BULGARIAN SQUAD AND RACE QUESTION-ANSWERING DATASETS

Simeon Monov, Detelinka Trifonova, Nikolay Pavlov, Andrey Nikolov

Abstract. *Currently multiple question-answering (QA) datasets exist, and SQuAD and RACE are very important for evaluating different LLM models on QA tasks and are also used for fine-tuning smaller models to perform question answering. These datasets are available in English language only.*

In this work we translate these models to Bulgarian language and evaluate their performance on both QA tasks and fine-tuning of Bulgarian language models. We present the results of our findings.

Key words: NLP, Bulgarian dataset, question answering, LLM.

Acknowledgments

This paper is partially supported by project MUPD23-FMI-009 of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

Simeon Monov[1], Detelinka Trifonova[2], Nikolay Pavlov[3], Andrey Nikolov[4],
[1, 2, 3, 4] Paisii Hilendarski University of Plovdiv,
Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4003 Plovdiv, Bulgaria
Corresponding author: smonov@uni-plovdiv.bg