Investigating the vulnerabilities and effects of prompt-tuning on pre-trained language models

Motivation

- Pre-trained billion-parameter language models are expensive to fine-tune.
- Alternative method: **parameter-efficient tuning**, yielding similar results.

Research question

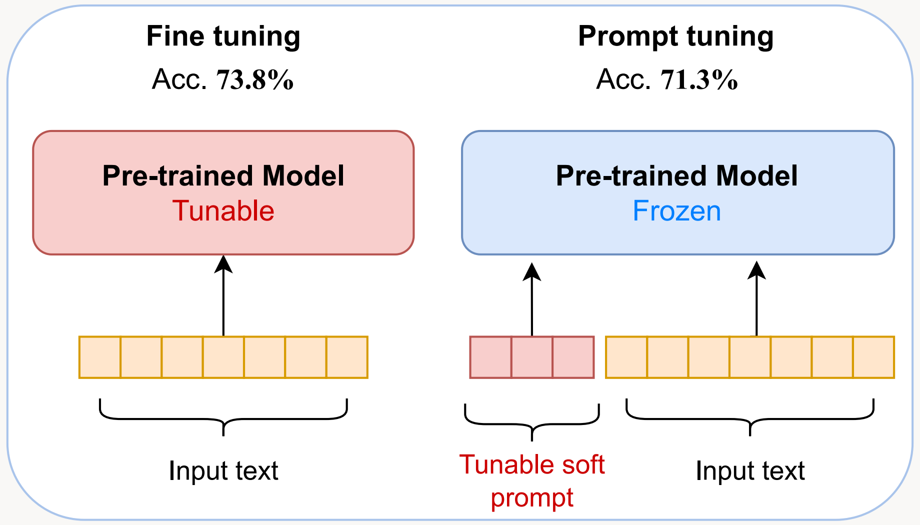
- Where do parameter-efficient tuning methods attribute their effectiveness?
- How robust are they against malicious attackers?

Method

Task: Sentiment classification (negative, neutral or positive)

Language model: BLOOM auto-regressive, decoder-only transformer

Visualisation: Embedding $e \in \mathbb{R}^{1024}$ of input's predicted sentiment (PCA, t-SNE projections)



Fine tuning
Acc. 73.8%

Prompt tuning
Acc. 71.3%

Pre-trained Model
Tunable

Pre-trained Model
Frozen

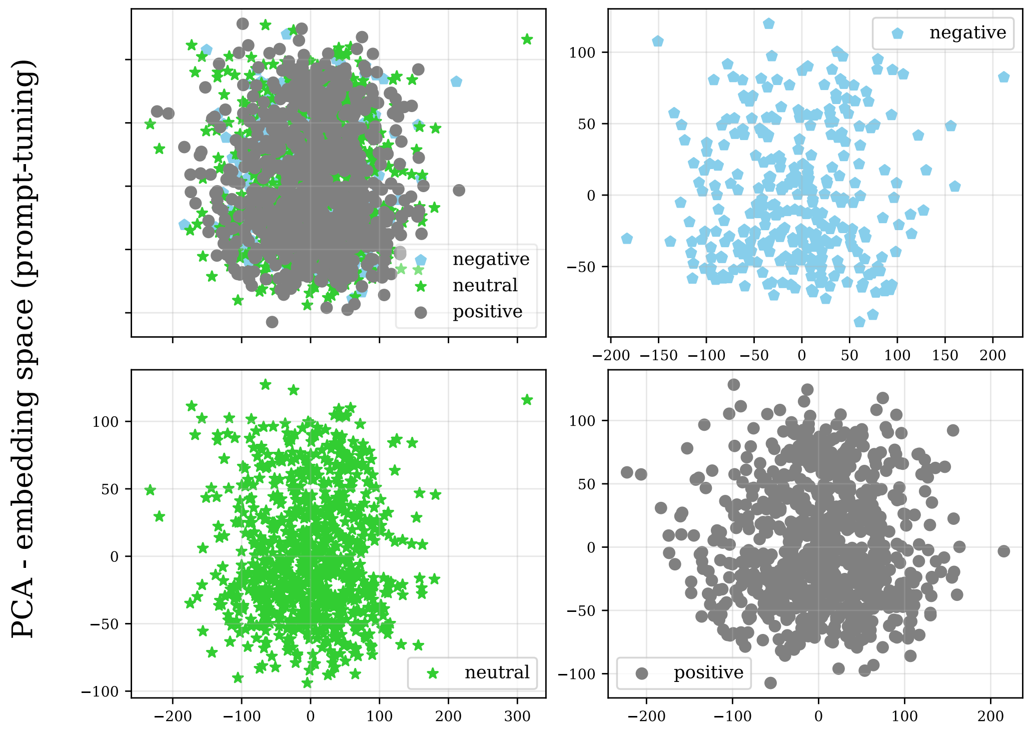
Input text

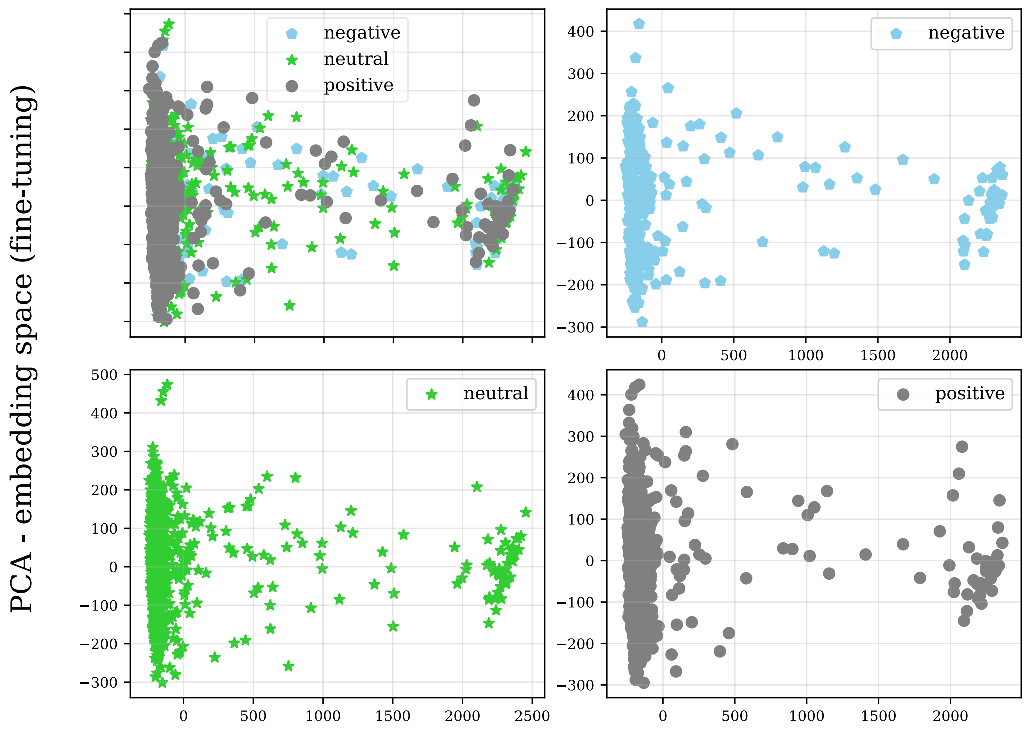
Tunable soft prompt

Input text

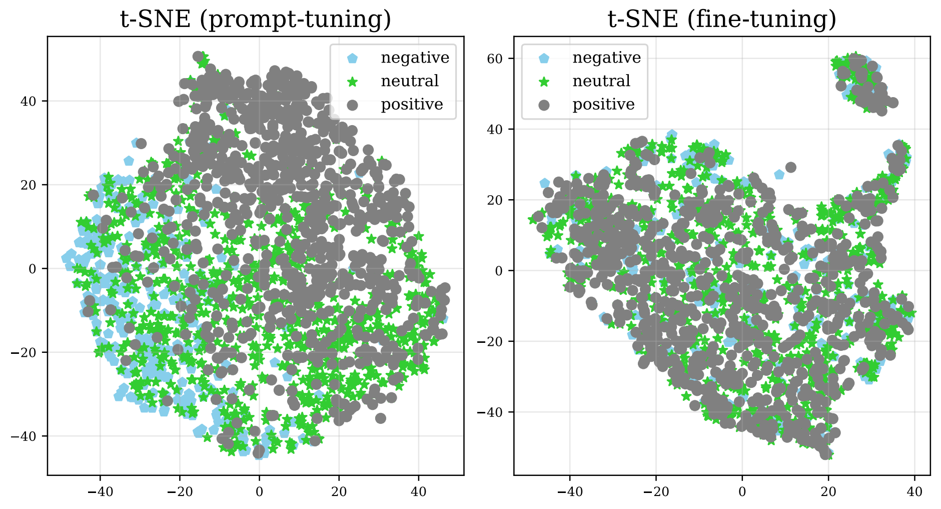
Results

- **Prompt-tuned sentiment classifier results to more dense clusters compared to fine-tuned classifier**





- **SImilar clustering observed by t-SNE**



Upcoming work

- Running TextFooler, black-box and semantics-preserving adversarial attack
- Interpretation of clusters w.r.t. the (benign and adversarial) input texts
- Implementing novel prompt-based attack to explore prompt-tuning vulnerabilities

Fotini Deligiannaki
fotini.deligiannaki@dlr.de
Institute for AI Safety & Security, DLR