Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Aleksandr Rubashevskii

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Efficient active learning and fact-checking via predictive uncertainty quantification for natural language processing

Supervisor: Assistant Professor Maxim Panov

Name of the Reviewer: Mikhail Beliaev

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	**Date: 17-12-2024**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

This thesis focuses on the problem of efficient uncertainty estimation for natural language processing applications. One of the tasks is related to batch active learning, where uncertainty quantification is used to select a set of samples for annotation, augment the training set, and improve the quality of the trained model. Another task is related to fact-checking using uncertainty estimation, where the assessment of the correctness of the generated information is based on the uncertainty values of the generated model.

There are several main contributions of the thesis. First, it is a batch active learning method that selects many samples for annotation in a single step based on the

informativeness and diversity of the samples taken. The proposed method is based on an approximation of the BatchBALD method using the BALD criterion and pairwise mutual information between samples. The proposed method is significantly faster than the original one and shows leading performance on multiple domains by avoiding duplicate sampling.

Second, it presents a fact-checking method using uncertainty quantification that focuses only on the uncertainty of the generated statement and does not consider the uncertainty of the stylistic features of the generating model. The proposed method outperforms alternatives for fact-checking of generated biographies for various models and multiple languages. An important addition to the developed method is the design and implementation of a general pipeline for fact-checking of white-box language models through uncertainty estimation.

Third, multi-stage fact-checking annotation methodology is proposed, as well as corresponding annotated dataset for generations of ChatGPT model. Moreover, the proposed methodology can be transferred to any other language model.

The presented work includes publications at A/A* level conferences: SIAM SDM, ACL Findings, EMNLP Demo, EMNLP Findings, which also confirms the significance of the presented research and obtained results. I consider the work done to be of high quality and recommend it for defense.

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*