Statistics and Actuarial Science Seminar

Title: Confidence in the large language models

Speaker: Yudi Pawitan (Karolinska Institutet)

Date: Thu 26th September 2024 at 3:00PM

Location: E0.32 (beside Pi restaurant)

Abstract: There is a growing literature on reasoning by large language models (LLMs), but the discussion on the uncertainty in their responses is still lacking. Our aim is to assess the extent of confidence that LLMs have in their answers and how it correlates with accuracy. Confidence is measured (i) qualitatively in terms of how persistent they are in keeping their answer when prompted to reconsider, and (ii) quantitatively in terms of self-reported confidence score. We investigated the performance of three LLMs – GPT4, GPT4-turbo and Mistral – on two benchmark sets of questions on causal judgement and formal fallacies and a set of probability and statistical puzzles and paradoxes. Although the LLMs showed significantly better performance than random guessing, there was a wide variability in both their qualitative and quantitative confidence. We observed positive correlations between qualitative confidence and accuracy, and between qualitative and quantitative confidence. However, the material effects of prompting on qualitative confidence and the overconfidence when explicitly asked for their level of confidence indicate that the current LLMs do not have any internally coherent sense of confidence.