Model Risk Management for LLMs: Model Monitoring

Abstract

As GenAI grows in capability, more and more banks are considering either introducing it into their workstreams in some fashion or expanding its usage where it is already present. Here we propose a selection of guidelines to fit into an overarching framework, intended to bring LLMs under the Model Risk Management umbrella, by considering the shared properties and shared risks they have in common with traditional models. We introduce some bespoke techniques and tools designed specifically to practically monitor LLM performance and discuss what their implementation might look like.

In particular, we consider four different areas in which we want to monitor LLM risk:

- Validity,
- Bias,
- Toxicity, and
- Privacy.

We look at three different types of drift, namely

- Data drift (changes in model inputs or in this case use prompts),
- Prediction drift (changes in the relationship between inputs and outputs), and
- Concept drift (changes in model outputs or in this case the LLM responses).
- which can be detrimental to model performance and so need to be monitored.

We will look to cover eight different techniques in the presentation, each of which can help mitigate against a combination of the risks and drifts.

Authors & Affiliations

Dr Ed Gallagher[1], Mr Jakob Kisiala[2]
[1]True North Partners, London, United Kingdom. [2]True North Partners, Edinburgh, United Kingdom