|  |
| --- |
| Can LLMs detect deception? An exploratory study from a linguistic cues perspective in a multilingual context  **Papantoniou Katerina**1#,, **Papadakos Panagiotis 2** 2\* and **Plexousakis Dimitris 3** 2  1 FORTH  # Presenting author: Panagiotis Papadakos email:papadako@ics.forth.gr  \* Corresponding author: Panagiotis Papadakos email:papadako@ics.forth.gr |

Abstract

Deception is prevalent in human societies and can have a major impact in people’s lives. However, humans are poor judges of deception due to human nature biases. This is perplexed even more when deception is considered in a cross-cultural context. Lately, there is an increased interest for automated Deep Learning (DL) deception detection classifiers, especially in the form of generative pre-trained transformers language models that include Large Language Models (LLMs). These models exhibit remarkable capabilities in various Natural Language Processing (NLP) tasks and can generate human-like text. Our work explores the ability of LLMs to detect deception in a cross-cultural setting, where language is used as a proxy for culture. Specifically, we examine zero- and one-shot learning settings, and fine-tune the models for the deception detection classification task. In addition, we explore their ability to generate content that imitates human deceptive verbal content and whether they are aware of the subtle differences of deceptive language across cultures [1]. In this context and in combination with our first objective we examine the introspection of the models for the generated text and also their awareness of the deceptiveness/truthfulness of the text generated by a different LLM. Finally, we also examine whether LLMs can exploit the current deception detection scientific knowledge in a RAG setting that is provided as context to the LLMs for improving their deception detection capabilities. The quality of explainability for all tasks was also explored via a frequency based text analysis, to model what knowledge these models claim that they hold about deception and their reasoning capabilities.

**REFERENCES**

[1] K. Papantoniou, P. Papadakos, T. Patkos, G. Flouris, I. Androutsopoulos, and D. Plexousakis. Deception detection in text and its relation to the cultural dimension of individualism/collectivism. In Natural Language Engineering (2022