FACULTY
OF SOCIAL SCIENCES
Charles University

Diploma Thesis Evaluation Form

Author: Soňa Milová

Title: Failure Modes of Large Language Modes

Programme/year: ISSA, 22/23

Author of Evaluation (supervisor/~~external assessor~~): Petr Spelda

Criteria	Definition	Maximum	Points
Major Criteria			
	Research question, definition of objectives	**10**	9
	Theoretical/conceptual framework	**30**	28
	Methodology, analysis, argument	**40**	38
Total		***80***	75
Minor Criteria			
	Sources	**10**	10
	Style	**5**	5
	Formal requirements	**5**	5
Total		***20***	20
TOTAL		***100***	95

Evaluation

Major criteria:

The dissertation analyzes several deficiencies of large language models (LLMs) that have security/safety and generally societal implications. The state-of-the-art in studying the risk surface of LLMs is captured well and its social impacts explained clearly. The strongest part of the dissertation is its analysis of alignment methods, using reinforcement learning-based approaches as the main case. Limits of alignment methods with respect to previously analyzed deficiencies are identified correctly. Equally correct is the conclusion that the risks involved in developing LLMs do not have strictly technical solutions.

Minor criteria:

All minor criteria are observed in full.

Based on the anti-plagiarism software checks, it is formally confirmed that the submitted thesis is original and, to the best of my knowledge and belief, does not, in an ethically unacceptable manner, draw from the works of other authors.

Overall evaluation:

The analysis offered in the dissertation is timely, considering the emerging debate on safety/security of LLMs. All presented arguments are well informed and rely on a body of relevant literature. Finally, it needs to be emphasised that by correctly combining technical content with security debates, the dissertation represents a well above average analysis.

Suggested grade: A

Signature

Mgr. Petr Špelda, Ph.D.

Digitálně podepsal Mgr. Petr Špelda, Ph.D.
Datum: 2023.06.14 15:37:19 +03'00'