# Master thesis problem statement

University of Tartu

March 16, 2023

## 1 Problem:

Several studies showed that current NLP state-of-the-art models have shortcoming which is related to the model not paying attention to the input fully(only for some anchoring word), and correspondingly provide an answer from parametric knowledge(i.e. weights). This is undesirable behavior because input can be perturbed or factual knowledge can be altered, that predicted answer wouldn't make sense. In [Welbl et al., 2020] the question was perturbed, and in [Neeman et al., 2022] factual information in context was substituted with counterfactual knowledge. They observed that model instead of paying attention to the given input, rather rely on some anchoring words(in most of cases unrelated to an answer) to make predictions.

## 2 Goal:

Steer a model to pay attention to the input.

## 3 Solution:

We want to study effect of combination of methods such as data augmentation and parameter efficient fine-tuning. Data augmentation already proved its effectiveness in [Neeman et al., 2022].

### 3.1 Parameter efficient fine-tuning

This approach has advantage in contrast with fine-tuning, such as:

1. Preserving parametric knowledge, i.e. avoid catastrophic forgetting problem. Maybe it would be beneficial, but we don't know yet.

2. The performance is comparable with fine-tuning, and only small fraction of extra parameters are trained.

3. Modularity(only in case of prefix-tuning). Maybe it would be possible to combine various prefixes that were separately trained on factual, counterfactual, empty and random datasets. Potentially we could mix them up to get similar performance as a model that was fine-tuned on mix of datasets. The main benefit is faster training.

### 3.2 Concerns

1. Why we want to use parameter efficient fine-tuning? It is not clear, sounds like something we decided out of the blue. I did the additional literature review where I was looking for papers that have "parameter efficient fine-tuning"/"soft tuning" and "generalization". But this combination is not the point of those papers. Instead they are stressing that performance is comparable with only small fraction of trained parameters. However, no one mentioning that generalization increased.

2. Considering previous concern the word 'generalization' in my thesis topic name, would raise a question like 'Why is he talking about generalization improvements if the results are only comparable with [Neeman et al., 2022]?'

## 4 Contribution

- Study parameter efficient fine-tuning on model ability to disentangle models knowledge together with data augmentation.

# References

[Neeman et al., 2022] Neeman, E., Aharoni, R., Honovich, O., Choshen, L., Szpektor, I., and Abend, O. (2022). Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering.

[Welbl et al., 2020] Welbl, J., Minervini, P., Bartolo, M., Stenetorp, P., and Riedel, S. (2020). Undersensitivity in neural reading comprehension.