Attack vectors and security risks of large language models

Alexander Zwisler[1]

1- University Leipzig - Faculty for Mathematics and Computer Science

Abstract. Type your 100 words abstract here. Please do not modify the style of the paper. In particular, keep the text offsets to zero when possible (see above in this 'SeminarV2.tex' sample file). You may *slightly* modify it when necessary, but strictly respecting the margin requirements is mandatory (see the instructions to authors for more details).

1 Introduction

Large language models (LLMs) are currently transforming the way we interact with computer systems. With their unseen capabilities in natural language processing tasks they are a valueable system to integrate with a lot of existing systems. We currently see a fast paced integration of LLMs with pretty much all systems.

1.1 Large language models

What are large language models

2 Prompt Injections

This SeminarV2.tex file defines how to insert references, both for BiBTeX and non-BiBTeX users. Please read the instructions in this file.

3 LLM integrated systems

This SeminarV2.tex file defines how to insert references, both for BiBTeX and non-BiBTeX users. Please read the instructions in this file.

3.1 Indirect Prompt Injections

4 Conclusion

References

[1] S. Haykin, editor. *Unsupervised Adaptive Filtering vol.1 : Blind Source Separation*, John Willey ans Sons, New York, 2000.

[2] N. Delfosse and P. Loubaton, Adaptibe blind separation of sources: A deflation approach, *Signal Processing*, 45:59-83, Elsevier, 1995.