A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation

2023-08-07 10:48:31

Authors

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun WU, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, Mustafa A. Mustafa, Saddek Bensalem

Abstract

Large Language Models (LLMs) have exploded a new heatwave of AI, for their ability to engage end-users in human-level conversations with detailed and articulate answers across many knowledge domains. In response to their fast adoption in many industrial applications, this survey concerns their safety and trustworthiness. First, we review known vulnerabilities of the LLMs, categorising them into inherent issues, intended attacks, and unintended bugs. Then, we consider if and how the Verification and Validation (V&V) techniques, which have been widely developed for traditional software and deep learning models such as convolutional neural networks, can be integrated and further extended throughout the lifecycle of the LLMs to provide rigorous analysis to the safety and trustworthiness of LLMs and their applications. Specifically, we consider four complementary techniques: falsification and evaluation, verification, runtime monitoring, and ethical use. Considering the fast development of LLMs, this survey does not intend to be complete (although it includes 300 references), especially when it comes to the applications of LLMs in various domains, but rather a collection of organised literature reviews and discussions to support the quick understanding of the safety and trustworthiness issues from the perspective of V&V.

Keywords

Large Language Models, Safety, Trustworthiness, Verification and Validation