Ethics for NLP: SS 2024 **(Due: 03 July 2024, 11:59pm)**

Homework 3

Name: Julian Steiner *Matriculation No.:* 2669944

Problem 3

The other reason could be possible data contamination. It could be that the data from the sentiment analysis benchmark is now part of the training data for the newer version of the large language model.

To find out which reason behind the improvement is more likely we have two options. First we could check the "LLM Contamination Index". This is a database of contamination evidences for LMs. If we do not find any information about the sentiment analysis benchmark data set and a possible contamination, it would still be possible that we create some new sentiment analysis data and check if the newer version shows substiantially substantially lower performance on this new data. If yes, there might have been data contamination.

Problem 4

The common issued is called "Hallucination". Hallucination occurs when an AI model generates false information and present its as a fact. The three possible examples in the text are:

- Incorrect author assignment for publications. The response of the LLM-generated texts claims Iryna Gurevych has published the following works: "Text Classification and Clustering" (2006), "Natural Language Processing and Information Retrieval" (2011), "Machine Learning for Text Analysis" (2017). But she was not involved in any of these books as an author. No reference to this could be found on her profile page at TU Darmstadt under the publications[1] section.
- The AI-generated text also contains incorrect information in the research contributions of Iryna Gurevych. It is most likely true that Iryna Gurevych has made a significant impact on her field of research through her many publications in this field. But it is not true that she developed the concept of "topic modelling" for text analysis and designed the "Latent Dirichlet Allocation" (LDA) algorithm for topic modelling. The "Latent Dirichlet Allocation" algorithm was designed by David M. Blei, Andrew Y. Ng, Michael I. Jordan and published in the Year 2003[2]. Iryna Gurevych used this algorithm in two of her works[3][4]. Additionally, I could not find that she designed the concept for topic modelling for text analysis. However, she has published a few papers related to topic modelling[5][6][7].
- Another "invented" content of the AI is the "Gurevych model" for named entity recognition. I could not find any evidence that this model exists.

[1] TU Darmstadt profile of Iryna Gurevych
[2] Latent Dirichlet Allocation, David M. Blei, Andrew Y. Ng, Michael I. Jordan (2003)
[3] Unsupervised Latent Dirichlet Allocation for supervised question classification (2018)
[4] Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project (2015)
[5] Extrinsic Evaluation of Topic Models on Unknown Corpora (2015)
[6] Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project (2015)
[7] Focusing Knowledge-based Graph Argument Mining via Topic Modeling (2021)