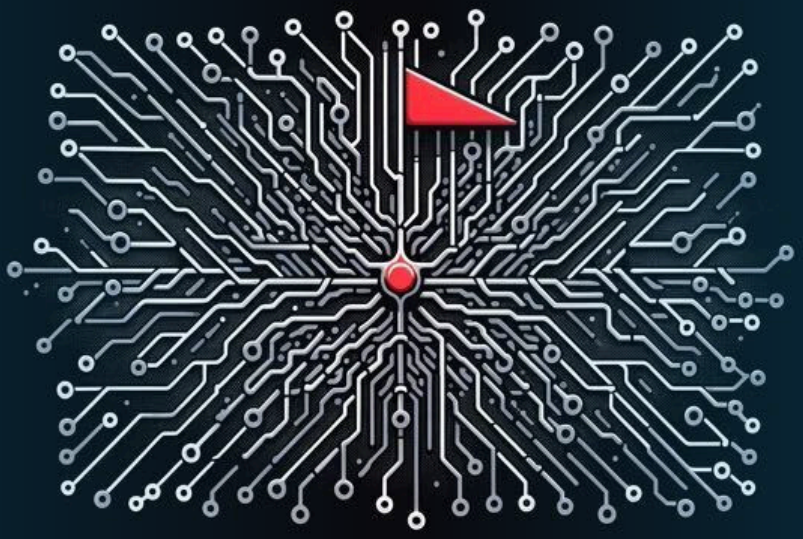
QUANTIFIND

Choosing Your AI Battles: Precision LLMs in Risk Assessment, from Anti-Money Laundering to Vendor Vetting

John Stockton, Co-founder, Quantifind November 2023



Both the financial and government sectors must mitigate the risks of working with vast, globalized sets of people and organizations. Whether to comply with AML/KYC (Anti-Money Laundering/Know Your Customer) regulations or to de-risk vendors within a supply chain, any leader who manages the risk exposure of their institution needs to balance growth opportunities with risk management innovation. While Large Language Models (LLMs) are new and come with risks of their own, there is an increasingly undeniable case that they will be a valuable contributor to risk assessment products of the future.

To help risk management practitioners make informed decisions and cut through the hype around AI solutions, this paper presents a framework for assessing where LLMs could be responsibly used in a risk assessment technology stack. To underscore specific challenges, the overview is followed by the discussion of a single use case within Quantifind's Graphyte platform, where LLMs are used to "structure unstructured data" for risk-labeling entities.

The Promise and Peril of LLMs for Risk Assessment

Large Language Models such as GPT4 have recently transformed the landscape of AI, through products like ChatGPT and its competitors. These LLMs are trained over vast amounts of data, in a process that encodes the contents and patterns of a corpus into the weights of an extensive multi-layer neural network. Aided by recent architectural innovations (transformers), high-power hardware (GPUs), and massive amounts of training data, LLMs have enabled many new AI applications, particularly those involving natural language processing (NLP) tasks.

While the excitement is palpable, it is essential to approach the hype with a critical perspective. Responsible AI challenges, including biases, hallucinations, lack of explanations, and information leakage, must be carefully addressed. The hype often generates expectations that may not align with the current capabilities of AI, emphasizing the importance of informed and realistic assessments when integrating these technologies into various domains.

Even with optimally trained models, how the model is designed into a product can make or break the implementation. The details of where an LLM is used and how it is deployed can be the difference between creating a truly differentiated product versus a distracting "me too" AI feature that introduces more problems than it solves.

The strategic adoption of a suitable AI language model, when executed with careful consideration, can revolutionize risk management strategies and augment decision-making workflows. The overall impact in enhancing a business operation hinges on a nuanced evaluation of the model's accuracy, speed, cost, scalability, and compliance adherence. While some implementations may prioritize accuracy, potentially leading to compromises in cost-efficiency and transparency, judicious design and targeted application of the language model can mitigate these issues.

A surgical choice of how an LLM is applied can simultaneously reduce false positives, monitor emerging threats, and adhere to regulatory standards. A precise LLM can enable organizations to streamline routine tasks and make more informed decisions swiftly. Such an implementation can significantly reduce risk exposure and boost productivity, for investigators and developers.

Quantifind and Precision LLMs

Quantifind is at the forefront of developing innovative software using precise language models (PLLM) to evaluate the risk profiles of entities (people and organizations) with the highest degree of speed and accuracy. Quantifind empowers its users to make data-driven decisions with a precise risk intelligence platform, Graphyte, that identifies entity associations with fraud, financial crimes, foreign investments, and many other risk factors. Because its models and risk-based use cases depend heavily on understanding the context within unstructured textual data, there are countless opportunities to leverage the intelligence gains made possible by recent developments in Large Language Models.

There are abundant opportunities and potential integration points for large language models, but therein lies the challenge. There is a "tyranny of choice" that any product developer faces when told by upper management to sprinkle "AI pixie dust" on their products or processes.

The pressure to keep up with the marketing hype may lead to the premature and misplaced application of LLMs. This is especially problematic in the regulated risk assessment industry, where poor decisions are high-impact and high-cost.

The impact of any potential implementation should be weighed against two considerations: economic cost and model performance. **Despite their promise, many LLM implementations are still too inaccurate, slow, and expensive for scaled-out deployments.** On the performance side, even when accuracy numbers are promising, ensuring rigorous model control from the outset of any integration is critical to mitigating the risks of hallucinations and unexplainable results. This process of "baking-in" trust was discussed in our previous work on responsible AI, and Quantifind is actively extending it to the domain of LLMs in risk management. These concerns can be effectively mitigated by making a wise choice about the class of LLM that is used and the way it is integrated.

Categorizing LLM Integration Opportunities

With so many possibilities for LLM integration, this section lays out a framework for where and how LLMs can be used in a risk assessment stack such as the one that powers Graphyte. The description of these tasks below helps explain how LLM-enabled automation can improve efficiency and effectiveness. As shown in Figure 1, these tasks are categorized by *who* they are directly helping, *when* they are used, and the ultimate impact of potential integrations. In each example, a human is in the loop, either as the one being helped, or the one doing the helping.

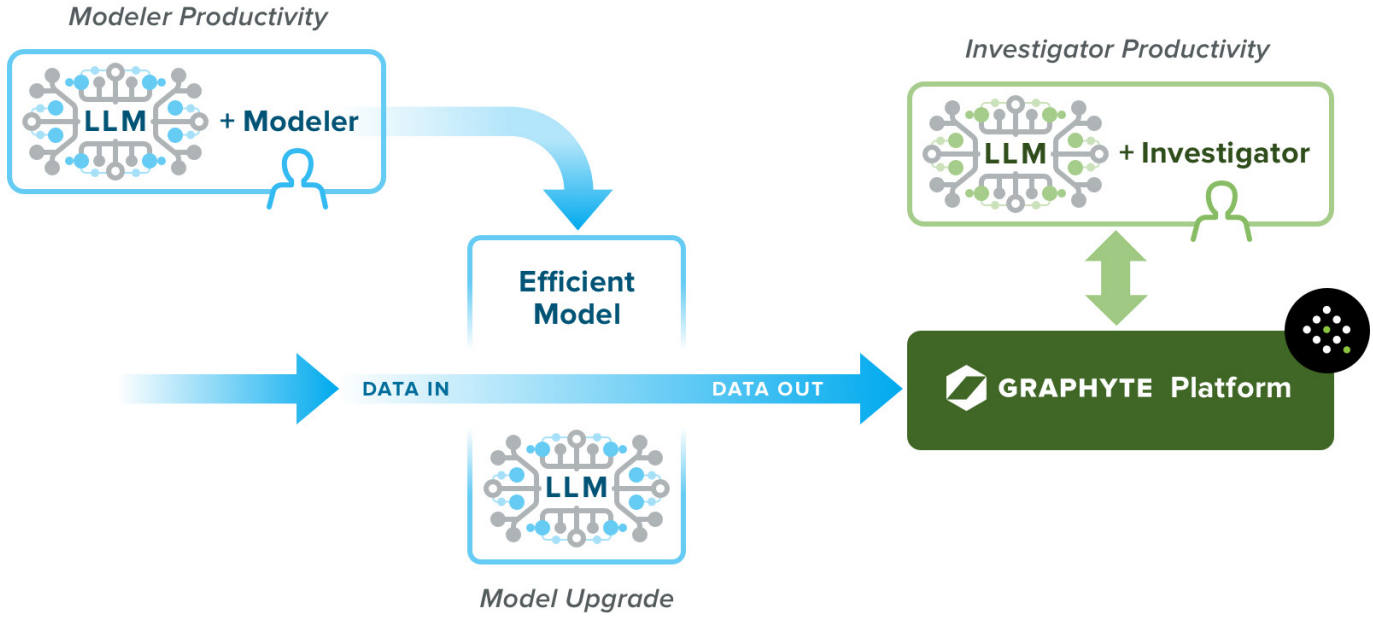


Figure 1: Integration Opportunities for LLMs

Modeler Productivity

When: **Model Time**

Impact: **Higher performance models, without LLM-specific risks**

Quantifind's data science team harnesses the power of LLMs to enhance traditional model-building processes. As experienced pragmatists, Quantifind uses the component technologies inherent to language models to achieve superior performance in many critical tasks, including domain-specific embeddings to assist in risk labeling. Instead of completely entrusting a critical task to an LLM, certain important sub-tasks like featurization (e.g., embeddings for topic models) and training data generation (whether to create synthetic examples or replace human-generated labels over real data) can be assigned. This approach allows the training of non-LLM models that are more efficient and controllable in production. It allows a product to secure the upside of LLMs while avoiding the risk of implementing LLMs in production, where hallucinations and lack of provenance may be showstoppers. LLMs also play a role in baselining, providing insights into potential model performance limits, but properly weighed against the speed, control, and efficiency of existing models. Think of this like the veteran presence on a team: even if they sit out of the game because they are not fast enough, they can play a pivotal role in training younger players.

> **Some LLMs are like this veteran player, capable of training other more stream-lined models with higher pipeline (game-time) performance potential.**

This form of semi-manual or automated model compression is an active, and promising, research area. Most models do not have to be AGI (artificial general intelligence). The key learnings from an LLM can be distilled into something that works with much higher speed performance and lower cost because of a specialization trade-off. All of these approaches can be thought of as productivity enhancements for model creators. Just as programmers use LLMs such as Microsoft's Copilot to enhance "Developer Productivity," LLMs can also be deployed alongside data scientists in these ways to amplify "Modeler Productivity."

Model Upgrades

When: **Pipeline Time**

Impact: **Potentially optimal model performance for isolated tasks**

Certain scenarios warrant "handing the keys" to an LLM for part of the production model stack. For example, an LLM is integrated into part of Quantifind's ETL (extract, transform, load) pipeline to create artifacts and indexes that optimize the efficiency of its real-time platform, Graphyte. This can consolidate many existing models and streamline development significantly. However, model compression is still in its infancy, and the cost of using an existing commercial LLM (whether self-hosted or via API) to fully process, contextualize, and understand the scale of news articles that power Graphyte would be too expensive. Also, downside risks, including hallucinations and leakage, should be mitigated through transparency and validation testing.

> **Specifically focused domains with manageable data volumes present viable opportunities for LLM integration.**

The concluding section of this paper presents a use case illustrating the application of LLMs to convert unstructured articles into structured data tables.

Investigator Productivity

When: **Query Time**

Impact: **Increased user efficiency through the transparent automation of "last mile" manual tasks**

Both of the above LLM use cases are less visible to users, who may not know when they benefit from LLMs. However, multiple uses of LLMs put them directly on the "service staff," enabling users to access them directly and in real-time.

> **As data scientists can use LLMs to enhance "Modeler Productivity," risk investigators can use LLMs as an assistant to enhance "Investigator Productivity."**

One such direct use case is profile summarization. In some risk investigations, a user may be confronted with the unwieldy manual task of turning hundreds of relevant articles into a short, actionable summary. Instead, LLMs can distill the essential points of those articles or even prepare a draft summary that expedites the investigator's official reporting process. Other assistive features could include article clustering, single article summarization, and text annotators that save the human user the hassle of fully processing superfluous information. This digital assistant accelerates manual tasks, further bolstering the inherent efficiencies of the base platform but without taking the investigator out of the driver's seat. Proper product design should render it obvious to the user exactly when AI is used in this manner because some applications may not allow AI-generated content.

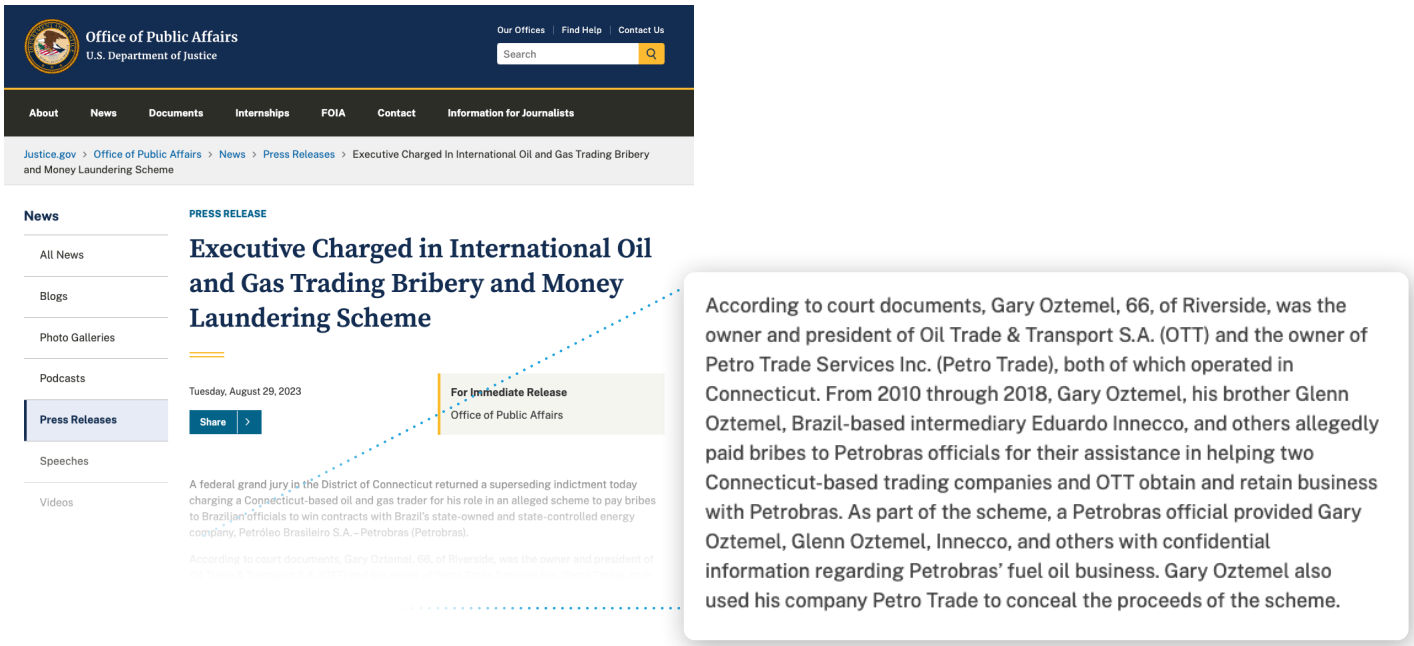
These are only a few examples, and many other applications exist. One could, for example, use an LLM with human feedback to curate data into increasingly "exquisite" intelligence within a knowledge graph in the data preparation phase itself, iteratively removing all entity resolution and risk relevance mistakes. Still, the framework outlined above provides a basic roadmap for integrating LLMs into existing stacks without necessitating a complete architectural overhaul. This is a useful perspective for product developers and senior risk management teams who help inform product development.

Focused Example: Structuring the Unstructured

A focused "Pipeline Time" application of LLMs within the Quantifind pipeline, is illustrated in *Figure 2.* Quantifind processes a massive corpus of unstructured documents (including news articles, court documents, and other sources of natural language text) in multiple languages, and the output of this pipeline enables the real-time Graphyte Platform. Within the pipeline, a suite of models for entity recognition, relationship extraction, and metadata extraction are applied. These models effectively turn the unstructured data (text) into structured data (tables): a "Text-to-Table" workflow. These models could be replaced with a catch-all, fine-tuned LLM approach; however, they would be prohibitively expensive to run, given the cost and speed associated with typical LLM performance.

A more pragmatic strategy involves concentrating our efforts on a smaller yet significant domain. For example, Quantifind leverages an LLM approach on Department of Justice (DOJ) press releases, which is a good target database due to the consistent structure of each document. Each DOJ press release always has the same pattern (bad guys listed in a certain way, good guys in a different way, etc). As with other government text documents (SEC, patents, etc.), these articles are semi-structured, and not as variable as the unstructured content found in less formal data sets. This structure can be leveraged by both traditional algorithms and LLMs to outperform more generalized approaches that fail to explicitly recognize and leverage this structure.

As seen in *Figure 2*, the objective of the "Text-to-Table" workflow is to take every article and convert it into a table with designated columns, including Entity Name, Alias, Type, Location, Age, Role, Crimes Committed, Relationships, Job Title, etc. To achieve this overall goal, one could choose to fine-tune an LLM, explicitly training it over these articles. Or one could use prompt engineering with a pre-trained LLM.



Name	Type	Age	Location	Alias	Crime	Relationships
Gary Oztemel	PER	66	Riverside, CA	…	Money Laundering, Bribery, FCPA	…
Oil Trade and Transport S.A.	ORG	NA	…	OTT	…	…
Eduardo Innecco	PER	…	Brazil	…	…	…
…	…	…	…	…	…	…

Figure 2: Text-to-Table Workflow

Several challenges arise in the latter approach. Putting aside plumbing and efficiency issues, the primary challenges and items to consider while designing prompts include:

- **Precision:** Is the information in each column accurate?
- **Recall:** Does the output table include an entry for every entity mentioned? How many columns are "filled in"?
- **Information Leakage:** How do we ensure that the LLM is not bringing in information from outside the article? For example, suppose an article is about a prominent person. How do we prevent the model from pulling in the information not contained in the article but within the LLMs own training data?
- **Hallucinations:** How do we ensure that the model does not make up results not contained in the article?
- **Output Consistency:** Does the LLM always produce the same output type? (e.g., properly formatted CSV or JSON).
- **Prompt Strategies:** While prompt strategies will likely become more deterministic over time, there is an art to them for now, and multiple approaches should be attempted. For example, chain-of-thought and few-shot prompting are strategies that can be considered to specify points of emphasis, such as how much attention to apply to certain entities or themes versus others.

One backstop for leakage, hallucinations, or any other imprecision is provided by provenance: in Graphyte users are always given access to the raw data in the user interface. In this way, Graphyte's transparent UI design relieves pressure on the model. While effective prompting is critical, perfect prompting is not achievable; embedded provenance features in the interface allow users to conveniently validate the most critical findings. These design patterns of "augmented intelligence" allow the solution to be distributed across development teams beyond the model developers alone.

After using systematic prompting approaches, Quantifind achieved high-performance metrics in the Text-to-Table task. Performance on people entities surpasses 90% accuracy, while also exhibiting enhanced recall compared to traditional models. Organizations were more difficult and less susceptible to certain heuristics because they are often indirectly involved in DOJ articles and less frequently the direct target of prosecution. Therefore, organization entities require direct fine-tuning of an LLM in order to achieve comparable accuracy.

Learning to Learn with LLMs

The choice of when and where to use LLMs is dependent on where they can provide the most benefit while minimizing cost and aberrant behavior. However, it is also important not to overanalyze the problem. The field will not cease to keep moving, and the only real way to begin a culture of learning and adapting is to recognize low-hanging opportunities to deliver high-impact wins. Even if the initial application is small, the establishment of any model in the new paradigm of LLMs will help an organization better know how and when to scale them to more complex and mission-critical applications. Starting the journey is more important than getting it completely right, especially if risks can be mitigated through careful consideration of model transparency, validation, and the pragmatic choice of a starting point.

Contact Us

Email contact@quantifind.com to discuss the pragmatic application of AI to address your specific risk assessment challenges.