motywacji dla podjęcia badań opisanych w kolejnych rozdziałach.

W drugim rozdziale (na podstawie pracy P1, ICML'2024) zaproponowano zmniejszanie wielkości ekspertów przy jednoczesnym proporcjonalnym zwiększaniu liczby ekspertów, do których kierowany jest token. Owa zmiana nazywana w pracy granularnością jest liczbą naturalną. Badano jaki ma ona wpływ na prawo skalowania. Jest ono związane z tym, że „optymalna" wielkość modelu wyrażona całkowitą liczbą parametrów jest zależna od wielkości zbioru uczącego, tj. od liczby tokenów w zbiorze uczącym. Oznacza to, że zwiększając zbiór uczący powinniśmy odpowiednio zwiększać wielkość modelu, aby zachować podobny poziom „wyuczenia" rozumiany jako pożądany poziom końcowy funkcji straty. W pracy wykonano szereg eksperymentów (patrz Tab2.3), których celem było zbadanie zarówno samego zjawiska skalowania jak i doboru „optymalnych" hiperparametrów. Najważniejszym wg mnie wnioskiem zawartym w tej pracy (który oczywiście wymagał wielu innych elementów) jest pkt. 2.5.4 (Tab. 2.1 i Fig. 2.6) oraz Tab. 2.2. Dodatkowo, w pkt. 2.8 można odnaleźć interesujące analizy dotyczące nieco szerszego kontekstu wykonanych eksperymentów. Wartościowy jest także Fig. 2.3b i Fig. 2.5, ale dotyczą one tylko trzech konkretnych wielkości zbioru danych. Całość badań w tym rozdziale należy ocenić wysoko.

W kolejnym, trzecim rozdziale zaproponowano rozszerzenie zjawiska skalowania na wersję z ustaloną wielkością pamięci, czyli w istocie o łączną wielkość modelu MoE, w postaci zdefiniowanego nieco innego prawa skalowania. W szczególności wyznaczono 'optymalną' wartość liczby ekspertów. W jasny sposób wyszczególniono 4 najciekawsze obserwacje wynikające z przeprowadzonych szerokich eksperymentów, a przede wszystkim wyznaczono jeden zestaw 'optymalnych' hiperparametrów (Tab. 3.3), dla których podano błąd odchylenia między formułą a uzyskanymi wynikami eksperymentalnymi i widać to dość dobrze na Fig. 3.5. Wartościowym elementem jest dodatkowa weryfikacja inferencji na 10 zadaniach wzorcowych (benchmarkowych) zawarta w pkt. 3.11. Całość jest bardzo dobrze napisana i na bardzo wysokim poziomie, klarowna i kompletna, biorąc pod uwagę to, że jest to artykuł konferencyjny z ograniczeniami na liczbę stron.

[3] Jest to także wymagane np. przez Elsevier w osobnej sekcji *highlights*.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

Rozdział 4 powstał z treści nieopublikowanego artykułu, który został wskazany jedynie poprzez listę autorów i tytuł[4]. Dodatkowa kwerenda w sieci pozwala zidentyfikować go jako odrzucony na ICLR 2024, dzięki czemu można przynajmniej określić mniej więcej okres wykonania badań. Całość tego rozdziału dotyczy – nie nowego – pomysłu adaptacji *learning rate* osobno dla różnych komponentów transformera: *embedding*, *unembedding*, *router* oraz *self-attention*, przy czym owej adaptacji dokonano osobno dla zwykłego transformera oraz MoE. Można to umiejscowić w tematyce często określanej jako *learning rate scheduling*. Badania wykonano na kilku konfiguracjach i generalnie raczej stosunkowo małych modelach. Wyniki są ciekawe, ale nie super nowatorskie, zaś całość została dobrze zaplanowana i przeprowadzona.

Rozdział 5 to treść artykułu zaakceptowanego na główną konferencję NeurIPS 2024, przy czym Doktorant jest piątym z dziesięciu autorów, trzech głównych to inne osoby a sam swój udział określił na 10%. Wprowadzono w nim nową koncepcję mieszaniny tokenów (*mixture of tokens*), tj. dodano nowy router tworzący odrębną mieszaninę tokenów dla każdego eksperta w architekturze MoE. Jest to bardzo ciekawa i nowa koncepcja. Przetestowano kilkanaście konfiguracji a wyniki wykazały większą szybkość uczenia dla zaproponowanych architektur. Podobnie jak w rozdziale 3 także tutaj przeprowadzono walidację inferencji na trzech zbiorach testowych.

Rozdział 6 jest dość tajemniczy ze względu na odniesienie do jego źródła. W pkt. 1.8 wskazano listę 10 autorów, tytuł i udział Doktoranta na 15%. W bibliografii z kolei mamy 5 autorów, ten sam tytuł i tylko rok 2024. Dodatkowa kwerenda w sieci ujawnia fakt opublikowania tej pracy jako preprint arXiv oraz akceptację jako poster na warsztatach *Workshop on Understanding of Foundation Models (ME-FoMo)* przy ICLR 2024, zaś obie wersje posiadają 10 tych samych autorów. Nie bardzo wiem jak należy rozumieć kwestię autorstwa tego rozdziału?

Sama koncepcja zaproponowana w tym rozdziale czyli połącznie *Space State Model* (Mamba) z MoE jest bardzo interesująca. Głównym wynikiem badań było wykazanie, że zaproponowana architektura szybciej się uczy (Fig. 6.1), w porównaniu do zwykłego transformera i do zwykłej Mamby. Nieco otwartym pytaniem jest tutaj jednak kwestia inferencji, gdyż zaletą Mamby jest właśnie większa szybkość inferencji a nie uczenia. Bardzo wartościowe są rozważania w pkt. 6.9 dotyczące kompresji historii do skończonego stanu ukrytego w Mambie a co za tym idzie być może ograniczonej zdolności do kopiowania tokenów, co z kolei może mieć przełożenie na mniejsze zdolności do uczenia przez kontekst (*in-context learning*). Obserwacja ta nie została jednak bliżej zbadana. Nieco podobnie do pozostałych rozdziałów pewien niedosyt budzi brak głębszego porównania do innych alternatywnych architektur znanych z literatury. Można więc ten rozdział traktować jako świetny punkt wyjścia do dalszych badań, w tym także dotyczących efektywności wnioskowania; perplexity świadczy o tym, że model jest pewien tego

[4] Dlaczego nie podano żadnego opisu bibliograficznego w bibliografii oraz odpowiedniego do niego odesłania?

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

co generuje, co nie jest równoznaczne z tym, że dobrze generuje treści i że są to treści właściwe.

V. Problemy, pytania i uwagi dyskusyjne

Problemy i pytania

1. W pracy brakuje mi nieco szerszego spojrzenia na proces pretreningu wielkich modeli językowych (oraz MoE). Dotyczy to m.in. różnych hiperparametrów procesu uczenia i czynników na nie wpływających, takich jak dla przykładu *global* i *mini/micro batch size*, *accumulation steps*, *curriculum learning* i związana z tym jakość i różnorodność danych (jest to do pewnego stopnia zawarte w czynniku c – entropia wzorów dotyczących zjawiska skalowania, ale bez dyskusji na ten temat), rozkłady różnych języków naturalnych, rodzaj tokenizera, liczba epok (często tylko 1), długość kontekstu, zjawiska *spikes* (rzadko występujące, ale Autor doświadczył je w rozdz. 5), stronniczość (*bias*) ekspertów, precyzja reprezentacji i kwantyzacja, liczba bramek, liczba ekspertów przeznaczonych do bycia zawsze wybieranymi, hiperparametr zapewniający większe zbalansowanie przyporządkowania do ekspertów, nie tylko losowa inicjalizacja parametrów, liczba bloków transformera vs. szerokość warstw, itd. Nie wspominam tutaj nawet o jeszcze szerszym kontekście, czyli o jakości generowanych treści i czasie generowania a co za tym idzie koszcie inferencji[5]. Uwaga ta była by mniej istotna gdyby praca była złożona jako lista publikacji a nie monografia. W oczywisty sposób w artykułach konferencyjnych istnieje ograniczenie na liczbę stron (często także dotyczy to wielkości załączników), więc trudno w nich oczekiwać większych lub bardziej ogólnych wstępów, szerszych dyskusji, czy bardzo rozbudowanych badań. Monografia jednak dobrze, aby była nieco bardziej zwarta i spójna. Jeden, dość pobieżny rozdział 1, w małym stopniu wypełnia tę lukę.
2. W pracy duża jakość uczenia jest rozumiana jako mała strata (*loss*) lub *perplexity* (wtedy korzysta się ze zbioru testowego). Generalnie zgadzam się, podobnie jak większość badaczy, z tym, że ma to duży związek z jakością modelu w jego zastosowaniach, niemniej jednak czy jest to tożsame? Przykładowo, doświadczenia związane z kolejnością użycia (gorsze na początku, lepsze na końcu) w pretreningu danych o zadanej jakości pokazują, że ma to istotne znaczenie na efekt końcowy. Są to także nasze doświadczenia także z tworzenia modelu PLLuM. Może w ogóle to nie być widoczne w postaci różnej wartości funkcji straty, zaś entropia danych jest taka sama dla każdej kolejności. Oznacza to, że model może dobrze przewidywać następny token, ale może mieć

[5] W rozdz. 2 skupiono się jedynie na samym procesie pretreningu, ale w rozdz. 3 i 5 dokonano oceny wnioskowania na kilku benchmarkach, pkt. 3.11 i 5.6.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

większą tendencję do generowania bzdur, jeżeli uczenie dobrymi danymi jest na końcu. W związku z powyższym, rodzi się pytanie: czy wartość końcowa funkcji straty (lub perplexity) jest jedyną i najlepszą miarą jakości wyuczonego modelu (w pracy używano na to ogólnego pojęcia *performance*)?

3. Jak się ma idea granularności do koncepcji *multi-gate mixture of experts*. która była szeroko stosowana w problemach wielozadaniowych?
4. Dwa zasadnicze pytania dotyczące zjawiska skalowania (rozdz. 2) to:
   a. jak bardzo stabilne są uzyskane wyniki?
   b. jak są one różne od istniejących wtedy rozwiązań innych autorów; czy tylko to co zawarto w pkt. 2.10 czyli elementy porównania do Clark et al. 2022 było możliwe wtedy do osiągnięcia?

   Obie analizy dałyby wgląd w to, jak bardzo uniwersalne są uzyskane wyniki, zwłaszcza badając inne czynniki, które mogą na nie wpływać – patrz pierwsza uwaga dotycząca innych czynników.
5. Jakie głębsze wnioski można by wyciągnąć z analizy zadań downstream (pkt. 3.11), np. dlaczego w niektórych zadaniach rozważane modele MoE są gorsze pod względem perplexity od modeli gęstych? Czy ta obserwacja nie jest sprzeczna z kategrycznym stwierdzeniem w tytule i treści pkt. 2.6.3 *MoE is Always More Efficient*?
6. Dlaczego nie porównano między sobą wszystkich architektur opisanych w pracy, w tym np. *MoE-Mamba* z *Mixture of Tokens*?
7. Jak się mają wykazane względem *Mamby* zyski w szybkości uczenia *MoE-Mamby* do rzeczywistych zysków na koszcie pretreningu?
8. Dlaczego nie podano żadnego opisu bibliograficznego w bibliografii oraz odpowiedniego do niego odesłania związanego z treścią rozdziału 4?
9. Wprowadzenie różnych nowych hiperparametrów do procesu pretrenowania poprawia szybkość uczenia, tj. szybsze osiągnięcie określonego poziomu straty. Jednak otwartym pytaniem jest, czy wynika to bezpośrednio z nowej koncepcji czy może większej elastyczności całego procesu, co jest spowodowane dodaniem nowego hiperparametru czyli kolejnego stopnia swobody?
10. Dlaczego wiele rzeczywistych modeli nie spełnia optymalnych wartości uzyskanych w pracy w tym np. liczby i wielkości ekspertów?

Uwagi dyskusyjne

11. Stwierdzenie „*MoE is **Always** More Efficient*" jest chyba kontrowersyjne, gdyż Doktorant może to twierdzić jedynie w ramach eksperymentów, które przeprowadził i ich założeń (nie wymienionych wprost), aby tak kategorycznie uogólniać swoje wyniki.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

12. Jak różne strategie adaptacji *learning rate* dla różnych komponentów mogłyby wpłynąć na poprawę szybkości uczenia (rozdz. 4)? Jaki wpływ na to ma charakterystyka *„mieszacza tokenów"*, który także się uczy?

Chciałbym podkreślić, że powyższe pytania i uwagi nie są świadectwem niskiej jakości rozprawy, ale w dużej mierze inspiracją wynikającą z dojrzałości zaprezentowanych analiz, z których, z kolei, mogą wynikać dalsze badania.

Uwagi formalne i redakcyjne

13. W streszczeniu zupełnie pomieszano numery rozdziałów z ich zawartością, tzn. zniknął w istocie rozdz. 1 a odniesienia do kolejnych powinny mieć o 1 większe numery, czyli *Chapter 1* to w istocie *Chapter 2*, itd.
14. W liście prac, na podstawie których powstała monografia (pkt. 1.8), brakuje części opisów dotyczących np. zakresu stron czy miejsca opublikowania preprintu P3 (OpenReview, artykuł wycofany przez autorów z ICLR 2025 po negatywnych recenzjach).
15. Wiele opisów bibliograficznych w sekcji Bibliography, która nie została umieszczona w spisie treści) nie posiada ważnych elementów takich jak np. rok – patrz praca Autora „Ludziejewski J., et al. „Scaling laws…" (czyli P1!!!), numery stron (bardzo wiele pozycji), miejsce wydania (np. Pióro 2024 – praca P5!!!, Zhao 2023a/b, Zhou 20222/23, Zhu 2024, Zoph 2022b i wiele innych), czy wydawca (nie jest to jednak zawsze oczekiwane pole).
16. Pewna liczba błędów językowych i redakcyjnych, np. brak odesłania w podpisie *Figure 1.2 (Figure from ?)*, przy czym sam podpis jest także zdecydowanie niewystarczający. Podobnie s. 63.
17. Dobra praktyka naukowa zakłada klarowność opisu, co także dotyczy wzorów, dla których dobrze jest wyjaśnić wszystkie występujące w nim symbole - patrz (1.1), np. nie jest pewne czy L to *loss*? D to wg opisu *dataset* czyli zbiór tokenów (bardziej lista?), czy może raczej *liczność* tej listy? Czym są α i β? Oczywiście można się tego domyślać, ale to nie jest dobry sposób komunikacji. Na rysunku Fig. 1.1 także mamy różne symbole np. L (loss?), $C_{\min}$, czyli *computing power* w petaflopdays? Dodajmy, że jest to w części przygotowanej przez Doktoranta a nie przeniesionej wprost ze współautorskich prac, co dodatkowo budzi pewne wątpliwości co wkładu Doktoranta do reszty treści przeniesionej ze współautorskich prac, gdzie nie ma takich niedociągnięć.
18. Niepotrzebna, pusta strona 102.
19. Mamy pkt. 1.4.1 a nie mamy punktu 1.4.2, co oznacza, że jest tylko jeden podpunkt w pkt. 1.4.
20. Tab. 6.5 jest nieco zniekształcona.

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434

Politechnika Wrocławska

Katedra Sztucznej Inteligencji

VI. Podsumowanie i ocena rozprawy

Podsumowując, należy stwierdzić, że rozprawa jest bardzo ciekawa, dojrzała i dotyczy ważnej, użytecznej i różnorodnej tematyki pretreningu modeli typu mieszaniana ekspertów. Bez wątpienia stanowi istotny wkład w dyscyplinę informatyka w dziedzinie nauk ścisłych i przyrodniczych. Osiągnięcia wymienione powyżej w pkt. III są nowe i wartościowe, co dotyczy także wysokiej ogólnej oceny szczegółowo zawartej w pkt. IV.

Całość świadczy bardzo dobrze o dobranym warsztacie badawczym i sporych umiejętnościach Doktoranta. Treść jest metodologiczne poprawna, na wysokim poziomie i mieści się w aktualnych kierunkach badań na świecie. Wyniki zostały dodatkowo zweryfikowane za pomocą trzech (a właściwie czterech) publikacji na najlepszych konferencjach naukowych. Praca ma bardzo duże znaczenie praktyczne, gdyż procesy pretreningu wielkich modeli językowych są bardzo kosztowne, więc ich poprawa jest bardzo istotna.

Pomimo pewnych niedopowiedzeń dotyczących wkładu Autora, dość szeroki zakres rozprawy, łącznie bardzo rozległe badania eksperymentalne, nowatorskość oraz poziom i dojrzałość badań moim zdaniem same się bronią, bez względu na szczegółową interpretację owego wkładu.

W związku z powyższym stwierdzam, że opiniowana rozprawa doktorska mgr Jana Ludziejewskiego spełnia wymagania stawiane w obowiązujących przepisach ustawy o stopniu naukowym doktora i wnoszę o dopuszczenie jego Autora do publicznej obrony.

Jednocześnie biorąc pod uwagę aktualność, nowatorskość oraz wysoki poziom przedstawionych badań nie wykluczam rozważenia wyróżnienia rozprawy. Jako jeden z warunków dalszego poparcia takiego wniosku stawiam m.in. kwestię pozytywnego wyjaśnienia wkładu Autora w prace będące podstawą głównych rozdziałów monografii.

HR EXCELLENCE IN RESEARCH

Evaluated by IEP INSTITUTIONAL EVALUATION PROGRAMME www.iep-qaa.org

Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji
Katedra Sztucznej Inteligencji
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
T: +48 71 320 24 54
www.pwr.edu.pl
ai.pwr.edu.pl
sekretariat.k46.wit@pwr.edu.pl
REGON: 000001614
NIP: 896-000-58-51
Nr konta:
37 1090 2402 0000 0006 1000 0434