Nr rejestracyjny: 2023/51/B/ST6/02914; Kierownik projektu: dr Łukasz Paweł Kuciński

Dopasowania dużych modeli językowych poprzez debatę oraz uczenie ze wzmocnieniem

Łukasz Kuciński

Sztuczna inteligencja (AI) stała się częścią naszego codziennego życia, głównie dzięki postępom w dużych modelach językowych (LLM) takich jak GPT. Modele te, które pomagają w zadaniach od poprawy produktywności po sterowanie robotami, mają głębokie implikacje dla społeczeństwa, biznesu i zarządzania. W związku z tym ich rozwój i użytkowanie są coraz bardziej monitorowane i regulowane, co widać na przykładzie inicjatyw takich jak Akt o AI UE czy Rozporządzenie Prezydenta USA Joe Bidena dotyczące AI.

Kluczowym aspektem rozwoju AI jest zapewnienie, aby te modele były dopasowane, tzn. działały zgodne z ludzkimi wartościami i intencjami. Tradycyjne metody polegają na szkoleniu tych modeli na podstawie preferencji ludzkich. Jednak nowe i obiecujące podejście polega na wykorzystaniu debat. W tej metodzie modele są szkolone do argumentowania i obrony swoich racji, angażowania się w dyskusję i przekonywania sędziów do swojego stanowiska. Format debaty nie tylko pomaga w umiejętności dotarcia do prawdy, ale także rozwija metody AI w zakresie zdolności formułowania złożonych argumentów oraz logicznego rozumowania.

Nasz projekt ma na celu wykorzystanie formatu debaty do szkolenia LLM. Inspirujemy się niedawnymi postępami w analizie debat, uczenia modeli oraz projektowania curriculum nauczania. Nasza metoda obejmuje ramy "nauczyciel-uczeń", w których jeden model AI opowiadający uczniowi, biorze udział w debatach, a oddzielny system AI odpowiadający nauczycielowi, nadzoruje i dostosowuje proces szkolenia. Zakładamy, że to podejście doprowadzi do odkrycia różnorodnych zasad debaty i lepszego dostosowania LLM do ludzkich wartości i intencji.

Projekt będzie skutkował protokołem szkolenia opartym na debatach, publicznym zestaw transkryptów debat, wyuczonym modelem AI oraz kodem źródłowym. Oczekuje się, że wyniki zwiększą zdolność AI do dostarczania dokładnych, niezawodnych odpowiedzi, tym samym poprawiając nasze zrozumienie świata i wspomagając podejmowanie decyzji. To badanie, skierowane na czołowe konferencje AI, może oznaczać znaczący krok naprzód w rozwoju wyrównania AI.