Poster: SafeLLM: A Novel Framework for Safety Evaluation of Large Language Models: A Case Study of Offshore Wind Maintenance Planning

Connor Walker, Callum Rothon, Koorosh Aslansefat, Yiannis Papadopoulos, Nina Dethlefs

AURA CDT

University of Hull

Abstract *Offshore Wind's (OSW) contribution to the renewable energy sector is paramount as the demand for global net-zero heightens. Estimations currently suggest that up to 1150 GW; 25 % of the world's usage being supplied from OSW by 2050 (Smith 2023). Up to one third of levelised cost of energy (LCOE) is belonging to Operations and Maintenance (O&M), driving competition to lower costs. There are previous reported cases whereby more than 500 alarms occurring within a single day at Teesside Wind Farm. Thus, making optimal O&M planning without automated systems unfeasible and somewhat impossible. Increased pressure on maintenance strategies to ensure safety and dependability in progressively harsher environments demonstrates the need to integrate dependable Artificial Intelligence (AI) and Machine Learning (ML) systems. To eliminate these issues, we recently introduced a specialised conversational agent trained to interpret alarm sequences from Supervisory Control and Data Acquisition (SCADA) and recommend comprehensible repair actions (Walker et al. 2023). Acknowledging the recent advancements of Generative AI and Large Language Models (LLMs), we expand on this earlier work, fine tuning LLAMA (Touvron 2018), using available maintenance records from EDF Energy. An issue presented by LLMs is the risk of responses containing unsafe actions, or irrelevant hallucinated procedures. In response to these issues, this paper proposes SafeLLM as a novel framework for safety monitoring of OSW, combining previous work with additional safety layers. Generated responses being filtered ensures that raw responses of this agent do not endanger personnel and the environment. The algorithm represents such responses in embedding space to quantify dissimilarity to pre-defined unsafe concepts using the Empirical Cumulative Distribution Function (ECDF)-based statistical distance measure including Wasserstein and Anderson-Darling. In addition, a secondary layer to identify hallucination in responses has been added; exploiting probability distributions*

to analyse against stochastically generated sentences. Combining these layers, the paper aims to fine tune individual safety thresholds based on categorised concepts, providing a unique safety filter. Finally, a human-in-the-loop feature layer is discussed, using expert-approved predictions for online reinforcement learning, improving safety over time. The proposed framework has potential to utilise the O&M planning for OSW farms using state-of-the-art LLMs as well as equipping them with safety monitoring that can increase technology acceptance within the industry