Ewaluacja dużych modeli języka

Filip Graliński, specjalność Sztuczna Inteligencja

1. Charakterystyka obszaru badawczego

   W ostatnim czasie nastąpił olbrzymi postęp w zakresie dużych modeli języka (ang. *Large Language Models, LLMs*). Duże modele języka mają coraz większy wpływ nie tylko na życie gospodarcze, lecz również społeczne. Przejawia się to przede wszystkim w postaci usług opartych na dużych modelach języka (np. ChatGPT), z drugiej strony pojawiają się coraz wydajniejsze modele otwartoźródłowe (np. Llama, Mistral). Szybkiemu postępowi w zakresie technik modelowania języka nie zawsze towarzyszy dobre rozumienia mocnych i stron modeli języka w szczególności w takich aspektach jak rozumowanie czy prawdomówność. W związku z tym proponowany obszar koncentruje się na zagadnieniach ewaluacji modeli języka – zarówno w zakresie metod ewaluacji opartych na algorytmach, ewaluacji manualnej czy w końcu ewaluacji przy wykorzystaniu samych modeli języka. Głównym celem jest znalezienie takich przykładów testowych, które pokazują (być może subtelne) słabości obecnych modeli języka. W ramach prac opracowane i zaimplementowane zostaną zarówno nowe sposoby ewaluacji, jak i nowe zbiory ewaluacyjne.

2. Motywacja

   Ewaluacja dużych model języka jest bardzo ważnym i wciąż niewyeskploatowanym polem badawczym. Nie wymaga również dostępu do olbrzymich mocy obliczeniowych, co jest konieczne w przypadku uczenia dużych modeli języka.

3. Obecny poziom badań i możliwości finansowania

   Ewaluacja modeli języka nie wymaga tak dużych zasobów obliczeniowych jak uczenie modeli. W związku z tym badania będą mogły być realizowane w ramach badań własnych pracownika bądź częściowo w ramach prac Centrum Sztucznej Inteligencji UAM.

4. Tematyka badawcza

   - zastosowanie zagadek logicznych w ewaluacji dużych modeli języka
   - wykrywanie błędów logicznych za pomocą dużych modeli języka
   - badanie zdolności reagowania dużych modeli języka na zmiany wiedzy
   - wykrywanie fałszywych wiadomości (fake news) za pomocą dużych modeli języka

5. Wymagania odnośnie członków projektu

   nd.

6. Literatura

   [1] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.
   [2] Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., ... & Li, B. (2023). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv preprint arXiv:2306.11698.
   [3] Gralinski, F., Jaworski, R., Borchmann, Ł., & Wierzchon, P. (2016). Gonito. net–open platform for research competition, cooperation and reproducibility. In Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language (pp. 13-20).