# An investigation of the vulnerabilities and effects of prompt-tuning to pre-trained language models

Fotini Deligiannaki[1] and Arne Peter Raulf[1]

Institute for AI Safety and Security, German Aerospace Center, Germany
{fotini.deligiannaki,arne.raulf}@dlr.de

**Abstract.** State-of-the-art language models are showing immense potential in natural language understanding. This is achieved by models being trained on massive text data sets containing hundrets of GBs of text data and by utilizing massive hardware resources to train models with billions of parameters. These Large Language Models (LLMs) are therefore provided pre-trained to the AI community and can be used as a baseline for many application scenarios that want to deploy a tuned version to solve specific target tasks. With the increasing amount of parameters in these models, fine-tuning is a non-trivial and costly problem that makes their adaptability difficult. Additionally, the manual search for optimal text prompts to be used for querying LLMs in few- and zero-shot learning can have high complexity [2].

These concerns have led to the development of parameter-efficient methods for tuning LLMs as well as automating the search of prompts for solving the target tasks. In the example of prompt-tuning [1], a task-specific prompt is trained, containing less than 0.1% of the total parameters of the pre-trained LLM, while the model parameters are freezed. After being optimized in the continuous embedding space, it is concatenated with the input text during inference, guiding the model to solve the task at hand.

Since this method shows promising results comparable to the standard fine-tuning pipelines, fundamental questions regarding LLMs have been raised, relevant to this project. On the one hand, the robustness of prompt-tuned models against malicious adversarial scenarios needs to be investigated since this method opens up new potential attack vulnerabilities and the manipulation of trained prompts [6]. On the other hand, there is little understanding of what task- or model-specific features are entailed in the trained prompts and how they relate to the information encoded within the layers of transformer-based LLMs, although the interpretation of the optimized prompts has been instantiated [1].

In our poster presentation we want to highlight our preliminary research on these key questions. Within the scope of the project, we experiment by prompt-tuning open LLMs such as the T5 [4] and BLOOM [5] models on popular classification tasks. For investigating the first question, we built on the TextAttack [3] framework to attack both prompt-tuned and complete fine-tuned models, while also reviewing other recent backdoor

and adversarial attacks outside its scope. We furthermore look into the feature representations of benign and adversarial text inputs specifically targeting the latent space of the LLMs' encoder and decoder hidden layers with a distribution analysis and visualization tools, in order to gain a perspective on the second question.

Within our work we intend to bring further development into the overall usefulness and robustness of prompt-tuning by understanding the underlying mechanisms of LLMs, and hopefully accelerate the responsible application of parameter-efficient tuning methods.

## References

[1] Brian Lester, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. DOI: 10.18653/v1/2021.emnlp-main.243. URL: https://aclanthology.org/2021.emnlp-main.243.

[2] Xiao Liu et al. "P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks". In: *CoRR* abs/2110.07602 (2021). arXiv: 2110.07602. URL: https://arxiv.org/abs/2110.07602.

[3] John Morris et al. "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 119–126.

[4] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2020. arXiv: 1910.10683 [cs.LG].

[5] BigScience Workshop et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 2023. arXiv: 2211.05100 [cs.CL].

[6] Lei Xu et al. *Exploring the Universal Vulnerability of Prompt-based Learning Paradigm*. 2022. arXiv: 2204.05239 [cs.CL].