INSTITUTE OF COMPUTER SCIENCE

Academy of Sciences of the Czech Republic

Martin Holeňa

Pod Vodárenskou věží 2, 182 07 Praha 8, phone: +420 266052921, fax: +420 286585789, e-mail: martin@cs.cas.cz
web: www2.cs.cas.cz/~martin

Review of the master thesis "Evaluation of gender bias of Large Language Models in natural contexts" by

Paul Mouret

The submitted master thesis „ Evaluation of gender bias of Large Language Models in natural contexts " reports the contribution of its author to evaluation of gender bias in large language models, both his contribution to a novel benchmark dataset, and his contribution to the evaluation methodology.

The first part of the thesis recalls large language models and evaluating in their context, as well as gender bias and its evaluation, but also parsing, coreference resolution, low-rank adaptation, and debiasing. It also surveys related work.

The second part explains the methodology of the author's research contributions. It includes the methodology of creating a novel benchmark dataset, relying primarily on the generation of relevant contexts and filtering out factually gendered context, two methodological contributions to bias evaluation − the bias score and the relevance score, the methodology of stereotype analysis, relying mainly on stereotypical scores, and the methodology of locating bias.

Finally, experimental results are presented in the third part of the thesis. They cover dataset creation, bias evaluation, stereoptype analysis, bias location and model finetuning. I appreciate the complex and comprehensive way of experimental evaluation, especially in stereotype analysis. It is only a pitty that the experiments are based only on the mean length of a context, and not alaternatively also on its median length, which is more robust.

As to the thesis itself, I'm satisfied with both its structure and its content. When reading the thesis, I came across only three disappointing points:

a) Although SwiGLU can be simply defined as a particular composition of functions, and this one-line definition is indeed recalled in the 3rd equation on p. 16, the thesis introduces an unnecessary complicated half-page definition.

b) It has not been explained how the normalization in Figures 3.3-3.6 has been performed and what was the probability space with respect to which it has been performed.

c) No justification has been given for the fact that a word is counted as gendered even if it is used in its fully neutral meaning, as was illustrated on the example “chap” on p. 66.

I suggest classifying the submitted thesis with the grade 1.

30.8. 2024

Martin Holeňa