Model Risk Management for LLMs: Model Monitoring

Dr Ed Gallagher & Jakob Kisiala
True North Partners LLP | tnp.eu
ed.gallagher@tnp.eu, jakob.kisiala@tnp.eu

1) Introduction

As LLMs' use increases in banks, it is becoming increasingly important to manage the associated risks through a rigorous MRM framework. In this poster, based on a white paper TNP have produced, we explore the classification, governance, and monitoring of LLMs within a broader MRM context. We propose an ensemble of techniques designed to enable effective and transparent monitoring of LLMs, based on a white paper we have written.

2) Proposed Techniques

We propose an ensemble of techniques to use for LLM monitoring, each of which helps across a range of different dimensions and with different amounts of efficacy and specificity. These techniques include, but are not limited to:

- Calculating explainable quantitative metrics to assess response quality, ranging from standard ML metrics to ones based on translation or on sentence embeddings in a vector space,
- Monitoring and detecting drift with statistical measures, on both an input- and output-based level, as well as a separate conceptual level,
- Implicit text comparison, to evaluate key properties of responses without needing to look at granular level at the response itself.

Monitoring Technique	Validity Checking	Bias Assessment	Toxicity/Misuse Evaluation	Legality/Privacy Guarantee
Standard ML Metrics	✓	X	X	X
Translation Metrics	✓	X	X	X
Second LLM as a Judge	✓	✓	✓	✓
Implicit Text Comparison	✓	✓	✓	X
Criteria Monitoring	✓	✓	✓	✓
User Interaction Data	✓	✓	✓	X
Drift Detection	✓	✓	X	X
Context Relevance	✓	X	X	X

Figure 1: a table showing the different dimensions across which our proposed monitoring tools can help. We note that the monitoring tools are not sufficient by themselves, as the model and its intended and potential use need to be considered in a wider context than that of monitoring alone.

3) Outcomes and Usage

The methods we propose are all fully automatable; hence, they can be scaled up when deployed without changing the fundamental approach.

LLM responses can be checked against a preselected set of tests before being sent to users. If a test fails (e.g. because the response contains inaccurate information or is not appropriate), the response can be retained for review & training purposes and a more suitable response regenerated to send to the end user. Simultaneously, wider-level metrics such as drift can be tracked over time and, when a threshold is breached, can be automatically flagged.

4) Further Considerations

Obviously, monitoring alone is not sufficient to fully manage the model risk associated with LLMs, or indeed with any model in general; our recommendations, therefore, are not intended as a "silver bullet" to manage all risks associated with LLM use. Rather, our position is that the monitoring guidelines we propose should be taken as one piece of a more comprehensive approach to risk management, which should include other key aspects (e.g. validation) too.

5) Conclusion

Because of the risks associated with LLMs, it is vital that they are monitored wherever they are used; our strategies provide an effective, robust, and explainable way in which to do this.
If you are interested in our white paper or want to discuss this topic more, please get in touch!

Selected References

1) True North Partners, LLM MRM Guidelines: Model Monitoring, *White Paper*
2) Sudjianto et al., Human-Calibrated Automated Testing and Validation of Generative Language Models: An Overview, *preprint, arXiv: 2411.16391*
3) Emeli Dral, How continuous testing keeps your LLM on track, *Talk, Pydata London 2024*

TRUE NORTH PARTNERS
FINANCE | RISK | STRATEGY