Allianz Global Investors

AI DECODED – INVESTMENT INNOVATION WITH BEST STYLES | OCTOBER 2025

From Word Counts to Transformers

Many people have been fascinated by the idea of using artificial intelligence (AI) for communication between men and machines. While HAL in Stanley Kubrick's *2001: A Space Odyssey* turned out to be less helpful to its human crew, its modern incarnation has been fascinating millions. With the success of ChatGPT, natural language processing (NLP) has taken center stage in the public eye as it has become synonymous with (generative) AI. More recently, models like DeepSeek have captured headlines – and even moved markets – highlighting the transformative power of large language models (LLMs) beyond chats. But this breakthrough didn't happen overnight. Generative AI and modern NLP are built on decades of progress, beginning with simple techniques like word counts and dictionaries. Let us take you on a journey from humble beginnings to the present and into the future.

What is a transformer model? Illustrating the technology with practical examples

The early approaches in the field of NLP captured broad structure but struggled to understand nuances. To move beyond simple word matching researchers had to win the fight against complexity (the "curse of dimensionality"[1]), by representing words as points in a (vector) space of reasonably low dimensions – so called word embeddings. After Bengio et al. (2003)[2] introduced neural networks to compute work embeddings models like Word2Vec significantly popularized and operationalized these embeddings. The striking observation was that this generation of models captures both semantic and syntactic relationships, making it one of the foundational techniques in modern NLP.

In our previous episode, we explored word embeddings like GloVe[3] (which at the time was the biggest competitor of Word2Vec) and how the Systematic Equity team applies it to turn earnings call transcripts and 10-K filings into investment signals aimed at enhancing risk-adjusted returns.

Dr. Jörg Hofmann
Director of Quantitative Research Systematic Equity

Tianxiao Fan
Product Specialist Systematic Equity

GloVe and other word embedding methods are powerful tools – they capture relationships between words and can often comprehend context surprisingly well, especially when applied to large, domain-specific datasets.

However, word embeddings still map each word on a single, fixed representation, regardless of the specific sentence or context in which it appears. After several failed attempts to overcome this limitation the 2017 paper "Attention is All You Need"[4] eventually introduced transformer models, which dynamically interpret words based on their surrounding context. For example, in traditional word embedding models, the word "bank" would have the same meaning whether it appears in "*river bank*" or "*bank account*". In contrast, transformer models can distinguish between these meanings by considering the entire sentence, understanding whether "bank" refers to the side of a river or a financial institution. Unlike static word embeddings, transformers read like humans – adaptively interpreting language based on context. Though more complex and resource-intensive, they offer a deeper and more flexible understanding of language.

At the heart of this innovation is a mechanism called self-attention, which lets the model weigh the importance of each word in a sentence, or even a document, relative to every other word. This means the model doesn't just look at words one-by-one but considers how all the words in an article relate to each other to grasp the full meaning.

Imagine you're describing a holiday: "I took my dog to the beach, and she loved playing with the waves." A transformer model understands that "she" refers to "my dog" and that "playing with the waves" is something enjoyable happening on the beach. It pays attention to these connections across the sentences, rather than interpreting each word separately. This ability to connect words and ideas across large chunks of text or even full documents helps transformers understand language much more like a human would, making them powerful tools for tasks like translation, summarization, and sentiment analysis. Transformer technology is the basic building block of state-of-the-art LLMs, where the term "large" refers to the billions or even trillions of parameters they contain.

LLMs in asset management

In the finance industry interpreting complex and nuanced text is critical. Earnings call transcripts, regulatory filings, and analyst reports, contain rich but unstructured information that traditional NLP-methods struggle to analyze effectively. Hence, the power of transformers caught the interest of investment professionals, since transformer models enable investment professionals to extract deeper insights from texts, improving decision-making and risk assessment.

There are several commercial and open-sourced LLMs available in the market, each designed to address a variety of needs – including those specific to finance. Among those, two of the most popular models are based on GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) architectures. Within the GPT model family, BloombergGPT[5] is tailored to financial applications. Similarly, within the BERT model family, FinBERT[6] has emerged as a version of BERT that has been specifically trained for understanding and analyzing financial texts.

BloombergGPT is a large language model developed by Bloomberg, designed specifically for the financial domain. Their model contains more than 50 billion parameters and was trained on a dataset of over 700 billion tokens (the basic unit of text). This includes around 360 billion tokens from their financial data sources (including news articles, research reports, and market data), along with 345 billion tokens from general-purpose datasets.

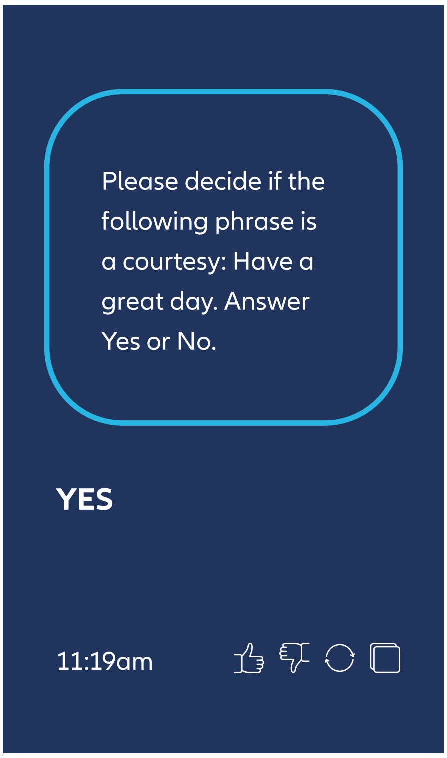
On the other hand, FinBERT is a compact, open-source transformer model based on the BERT architecture. As discussed in the paper by Huang, Allen H., Hui Wang, and Yi Yang in 2022, FinBERT was trained on a standard financial text corpus consisting of 2.5 billion tokens from corporate reports (10-K and 10-Q filings), 1.3 billion tokens from earnings call transcripts, and 1.1 billion tokens from analyst reports.

For sentiment analysis, FinBERT's final classification layer was further refined using 10,000 manually annotated sentences from analyst reports, labeled as positive, negative, or neutral. This targeted training enables FinBERT to detect nuanced sentiments, such as cautious optimism or concern, that often appears in earnings calls and regulatory filings.

Because FinBERT is built on the BERT architecture, it inherits the ability to understand context and meaning at a deep level. At the same time, it is smaller and more efficient than many other modern LLMs, making it practical for deployment in real-world financial systems. Additionally, its open-source nature allows for customization and integration into proprietary workflows.

Challenges and risks of solely relying on third party LLMs

However, FinBERT also has some known limitations. For examples, it sometimes misclassifies courteous phrases like "Have a great day" or "Enjoy your meal" as positive sentiment, while texts like "no problem" as negative sentiment, even though these are neutral in financial contexts. Such errors can be mitigated through fine-tuning or by using more advanced prompting techniques with newer LLMs like GPT-4, which can better interpret subtle nuances[7].

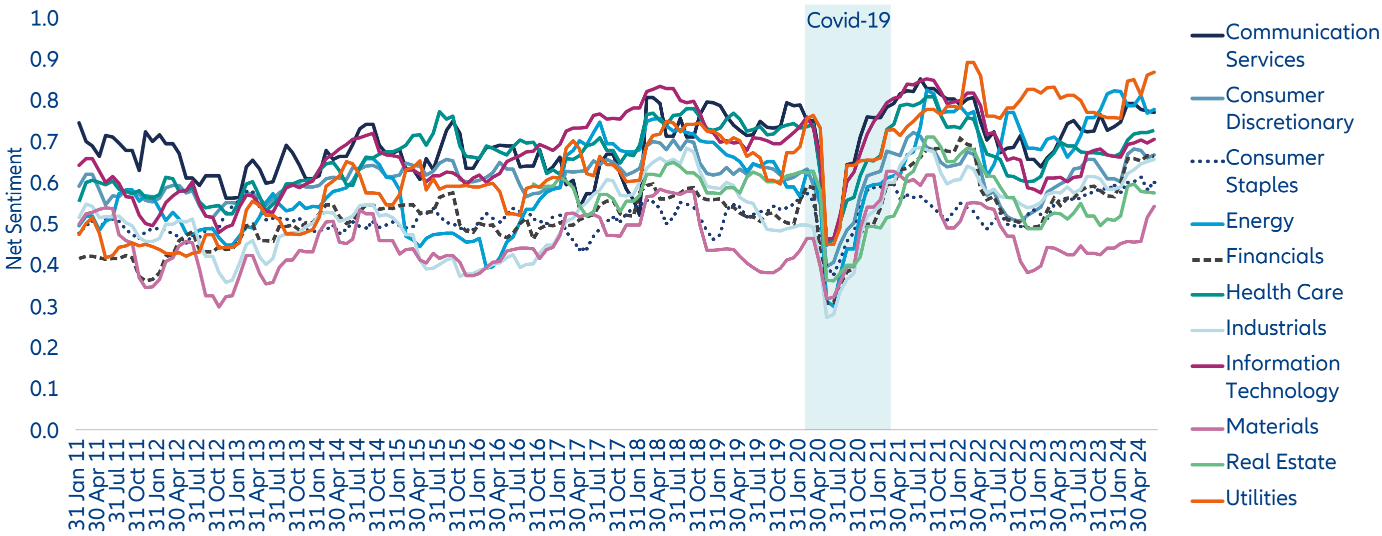


Source: Mistral AI, June 2025

While the previous examples were based on short sentences, we now turn to a more scalable approach to sentiment analysis suitable for longer financial documents, namely net sentiments, that works for documents of any length. To compute net sentiment, a document is first divided into chunks. Sentiment is then evaluated for each chunk individually. The results are aggregated from a chunk level to the full document level by comparing the number of positively scored chunks to negatively scored ones, yielding an overall net sentiment score[8].

Based on this net sentiment score, we had another observation which may challenge FinBERT's ability to accurately detect sentiment at the level of individual companies. While the model could identify negative sentiment during the COVID-19 pandemic (as depicted in the grey area in the chart below), it is important to note that some sectors consistently exhibit more positive sentiment than others. For example, communication services (yellow line in the chart below) and information technology (orange line) always exhibit more positive sentiment than Materials (purple line). This sector bias may require additional adjustments in the later stage of the investment process, particularly for investors who prefer to avoid sector-specific exposures because they do not believe taking such risks is compensated over the long term.

Exhibit 1: Average net sentiment of GICS sectors



Source: Allianz Global Investors, Systematic Equity team. Data as of 30/06/2024. Back-test. Please refer to the disclosures related to "Back-testings and hypothetical or simulated performance data " on the last page of this document. The hypothetical performance and simulations shown are for illustrative purposes only and do not represent actual performance; they do not predict future returns. Please see important information regarding back-testings and hypothetical or simulated performance data at the end of this document. For illustrative purposes only and does not represent actual performance of any client account. The information should not be relied upon as an indicator of future results

Relying solely on off-the-shelf models comes with additional limitations, which can be grouped into two main issues:

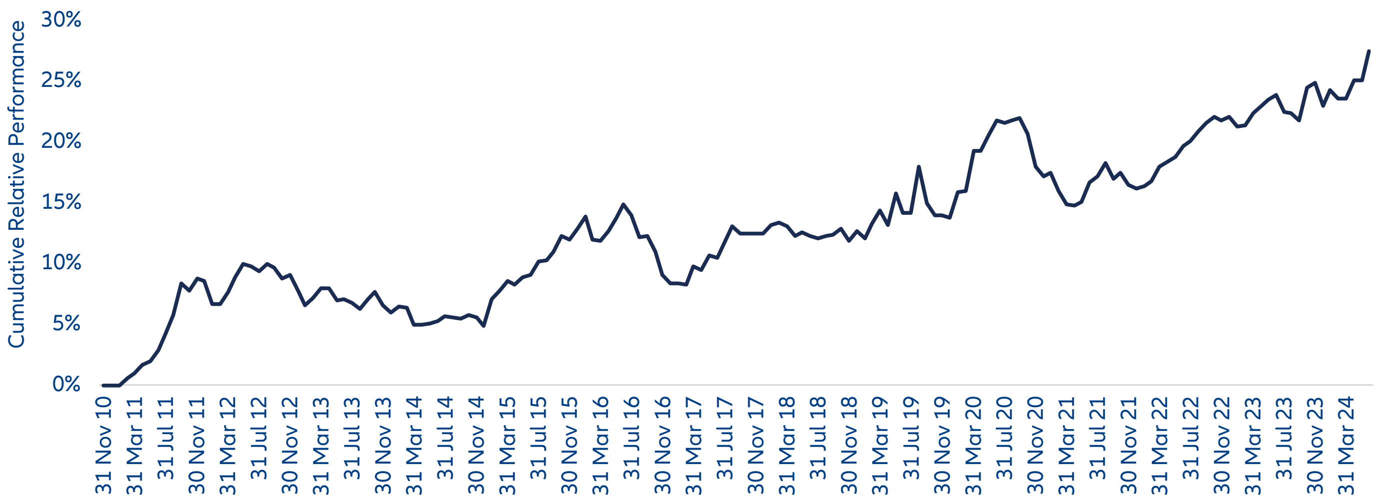
- **Issue 1:** In the early days, most NLP research was conducted within academic institutions, but over time, large corporations have taken the lead. This shift raises concerns around compliance, data ownership, data privacy, and copyright – critical considerations in the highly regulated financial industry.
- **Issue 2:** Additionally, these models continue to grow in size and complexity. However, as highlighted in the Chinchilla paper[9], simply increasing model size without a proportional increase in high-quality training data leads to diminishing returns. For our proprietary financial applications, there might not be enough domain-specific data to fully leverage these massive models efficiently.

Consequently, off-the-shelf models may not be the best choice to meet our unique requirements. To maintain control over data governance and optimize model performance in the financial context, we must conduct our own research and develop customized transformer models tailored to our strategy.

To address this, we have customized the FinBERT model specifically for our use cases, ensuring it was better aligned with the unique requirements of our investment process. We then conducted a comprehensive simulation using sentiment signals generated by our tailored FinBERT model. The result of our backtest was encouraging: the simulated strategy achieved an information ratio of approximately 0.6 (before cost) against the S&P 500 index (as shown below). This demonstrates that machine-evaluated sentiment, when properly calibrated and integrated into a disciplined investment framework, can contribute meaningfully to portfolio performance.

While these results are promising, we expect more to come. Thus far, this article has focused solely on sentiment. However, the potential of LLMs lies well beyond sentiment analysis. These models are capable of more sophisticated tasks, such as understanding casual relationships, or even reasoning. As LLMs continue to evolve, we see exciting opportunities to harness their capabilities for deeper insights and more sophisticated analysis.

Exhibit 2: Cumulative relative performance of our tailored BERT-based signal



Source: Allianz Global Investors, Systematic Equity team. Data as of 30/06/2024. Back-test. Please refer to the disclosures related to "Back-testings and hypothetical or simulated performance data " on the last page of this document. The hypothetical performance and simulations shown are for illustrative purposes only and do not represent actual performance; they do not predict future returns. Please see important information regarding back-testings and hypothetical or simulated performance data at the end of this document. For illustrative purposes only and does not represent actual performance of any client account. The information should not be relied upon as an indicator of future results

Can LLMs replace analysts? – Probably not!

Based on the promising results from our sentiment-based strategy, we went a step further and asked ourselves: Could this technology eventually replace a financial analyst?

A core function of fundamental analysis is building valuation models using data from balance sheets and cash flow statements. Traditionally, this requires skilled analysts to interpret and contextualize company fundamentals. To test whether LLMs could assist in this domain, we leveraged OpenAI's GPT-4 model via the OpenAI API, processing reports of thousands of companies in parallel.

Our experiments showed that LLMs can indeed extract key figures, compute ratios, and summarize broad trends with reliability. This aligns with research like the Chicago Booth paper[10], which demonstrated that LLMs can ingest cash flow statements and balance sheets to evaluate future earnings improvement.

However, recent studies underscore the limitations of current reasoning models. Shojaee et al. (2025)[11] systematically examined the abilities of large reasoning models and found they face a collapse in accuracy and consistency beyond certain levels of complexity. While LLMs excel at broad pattern recognition and can generate convincing reasoning and analytic chains for simple tasks, their computational reasoning deteriorates for nuanced and high-complexity problems, such as those frequently encountered in comprehensive equity analysis. The study further shows these models often fail to employ explicit algorithms and may produce inconsistent explanations when faced with complex scenarios.

This suggests that, although LLMs can reliably extract figures, identify trends, compute ratios, and provide analytical starting points, they are not yet able to fully replicate the depth, logic, and rigor of a skilled human analyst, especially when the financial situation demands nuanced judgement or the construction of bespoke investment rationales. In practice, LLMs should be viewed as a supportive tool but their output always requires careful review and interpretation by experienced investment professionals. Therefore, in our view, the human analyst's expertise, critical reasoning, and intuition remain indispensable in translating computational analysis into highest quality investment decisions.

Conclusion

In summary, this paper has shown how the evolution of natural language processing – from early word embeddings like GloVe to advanced transformer models such as FinBERT – is reshaping the landscape of quantitative finance. By moving beyond static, context-free approaches to models that understand language in its context, we have been able to extract actionable insights from unstructured financial data.

Our experience demonstrates that off-the-shelf models, while powerful, often fall short in meeting the specific demands of financial analysis, particularly when it comes to sector bias, compliance, and data privacy. By developing and tailoring our own models, we have meaningfully improved our sentiment analysis, as evidenced by the simulation that outperformed the S&P 500 on a sector-neutral basis.

As these technologies continue to advance, our focus will remain on refining our models, integrating new capabilities, and responsibly applying AI-driven insights to our Systematic Equity strategies like Best Styles and Powered by AI strategies. Ultimately, the combination of domain-specific AI and human expertise positions us well to decode financial narratives more effectively and to uncover new sources of value in ever-evolving markets.

Footnotes:

[1] Curse of dimensionality: it refers to the various challenges and complications that arises when analyzing and organizing data in high-dimensional spaces (often hundreds or thousands of dimensions). In the realm of machine learning, it's crucial to understand this concept because as the number of features or dimensions in a dataset increases, the amount of data we need to generalize accurately grows exponentially. Source: **https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning**

[2] **https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf**

[3] GloVe: Global Vectors for Word Representation. **https://nlp.stanford.edu/projects/glove/**

[4] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. arXiv:1706.03762

[5] **https://arxiv.org/abs/2303.17564**

[6] Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A Large Language Model for Extracting Information from Financial Text." Contemporary Accounting Research (2022). **https://doi.org/10.1111/1911-3846.12832**

[7] **https://arxiv.org/html/2306.02136v2**

[8] Net sentiment is calculated as log [(1+number of positive sentences) / (1+ number of negative sentenced)]. This logarithmic formulation helps smooth extreme values and ensures the metric is defined even when no positive or negative chunks are present.

[9] **https://doi.org/10.48550/arXiv.2203.15556**

[10] Kim, A., et al. (2024). Financial statement analysis with LLMs. arXiv:2407.17866

[11] Shojaee et al. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity **https://doi.org/10.48550/arXiv.2506.06941**

October 2025 LDS-250206