truera

QuantUniversity, LLC

Managing the risk of large language models in financial services

NOTES FROM AN INDUSTRY PRACTITIONER ROUNDTABLE
NEW YORK CITY, 3RD MAY 2023

truera

Context

Over the last 2-3 years, Financial Institutions (FIs) have been making changes to their internal policies, standards and processes to manage the risks arising from the adoption of ML models. However, in many FIs, these mechanisms, created for small scale experimentation and adoption, have failed to scale up over time. The challenge has become even more acute as FIs grapple with the implications of recent advances in Large Language Models (LLMs).

This roundtable brought together practitioners from 10-15 leading FIs (including 6 of the top 10 US banks). Representatives from across first line Data and Analytics/ AI teams, Model Risk, Data Management and analytics-intensive functions spent over 2 hours comparing notes on

- The most relevant LLM use cases
- Emerging risk considerations
- Approaches to managing such risks, including but not limited to Model Risk
- Implications for people, skills and ways of working

The event kicked off with short presentations by Agus Sudjianto, EVP and Head of Corporate Model Risk at Wells Fargo, **Sri Krishnamurthy**, founder of **Quant University**, and **Anupam Datta**, co-founder of **TruEra** and a former professor at Carnegie Mellon University. This was followed by an open, Chatham House rules discussion with all round-table participants.
The following pages summarize the key points of discussion during the 2–3-hour discussion. There was broad agreement on some topics, and differences in opinion on other

truera

Key Takeaways

LLM Use Cases

Q&A on internal knowledge corpus is the first “killer app”
Text summarization/ translation is popular too
“Traditional” NLP use cases are on steroids with LLMs
Writing/ reading code may be the Next Big Thing
Appetite for customer-facing use seems limited for now

Risks – New and Old

Accuracy/ hallucination concerns are top-of-mind
Security/ Privacy feature heavily with hosted LLMs
There is uncertainty over IP obligations and rights
Reputation risk is holding back direct customer use
Bias will become important with customer-facing use cases

Managing Risks – Model Risk and Beyond

LLM risk is an end-to-end system risk (not just model risk)
Most risk managers seem to be in exploratory mode
Access to embeddings seems essential for Model Validation
Testing output quality in generative AI is a major challenge

People, Skills, Ways of Working

Everyone agrees that the *nature* of roles will change
There is less alignment on the extent/ pace of job *losses*
“Democratization” of AI will have significant impact
Will Specialists win over Generalists? Jury is still out

truera

1. LLM Use Cases in Finance

The first wave of use cases in FIs are a mix of some in the "too difficult to solve" category, and others in which LLMs represent significant improvements over existing alternatives.

"Killer app" – information retrieval/ contextual search (q&a)

Q&A applications that allow staff to access an internal corpus of relevant knowledge in a simple and reliable way, with a much lower level of curation needed compared to alternatives. Examples: financial advisors querying internal research on investment opportunities; staff querying FI's internal policies to get HR or compliance advice.

FI has control over the quality and provenance of the internal corpus. Testing for quality is easier if the embeddings created by the LLM on the internal corpus are available.

Text summarization (potentially also with translation)

Creating document summaries, typically for review and use by an expert human – e.g., 'reading' multiple documents to create a Know-Your-Customer profile, translating and summarizing a 200-page report in German for a US-based analyst to consume in 5 minutes.

'Traditional' NLP apps - on steroids now!

Traditional NLP use cases – e.g., extracting relevant data from unstructured data sources for operational automation, sentiment analysis on customer complaints or external media reports, analyzing trading room communication for potential misconduct – set for a boom, as LLMs

- Promise greater accuracy on existing data extraction use cases
- Make NLP viable for smaller, "long tail" use cases (previously inaccessible due to need for labelled data and custom model development/ training)
- Add complementary 'last mile' generative AI capabilities on top of data extraction (e.g. text summarization) that increase end-to-end automation levels

Reading/ writing code – next big thing?

Significant medium-term potential in LLM applications that can translate code into non-technical language and vice versa.

Customer-facing applications – low appetite for now!

Potential to automate aspects of customer engagement – e.g., automating service requests – considered but mostly deferred for now. At this stage, FIs are choosing to attack lower hanging fruit (above) with fewer privacy, security, bias and reputation risks.

truera

2. Risks with LLM Use

Accuracy/ hallucination concerns dominate, but privacy/ security/ IP/ reputation risks matter too.

Accuracy/ "hallucinations"

Most LLM applications prone to hallucinations, weak in math and unable to provide traceability of answers, at least for now. Except where applying LLMs on FI's internal knowledge corpuses.

Accuracy-related risk perception changes significantly based on the end user. Responses from an LLM-based Q&A system could be:

- Low risk if being reviewed by an expert user before being acted upon
- Somewhat riskier if the end user is inside the FI but lacks domain expertise
- Very risky if put in front of an end customer.

Accuracy is also heavily impacted by the exact wording of the question. "English language programming" can be too flexible and may need to be translated to resemble more structured 'pseudo-code' to ensure repeatability/ reliability.

Security and privacy

Concerns around security and privacy breaches, due to the need to expose internal FI data to the LLM provider. Some FIs have opted to host LLMs inside their own networks (using open source LLMs or through specific arrangements with the likes of Azure/ Open AI).

IP considerations

Risks of inadvertent IP/ copyright breaches by the FI, in case the LLM was originally trained on external data without appropriate permissions. Uncertainty over IP ownership when content or code is created using LLMs.

Reputation

A particular source of concern in customer-facing use cases (e.g., a chatbot that is 'tricked' into using inappropriate language by an end-user trying to embarrass the FI). Concerns around unjust bias not yet significant in the group, since none of the participants is focusing on customer-facing use cases.

***

Beyond the risks themselves, there was also concern about the potential originators of such risks– e.g., fintechs attempting to compete with regulated entities but without the internal guardrails that the latter have in place; staff inside regulated FIs who find themselves using models for the first time due to the LLM-led "democratization" of AI.

truera

3. Managing LLM Risks

Broad agreement that the approach to risk management should be "end-to-end", rather than solely focused on model risk.

Overall - everyone still in experimental mode

A majority of the participants claimed that ChatGPT was 'banned' from use in their organizations, except in narrow supervised conditions.

Several participants acknowledged that they were still trying to put their arms around the risks of LLM. They were in learning mode, and had implemented ad-hoc controls (e.g., a single channel to approve any LLM use case around the FI) in the interim. One participant felt that it was important to let low risk use cases move forward, and use those experiences to flesh out the control framework for higher risk ones.

Model risk – still working through implications

Many FIs opined that MRM departments were under pressure from all parts of their organizations to review/ approve potential LLM use cases. Some felt unprepared for the sudden spike in interest.

Some participants expressed confidence that when the LLMs are used in the embedding space (e.g., with the info retrieval type use cases), it was possible to apply robust MRM controls. Some pointed out that this was already underway with BERT style models.

However, where the embeddings are not made available, outcome-based testing is the only option right now. This naturally limits the extent to which MRM can approve such models for high stakes use cases, and/or results in imposition of additional human controls to compensate. Outcome-based testing is hard since exhaustive testing is not possible in most LLM use cases. Different prompts and/or end-user queries can produce different outcomes with same underlying model.

The challenges of testing generative ai output quality

For Generative AI, an additional testing challenge is to identify the quality of the output itself. Human feedback, where available, is of course ideal. However, in many situations, that is not a viable or scalable option. Here, considering a combination of automated feedback functions may be helpful (e.g., relevance testing or sentiment analysis of the answers)

The analogy is to data labelling in traditional ML, which was initially all human, but subsequently used automated techniques. Similar programmatic approaches *may* provide a way to test the quality of Generative AI system outputs.

truera

4. The People Dimension

A sense that LLMs will have profound impact on the nature, and possibly the number, of jobs in the industry, but not much clarity or agreement beyond that.

Elimination of roles – unclear at this stage

Opinion seemed divided on if/ how quickly LLMs will result in large scale reduction of jobs. Some participants felt that there may be fewer roles in coding and data science; others foresaw significant reductions in operations. However, few offered a view on the timing or scale of such losses

Consensus that many existing roles will change

There was general agreement among participants that over time, several roles inside FIs will certainly change a lot. One participant highlighted the potential for LLMs to automate the interpretation of code and/or the generation of new code from natural language instructions. At current capability levels, this would still leave a role for coders, but much more in an expert 'checker' role.

Another opined that LLMs are democratizing data science (since "anyone can implement LLM based applications"), and this would have significant implications, positive and negative, for data scientists.

The talent pipeline challenge

One potential challenge is the need for apprenticeship in order to build the necessary domain expertise needed to work with advanced AI tools: if most of the 'grunt work' that was previously done by entry level/ junior staff can be automated, then how would the pipeline for tomorrow's experts be built?

Experts vs. Generalists: no consensus yet

An unresolved question was whether widespread use of LLMs will tilt the balance more towards generalists or specialists.

- Will LLMs empower generalists to overcome the disadvantages arising from their lack of specialist knowledge, or
- Will they further entrench the value of specialists who can complement their deep domain knowledge with the outputs from LLM applications