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): Vit Stritecky

Criteria	Definition	Maximum	Points
Major Criteria			
	Research question, definition of objectives	**10**	9
	Theoretical/conceptua l framework	**30**	28
	Methodology, analysis, argument	**40**	37
Total		***80***	74
Minor Criteria			
	Sources	**10**	10
	Style	**5**	5
	Formal requirements	**5**	5
Total		***20***	20
TOTAL		***100***	94

Evaluation

Major criteria:

The dissertation deals with social implications of generative artificial intelligence. The selected cases very well illustrate risks involved in developing capable language models. The literature selected to support the arguments about harms that can be realized with language models is relevant and represents a technically informed understanding of the issue. The critical analysis of alignment methods, the safety procedures used during the development, is much appreciated and correct. The conclusion about techno-solutionism skillfully utilizes the preceding analyses of risks and available safety methods.

Minor criteria:

There are no formal issues.

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:

This is an excellent dissertation that deals with a pressing security issue from a technological point of view. The connection to discussions in security studies is realized well and the conclusions offer important insights into risks involved in developing generative artificial intelligence.

FACULTY
OF SOCIAL SCIENCES
Charles University

Suggested grade: A

Signature

PhDr. Vít Střítecký, M.Phil., Ph.D.

Digitálně podepsal PhDr. Vít Střítecký, M.Phil., Ph.D.
Datum: 2023.06.14 14:26:49 +02'00'