QUANTIFIND

Practical AI Language Models in Risk and Compliance

A Large Language Model (LLM) Toolbox for Combating Financial Crimes

Dr. Ari Tuchman, CEO, Quantifind

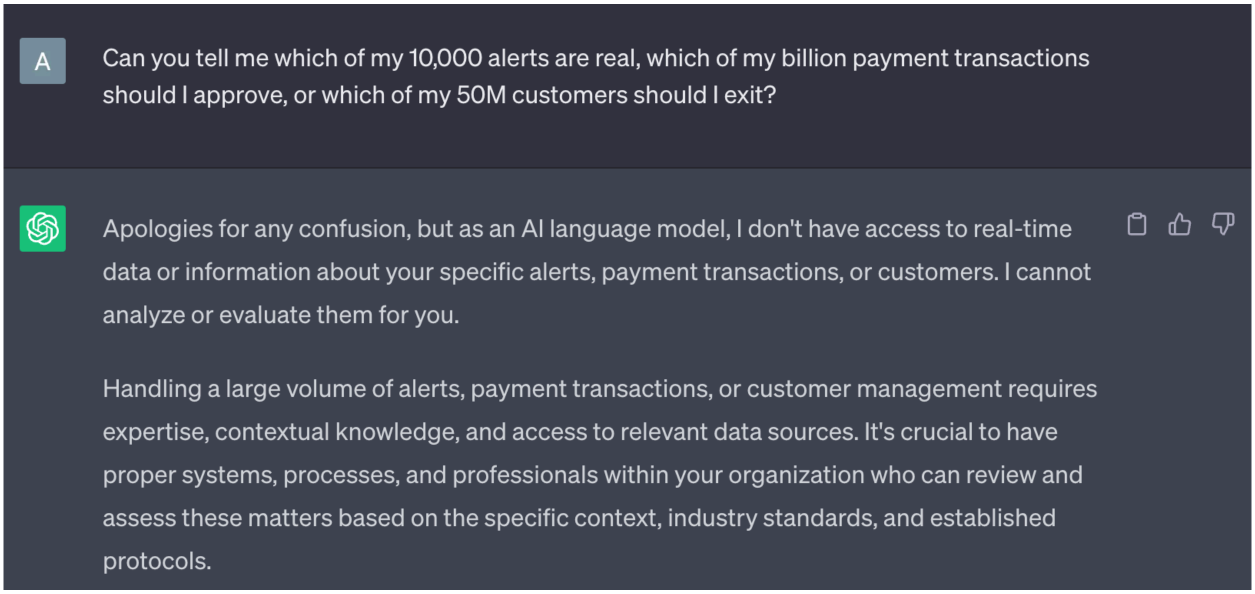
Practical AI Language Models in Risk and Compliance

A Large Language Model (LLM) Toolbox for Combating Financial Crimes

The explosion and subsequent ubiquity of ChatGPT, and the promise of AI Large Language Models (LLM) in general, has rapidly impacted our culture – ushering in both eager anticipation and simultaneous fears of an AI revolution. Start-ups and Fortune 100 companies alike have raced to ride the hype wave (whether for true impact or perhaps for valuations and stock prices) – evangelizing ChatGPT product integrations such as integrated recommendation engines and auto-generated customer email responses. Like with any paradigm shift, many of these integrations will turn out to have little practical application, at best, or be superficial marketing ploys at worst. Others appear poised to provide lasting disruptive innovation but may be years away from making an impact. However, some meaningful applications will be transformative.

One area where we are seeing immediate potential of impactful LLMs is in risk and compliance, specifically in combating financial crimes, discovering money laundering scenarios and fraud events, and even uncovering foreign malign influence in our economic markets. The race to automate much of the Anti-Money Laundering (AML) / Know Your Customer (KYC) process at financial institutions is a struggle to keep up with exploding transaction volumes, nested and opaque networks, and increased government regulation. Thus, LLM techniques offer immediate and welcome automated resources. They are already being driven towards production in Tier 1 banks and government applications – and have begun proving transformational in reducing false positive alerting and enhancing automated decisioning.

However, one cannot simply plug ChatGPT into an anti-money laundering or fraud prevention/detection process and expect to get meaningful results or remain in compliance with FinCEN regulations. A simple prompt to ChatGPT is illustrative:



ChatGPT is certainly correct that “It’s crucial to have proper systems, processes and professionals....” **Any AI solution must be part of a comprehensive system.**

Nevertheless, there are immediate and powerful aspects of LLM stacks that can be leveraged independently into a next generation AI solution for AML compliance and threat analysis. Many of these components have already been integrated into production AML solutions for transaction monitoring, alerts triage, and perpetual customer due diligence.

Before we explore why AI language models can be such powerful tools for financial crimes compliance, we need to delve into two key aspects of LLMs.

Large Language Models: Word Embeddings and Attention

The mathematical intricacies of LLM techniques are beyond the scope of this summary, but many subcomponents of these language models are being introduced into the jargon for compliance leaders. Even if only at a high level to respond to senior executives about innovation or to regulators about how a compliance process is leveraging “cutting edge AI,” understanding key terms can demystify the ChatGPT conversations around risk and compliance.

In the interest of focusing on practical aspects of this AI revolution, it is therefore worth drilling deeper into two of the more common buzz phrases: Word Embeddings and Attention.

Both techniques are relevant for defining and extracting entity risks from large unstructured databases. Therefore, both can be used to drive automation and to increase efficiency in recognizing which alerts and threats must be prioritized in combating money laundering and financial crimes.

Word Embeddings

Word embeddings are an input (the encoder) to Large Language Models and are learned over large bodies of text in the pre-training of the overall model. They represent an abstract mathematical way of expressing the occurrence frequency, context and location of words in a sentence or a document. These mathematical expressions can then be used to build a model predicting which word comes next or which word is missing. They are a powerful component of the ChatGPT solution which is used to generate reliable text ranging from college essays to code snippets.

In GPT-4 these word embeddings were trained over tens of gigabytes of data spanning sources from product manuals to ancient poetry. This broad training set has powered its broad applicability that is sweeping the world. However, every algorithm must truncate its vocabulary coverage at some point, and terms that occur very infrequently may not be well expressed in such word embeddings. This limitation greatly reduces the accuracy of understanding, reproducing, and contextualizing these under expressed terms. Further, words and phrases may take on entirely different

> **Quantifind has produced its own domain-specific word embeddings. These AML/KYC embeddings have been trained over domain-specific compliance and risk-relevant articles.**

meanings in different types of literature. To suit the most common use case, many publicly available language models are trained using the most general literature. Thus, even the presence of a key vocabulary term may not yield an accurate contextualized embedding. Without fine-tuning or training over appropriate data, an LLM may be too generalized.

For AML/KYC risk applications, Quantifind has therefore produced its own domain-specific word embeddings. These AML/KYC embeddings have been trained over domain-specific compliance and risk-relevant articles. While the complete embeddings of GPT4 are not available to the public, similar scale trained examples are available for comparison.

For example, the GloVe (Global Vectors for Word Representation from Stanford University) word embeddings do not include terms such as “cyberespionage” or “cryptoassets.” However, both of these terms are important in an AML/KYC domain-specific risk model. Similarly, in GloVe word embeddings, “AML” is closely associated with the term “COPD,” due to connected abbreviations representing Acute Myeloid Leukemia and Chronic Obstructive Pulmonary Disease. However, in a financial crimes and foreign threat domain-specific context, “AML” is more correctly paired with “anti-money laundering.” An overly broad training set can manifest hallucinations that provide the wrong context for a precision application such as AML/KYC. Similarly, “CBD” gets strongly associated with “Central Business District” when using generically trained word embeddings. However, in Quantifind’s domain specific embeddings, the associations with “Cannabidiol” is strongest, critical for extracting Marijuana related risk.

Quantifind’s embeddings have been shown to have approximately 50% more coverage for relevant terms associated with combating financial crimes risk, as compared to the GloVe word embeddings.

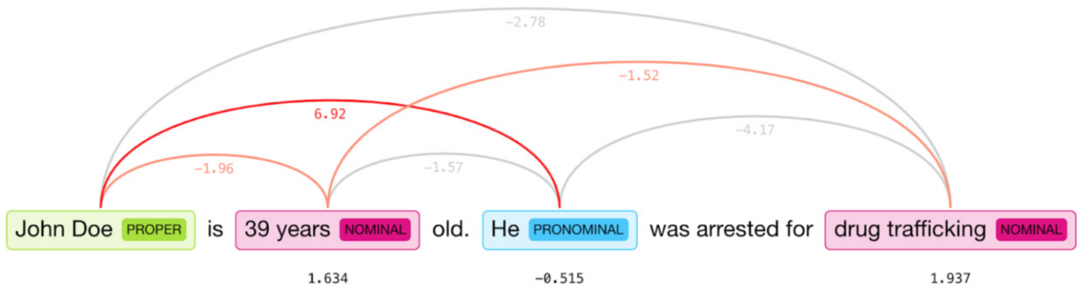
Quantifind’s embeddings have been shown to have approximately 50% more coverage for relevant terms associated with combating financial crimes risk, as compared to the GloVe word embeddings. Quantifind’s word embeddings currently leverage a powerful component of the LLM solutions pervading the market today, but they have been refined to be domain-specific. Therefore, they provide both higher recall and precision. This training improves the accuracy of risk relevance in determining alert triage in transaction monitoring and continuous monitoring, and directly improves the time saved through automation.

Attention

Attention is an innovative mechanism (comprising a neural network model) that can be highly parallelized, and it is the workhorse behind LLMs’ ability to scale. It does not require sequential optimizations to provide feedback one iteration at a time in order to refine output elements (from the decoder). Rather, this model can look at all past calculated elements simultaneously. This parallelization has drastically sped up training times which have powered the ChatGPT phenomenon. Furthermore, this Attention mechanism leverages the hyper-local environment of words to understand how, for example, a word in the beginning of a sentence connects to words in the end of sentence, or even to words in adjacent sentences. This model can leverage the word embeddings described above, with positional, grammatical and document level information encoded.

Attention mechanisms allow a deep learning model to implicitly learn parts of speech and grammatical dependencies between each word in a sentence, and determine their mathematical, weighted connectivity to other terms. To illustrate this explicitly using a dependency parser model, two sentences with their parts of speech detected and quantitative relationships assigned between words are shown below.

Dependency Parser Model



Token-to-token (tokens shown in colored blocks) dependency weights can be leveraged together with word embeddings to determine the relevance of risk typologies efficiently and accurately to a specific person or organization. Attention mechanisms can similarly learn which word structures are the most important in accurately assigning risk. In the examples above, “drug trafficking” is connected to the proper noun, John Doe, even though the antecedent is in a different sentence.

Attention mechanisms are highly powerful in assigning risk relevance by understanding the context of the sentences and can be further optimized by training over domain specific sources, as with the embeddings above.

While super-fast and scalable compared to earlier NLP (Natural Language Processing) models, for some financial transaction screening applications, neural-net Attention mechanisms are still not yet fast enough for the scale of transactions. Therefore, depending on the demands of scalable, real-time requirements, an Attention mechanism for contextual learnings could sometimes be replaced by a simpler dependency parser for capturing the most important contextual signals to be fed through a gradient boost model.

Balancing the precision degradation in predicting AML risk within a highly specified domain with the needs for high performance throughput is a critical component of launching an innovative, but pragmatic AI process as part of a larger compliance workflow.

The word embeddings described in the prior section represent how LLM techniques can be used to define accurate risk typology definitions. Attention techniques can be used to correctly assign those risks to entities being screened. A risk-based, practical compliance process can benefit immediately by both innovations, with care taken to implement the components that ultimately drive efficiency to the workflow – catching more bad actors with reduced resources.

Large Language Models for Risk and Compliance

Underlying the applicability of LLM solutions to augment risk and compliance processes is their ability to extract context from messy, unstructured data. In financial crimes applications, this most saliently refers to uncovering hidden risks in data sources that have traditionally been too noisy to efficiently integrate into a risk-based decisioning process. While extracting risks from text, such as news feeds, is currently leveraged across financial services compliance applications, this process often suffers from painfully inaccurate associations. These mistakes have material financial implications in terms of demanding additional manual resources to review, increasing delays in approving critical transactions due to false positives, and often mistakenly turning away desirable customers. Therefore, integration of disparate data sources such as adverse media has often been reserved for infrequent screens or extremely high stakes applications. However, with proper AI solutions including LLMs that focus on relevant risk topologies and which accurately associate relevant risks with individuals and organizations, these messy data sources can be unlocked.

Risk factors in documents such as news articles or regulatory press releases must be correctly extracted and assigned. However, the underlying documents often mention many entities which confounds a correct association and therefore poses a significant challenge in assigning risk factors to a linked entity (person or organization) mentioned in the document. **For example:**

- **Victims named in an article about violent crime should not be associated with the Violent Crime risk factor**
- **A judge sentencing drug traffickers should not be assigned the Drug Trafficking risk factor**

These entities are “innocent bystanders” that ineffective models will often label inappropriately.

An AI language model can power an effective entity-aware risk relevance model which identifies which entities in the document should or should not be associated with the risk factors that are present. AI provides the tools to attack this problem from a contextual basis, **as seen in the following examples:**

1. **...Jane Doe is an associate at Morgan, Lewis & Bockius LLP, where she focuses primarily on health care fraud and abuse matters...**
2. **...prosecutors will have to prove the Theranos founder acted with intent to defraud customers and investors....**
3. **... Saved In America, a nonprofit organization that says it has rescued more than 230 girls from sex traffickers, was accused of fraud ...**

In the first example, at the document level, there is a keyword match for the Fraud risk factor based, in part, on the phrase in italics “health care fraud.” However, these risk terms are clearly not relevant for the linked entity in bold, Morgan, Lewis & Bockius LLP. The document excerpt in the second example matches the Fraud risk factor again, due, in part, to the italicized text “defraud.” In this example, the risk factor is relevant for the linked entity, Theranos. The third example matches both the Human Trafficking (“sex traffickers”) and the Fraud risk factors, but only Fraud is relevant to the non-profit, Saved In America. These nuanced differences illustrate a key challenge in determining whether a risk factor assigned at a document or paragraph level is relevant for a particular entity mentioned. Determining this context to resolve risk relevance presents a perfect challenge for an AI language model and the techniques introduced above.

Conclusion

This brief summary of two LLM techniques in the world of risk and compliance is meant to highlight both the immediate gains as well as practical considerations in building a next generation AML/KYC process. There are many efficiency gains from AI language models to be leveraged, for example, by improving model development time with synthetic training data. However, one additional application that is often requested by practitioners is to streamline the generation of Suspicious Activity Report (SAR) narratives.

If the risk factor extraction is performed correctly using, in part, the techniques described above, then a profile for each entity, their transactions, as well as the risk associated with the entity across public domain data sources can be generated. This profile can be summarized automatically using the LLM infrastructure. The LLM model can be further trained on the domain of actual SAR filings, to mimic the SAR narratives more accurately. For some simple SARs, such as basic structuring (or “smurfing”) of transactions to avoid regulatory thresholds, these could be generated in a straightforward fashion. For more complex transaction patterns and public domain behavioral risks, further domain-specific training will likely be required, but could be rapidly implemented.

Additional applications of AI language models and LLM techniques to risk and compliance are inevitable and exciting. However, the requirements for domain-specific accuracy, domain-knowledge, and speed at scale will continue to require domain-specific AI products. These innovations will be part of a comprehensive solution containing exactly those “proper systems, processes and professionals” that ChatGPT so wisely recommends.

About Quantifind:

Quantifind’s AI-powered risk intelligence automation empowers risk analysts to investigate entities and uncover relevant risk signals with breadth, depth, accuracy, and speed. Quantifind‘s AI platform streamlines risk management workflows by delivering superior entity resolution, dynamic risk typologies, advanced knowledge graph technology, and by responsibly leveraging large language models. Quantifind is a top provider to tier 1 financial institutions, Higher Ed, Fintech, and government organizations for the following use cases: Know Your Customer (KYC), Enhanced Due Diligence (EDD), Alerts Triage, Continuous Customer Monitoring, and Supply Chain Risk Screening. These organizations benefit from the following proven outcomes: 95% reduction in false positives, 50% improvement in investigative efficiency, and daily screening of the entire stakeholder base with over 92% name resolution accuracy.