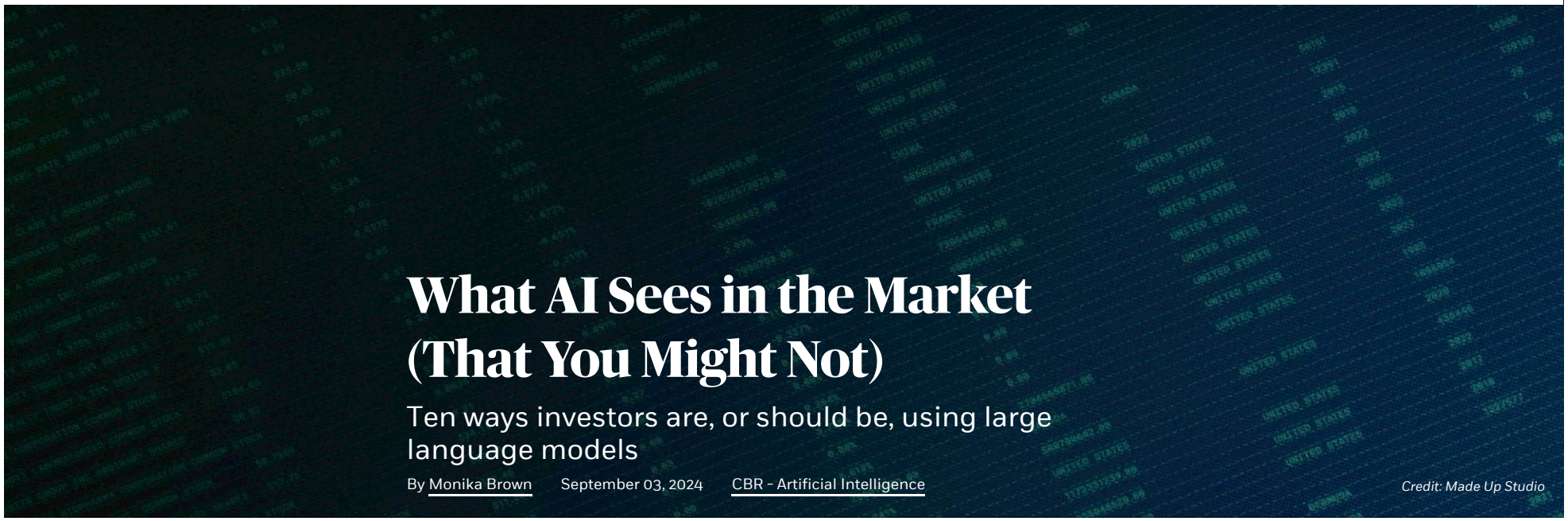
Chicago Booth Review



What AI Sees in the Market (That You Might Not)

Ten ways investors are, or should be, using large language models

By Monika Brown September 03, 2024 CBR - Artificial Intelligence

Credit: Made Up Studio

Great investors tend to be avid readers, always hunting for some piece of information to give them a financial edge. There are decades if not centuries of examples of pros who have combined something they've read—in a book, article, or regulatory filing—with their market experience to gain a lucrative insight. For one example, investment manager Jim Chanos's careful reading of Enron's regulatory filings, and his past experience with fraud detection, led him to suspect accounting irregularities at the company. He made $500 million when Enron filed for bankruptcy in 2001.

These days, though, even the most avid readers would have trouble competing with the volume of financial insights that artificial intelligence, in the form of

large language models, can uncover. LLMs have gained mainstream popularity thanks to OpenAI's ChatGPT, an advanced chatbot powered by a series of generative pretrained transformer language models. OpenAI has released several versions of its LLM, with GPT-3.5, GPT-4, and GPT-4o among the most recent.

Almost a decade ago, *Chicago Booth Review* published a feature titled "Why words are the new numbers" about a coming revolution in text analysis. That predicted revolution arrived, and it demolished the monopoly that numbers long held in forecasting models. Numbers are still important, of course—but text analysis is ascendant and everything is now potential data.

The candid speech during earnings calls? Data. The formal prose of annual filings? Data. News articles? Data. The entire internet? Data.

LLMs are trained on vast amounts of text covering a broad range of information and can apply their repositories of knowledge to evaluate new information. Where a human will depend on past experience and intuition, LLMs use data and patterns from their training.

And they operate at a scale that exceeds human capabilities, quickly analyzing mountains of text and allowing traders and investors to mine insights faster and more accurately than was ever possible. They can connect ideas from different parts of a text to create a better understanding of its overall content. LLMs can even be customized, trained to become experts on accounting irregularities—or, say, mall leases or risk management.

Every asset manager with a technology team now has the opportunity to wield—and profit from—an enormous knowledge base, and many are doing just this. Funds are using LLMs to read and glean insights from earnings call transcripts, 10-K regulatory filings, annual reports, social media, and streaming news headlines—searching for clues about a company's direction.

From the output of this text mining, LLMs can create direct trading signals (instructions to buy or sell) or develop new predictive variables for their forecasting models. If you hold actively managed funds in your retirement accounts, there's a good chance the pros running the strategies are harnessing the research power of LLMs.

It makes sense to ask whether the advantages of LLM strategies will disappear as soon as everyone else uses them too. That's been the outcome with arbitrage strategies—their returns fall when too many investors are chasing the limited opportunities. However, the opportunities here appear more bountiful than in arbitrage scenarios. With the field in its early stages, researchers are still finding new ways to apply AI to tease out investment insights and trading opportunities. Plus, new data sources that run the gamut from text to image, audio, and video are enabling the uncovering of information that is not so easily priced into the markets.

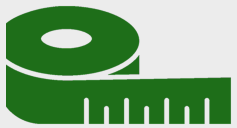
Researchers, like traders, are scrambling to stay ahead of the curve. Here are 10 of their recent observations.

How LLMs can help investors

Click below to jump to a section

LLMs can . . .




. . . produce useful summaries.

. . . identify corporate risk.

. . . see through boilerplate tax disclosures.

. . . measure and identify corporate policies.

. . . improve stock return predictions.

. . . combine insights from text with numbers to produce more accurate predictions.

. . . predict earnings more accurately than humans.

. . . signal stock crashes.

. . . be trained specifically for finance and other domains.

. . . find new information in disclosures.

1 LLMs can be trained specifically for finance and other domains.

Modern LLMs have significantly advanced the capabilities of natural language processing, essentially learning from giant data sets that represent a large swath of human knowledge. But some research indicates it may be possible to create more specialized, domain-specific LLMs that, at times, outperform the general-purpose models such as GPT-4.

Fine-tuning a smaller model has benefits beyond just customization for a particular task. It also lowers the computing costs, improves data privacy, and produces a tool that runs much faster than general-purpose models—possibly even on mobile devices.

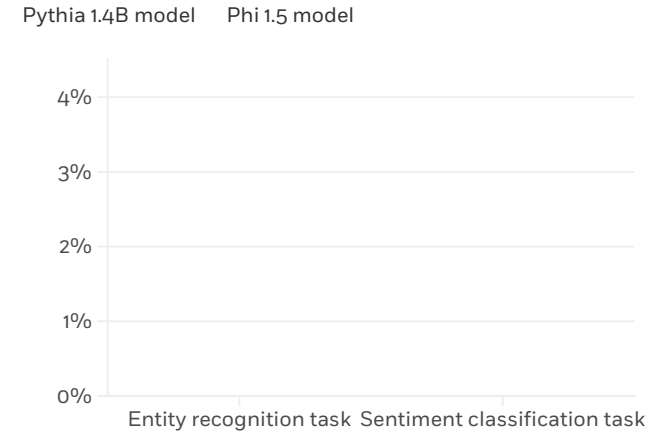
Motivated by this idea, Chicago Booth research professional Siyan Wang and Booth's Bradford Levy created a finance-focused data set, called BeanCounter,

which contains over 159 billion tokens extracted from corporate disclosures filed with the Securities and Exchange Commission. (A *token* is a word or part of a word.) For reference, OpenAI has disclosed that it trained GPT-3 using 300 billion tokens, primarily from Common Crawl, a nonprofit repository of online data. The researchers note that BeanCounter contains less than 0.1 percent of the data in Common Crawl–based data sets. What's more, they examined content directed toward various demographic identities and find that the content in BeanCounter tends to be substantially less offensive or harmful and more factually accurate.

Could a smaller data set ever produce an LLM whose performance could match that of GPT-4 or a similarly broad model? Wang and Levy say they have evidence that their LLM trained on BeanCounter actually does better. They used it to continuously pretrain two existing small, open-source LLMs. In finance-related tasks including sentiment analysis, the models pretrained on BeanCounter showed performance improvements over their base models. Both models also registered an 18–33 percent reduction in the level of toxic text generated after being updated with the data set.

Data quality matters a lot, says Levy, arguing that an LLM trained on fewer data points can perform well if they're high quality. The findings highlight the capabilities of smaller LLMs that are customized for various tasks or domains —and that work faster and cost less than large, generalized models.

Improvements in performance of finance-related tasks with additional training on BeanCounter data



Wang and Levy, 2023

CBR

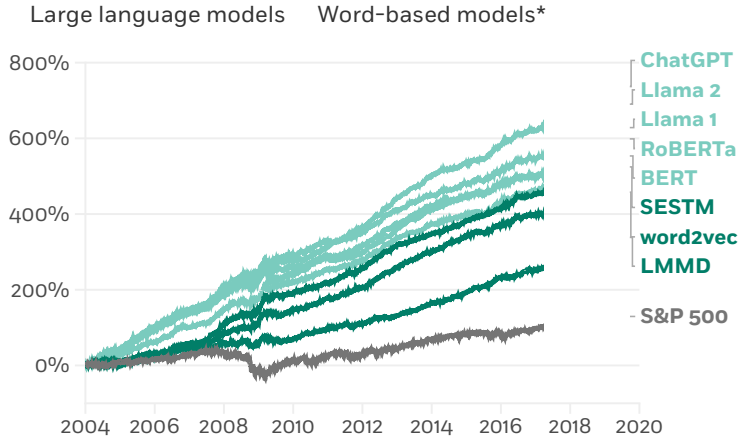
2 LLMs can improve stock return predictions.

When it comes to analyzing the sentiment of news articles, LLMs are far better than other models that came before them, research suggests. Booth PhD student Yifei Chen, Yale's Bryan T. Kelly, and Booth's Dacheng Xiu applied LLMs and traditional word-based models such as Google's word2vec to analyze the sentiment of business-news articles and their headlines in 16 global equity markets. And when they used the sentiment scores to make stock return predictions, the portfolio informed by LLMs outperformed those of word-based models. Thanks to their nuanced grasp of the meaning and structure of language, LLMs demonstrated a better comprehensive understanding of what was being said, and this led to a deeper interpretation of the news and greater predictive accuracy.

Models for predicting stock returns typically rely on variables focused on a company's characteristics, financial data, and historical returns. By creating a news sentiment variable and adding it to a predictive model, Chen, Kelly, and Xiu introduced an alternative data source, which also provided an opportunity for the model to capture additional data. For example, any information released overnight was missed by past return variables but contained in the sentiment variable.

Their research reveals a pronounced short-term momentum effect linked to news and suggests LLMs may offer promising opportunities for investors wanting to capture news sentiment in their models. Their simulations for larger LLMs such as RoBERTa (similar to one of the best-known LLMs, BERT—bidirectional encoder representations from transformers—but trained on a larger and more diverse data set) and Llama 1 and Llama 2 by Meta (similar to OpenAI's GPT-based models) achieved exceptional risk-adjusted returns. They saw Sharpe ratios above 4, a level that proprietary trading funds eagerly seek.

Cumulative returns of portfolios sorted by sentiment scores from different models



*SESTM = Sentiment extraction via screening and topic modeling; LMMD = Loughran-McDonald Master Dictionary

Chen et al., 2023

3 LLMs can produce useful summaries.

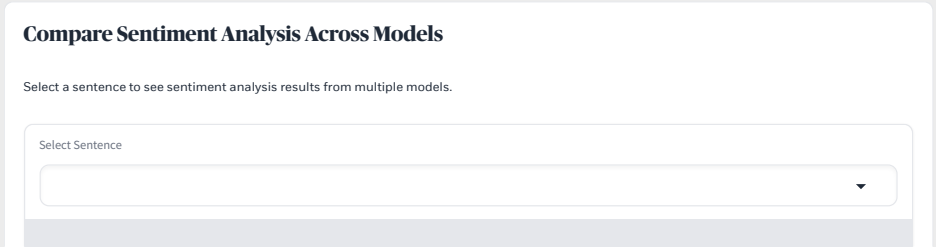
Companies disclose a lot of unstructured textual information in annual reports, with the management discussion and analysis sections found in 10-K filings being a salient example. ChatGPT can quickly distill the gist of what's being shared by summarizing both the MD&A and earnings call transcripts, research demonstrates.

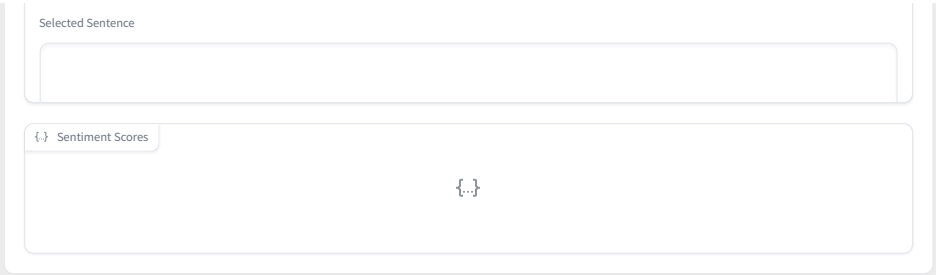
In a study, Booth researchers Alex Kim, a PhD student; Maximilian Muhn; and Valeri Nikolaev demonstrate how using GPT-3.5 Turbo can enhance clarity by stripping away boilerplate language, generic phrases, and less relevant details, offering a more accurate reflection of investor-relevant sentiment contained in complex corporate disclosures.

The researchers find that the sentiment of the GPT-3.5 Turbo-based summaries of the earnings announcements and 10-K filings, as opposed to the sentiment of the raw text, better explained the contemporaneous abnormal returns that resulted from investors reacting to these events. This suggests an opportunity for investors to use LLM summaries to enhance signals for trading around earnings calls and the release of 10-Ks. (For more, read "ChatGPT could help investors make more informed decisions.") Indeed, over the past year, several AI startups have emerged that generate summaries and allow customers to query corporate filings and communications.

How does AI decide if news is good or bad?

Select a sentence from the dropdown below to see how various models assess its sentiment. Notice how BERT, despite being a relatively small LLM, is more adept at capturing contextual nuances (particularly negative ones) than the three machine learning models. This example showcases the different capabilities of models for sentiment analysis, but is not an exact reproduction of the research methodologies discussed in this article.





4 LLMs can identify corporate risk.

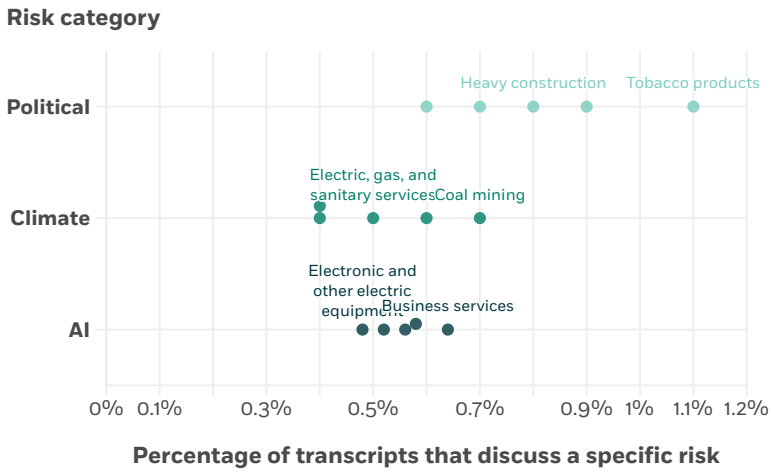
Some information is relatively straightforward to extract from corporate filings and earnings calls, but other information is trickier—such as certain types of risks facing a company. After all, managers on a public earnings call aren't usually keen to highlight their challenges.

However, in another paper, Kim, Muhn, and Nikolaev demonstrate the potential of LLMs to detect hard-to-quantify corporate risks, even when those risks are only indirectly disclosed. Their research suggests that an LLM can make inferences by using the vast amount of information on which it is trained to pick up on complex and nuanced relationships between statements scattered throughout a transcript's text.

The research focused in particular on political, climate, and AI risks affecting corporations. In the past, a number of different researchers have tried to pull insights on corporate risk from earnings call transcripts using natural language processing. They've had somewhat limited success due to executives' careful language choices and the algorithms' inability to understand the deeper context of what's being discussed. For example, a call transcript may contain a risk-related discussion without explicitly mentioning risks anywhere.

But GPT-3.5 Turbo connected statements made throughout a transcript and leveraged its vast knowledge base to infer, or read between the lines to discover, the risks, find Kim, Muhn, and Nikolaev. As a result, they say, the risk measures GPT-3.5 Turbo produced were capable of more accurately predicting volatility in a company's stock price following its earnings call. The LLM was even able to capture newly emerging risks that were not commonly seen in its training data, including risks associated with AI itself. (For more, read "AI reads between the lines to discover corporate risk.")

Industries with the highest LLM-based risk assessment scores



Kim et al., 2023

CBR

5 LLMs can find new information in disclosures.

In a different spin on making sense of lengthy, dense corporate disclosures, Booth's Anna Costello, Levy, and Nikolaev developed

LLMs that can spot new information. Rather than summarize documents, their method retains the original content and instead highlights the portions that are likely to be surprising.

The researchers first built a base LLM pretrained on financial disclosures from various companies, text that altogether totaled more than 35 billion tokens. They then created a firm-specific LLM for each company by further training the base model on that company's past regulatory filings. By using only contemporaneous data, the researchers made sure their measure learned solely what investors could have known about a company at the time.

Information theory holds that surprising events are those that investors assign a relatively low probability of happening. Along these lines, LLMs work by modeling the probabilities around the next word (or partial word) in a sequence of text. The researchers precisely measured the level of surprise associated with each word in a filing relative to the content on which the LLM had been trained.

They then used the notion that prices should fully reflect all publicly available information to validate their measure, finding that it explains a large portion of the short-term market reaction to corporate filings and is predictive of future returns. This future predictivity is small, however—consistent with market efficiency and limits to arbitrage, they write.

While this work is focused on corporate disclosures, the researchers say that their method is general enough to be applied to other settings such as supply-chain contracts and legal documents, or even other modalities such as images and video. Given the novelty of the method, notes Levy, the University of Chicago is pursuing a patent on the technology.

Example of new information identified by a model fine-tuned on Apple's disclosures

In a disclosure filed by Apple that mentions the iPhone for the first time

Degree of new information Low Medium High

CUPERTINO, California—January 17, 2007—Apple® today announced financial results for its fiscal 2007 first quarter ended December 30, 2006. The Company posted record revenue of $7.1 billion and record net quarterly profit of $1.0 billion, or $1.14 per diluted share. These results compare to revenue of $5.7 billion and net quarterly profit of $565 million, or $.65 per diluted share, in the year-ago quarter. Gross margin was 31.2 percent, up from 27.2 percent in the year-ago quarter. **International sales accounted for 42 percent of the quarter's revenue.**	Low Medium Low
Apple shipped 1,606,000 Macintosh® computers and 21,066,000 iPods during the quarter, representing 28 percent growth in Macs and 50 percent growth in iPods over the year-ago quarter.	Low
"**We are incredibly pleased to report record quarterly revenue of over $7 billion and record earnings of $1 billion," said Steve Jobs, Apple's CEO**. "**We've just kicked off what is going to be a very strong new product year for Apple by launching Apple TV and the revolutionary iPhone.**"	Medium High
"**We generated over $1.75 billion in cash during the quarter to end with $11.9 billion," said Peter Oppenheimer, Apple's CFO**. "Looking ahead to the second fiscal quarter	Medium

of 2007, we expect revenue of $4.8 to $4.9 billion and earnings per diluted share of $.54 to $.56."

Apple will provide live streaming of its Q1 2007 financial results conference call utilizing QuickTime®, Apple's standards-based technology for live and on-demand audio and video streaming. The live webcast

Low

There's more to this story

The Evolution of AI in Finance

The technological landscape has changed rapidly for investors.
CBR - Artificial Intelligence

Images and Audio Are Now Data Too

The revolution that turned words into analyzable data continues to progress.
CBR - Artificial Intelligence

6 LLMs can predict earnings more accurately than humans.

Financial-statement analysis requires quantitative skills, logical reasoning, critical thinking, and complex decision-making—so one might think that it's a domain in which humans still have a leg up on LLMs.

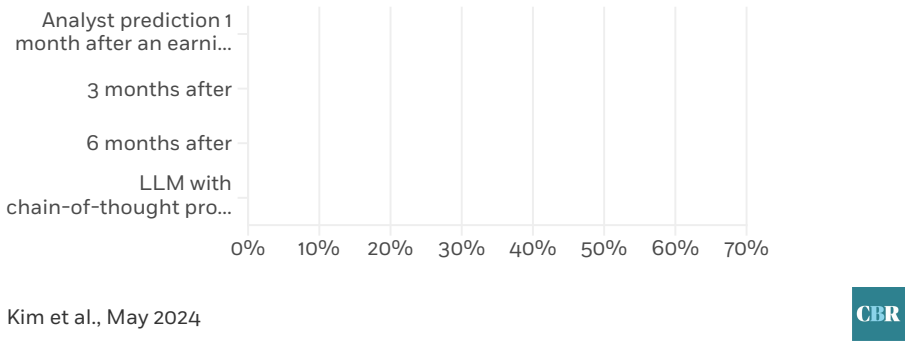
Research suggests this may soon change. Kim, Muhn, and Nikolaev find that LLMs, and specifically GPT-4 Turbo, can simulate professional financial-statement analysis, and in a way that outperforms humans.

The researchers provided GPT-4 Turbo with anonymous balance sheets and income statements with the management discussion and analysis sections removed so that the LLM did not have any textual clues. They standardized the statements by making sure all labels matched a template and replacing dates with *T*, *T-1*, and the like.

Then they used a "chain of thought" prompt to instruct the LLM to solve problems step-by-step and with reasoning as a human would. They wanted it to use the same thought process as an analyst in noting trends in the statements and computing key financial ratios. While earlier LLMs including GPT-3.5 are notoriously bad at math, the newer GPT-4 Turbo leveraged its understanding of math concepts and combined its computations with economic reasoning to deliver insights about companies, the researchers find.

Its predictions outperformed the consensus forecasts of professional financial analysts (60 percent accuracy versus 53 percent). Furthermore, its accuracy was on par with a sophisticated machine-learning model specifically trained to predict the direction of earnings.

The paper suggests LLMs have a relative advantage over analysts, who in some instances may struggle to come up with an accurate forecast (and hence issue differing forecasts) or display bias. In further study, the researchers find that human analysts' forecasts complemented GPT-4 Turbo's forecasts, indicating that professionals still provide valuable insights about companies and markets that aren't reflected in financial statements. They conclude that LLMs have the potential to play a central role in financial decision-making—complementing humans rather than replacing them.



7 LLMs can signal stock crashes.

LLMs can predict key financial and economic indicators, finds a study by Booth's Leland Bybee. His research creates a method for doing so by applying an LLM to news articles and then forecasting financial and economic measures such as the S&P 500 and the Consumer Price Index.

Applying the method to 100 years of news articles, Bybee produced a time series of economic beliefs. The predictions made by the LLM aligned closely with those recorded in investor and CFO surveys, as well as with equity fund flows.

And when he used the method to investigate behavior during financial bubbles, he finds that the more sentiment (rather than fundamentals) fueled a rise in an industry's stocks, the higher the probability of a crash and lower future returns. This suggests that sentiment-driven mispricing can predict bubbles.

Bybee tested these findings with an estimated trading strategy that held portfolios of stocks that were all from the same industry, were generally rising, and were predicted not to crash. That prediction was based on a cutoff threshold in the sentiment measure the researcher produced. This LLM strategy successfully avoided 80 percent of the industries that had sharp downturns and significantly outperformed a similar but more naïve strategy.

Cumulative returns on a portfolio of industry stocks with low crash risk

Bybee, 2023

CBR

How does AI parse corporate communications?

Below, a ChatGPT prompt analyzes earnings call transcripts from one quarter to the next and determines the direction and extent of the company's policy changes. The highlighted phrases are those that the LLM found to be

significant in its analysis. This example demonstrates how LLMs can identify policy shifts in earnings call transcripts, though it differs from the research described in this article, which focuses more on language discussing future changes.

ChatGPT's analysis of earnings call transcripts on capital spending over the next year

Citicorp | Chevron | Intel

Q4 2022

Small increase: The company has indicated a focus on investing in transformation and technology, with a 5% increase in technology-related expenses. They emphasize that these investments are crucial for enhancing their infrastructure and efficiency. This indicates a small increase in capital expenditures, particularly focused on technology and transformation initiatives.

. . . **And we are investing in technology across the firm, with total technology-related expenses increasing by 5%.** While we recognize this is a significant increase in expenses, these are investments we have to make, and these investments will make us a better, more efficient company in the future. . . .

Q1 2023

Unchanged: The company's focus remained on maintaining strong liquidity and capital, with no significant changes in capital expenditure levels mentioned. The emphasis was on continuing existing strategies rather than expanding investments.

. . . About 35% of our balance sheet is in cash and investment securities which contribute to our $1 trillion of available liquidity resources. **And at the end of the quarter, we had an LCR of 120%, which means we have roughly $100 billion of HQLA in excess of the amount required by the rule to cover stressed outflows.** And you can see the details of this on page 27 in the appendix. But just as important as the quantum of liquidity is the composition and duration of that liquidity. . . .

8 LLMs can see through boilerplate tax disclosures.

Tax audits are hugely important for companies and their investors, but you wouldn't know it from the boilerplate tax disclosures companies voluntarily make in corporate filings. Is an audit imminent? Has it just concluded? Are authorities about to levy a fine or challenge a company's tax-planning strategies? This has all been difficult for investors to figure out.

However, LLMs can make sense of the hard-to-parse disclosures and extract useful signals for investors, suggests research by City University of London's Ga-Young Choi and Booth PhD student Kim.

The researchers used GPT-4 to analyze about 20,000 10-K filings from 2010 to 2021, extract each company's relevant tax and audit information, and track the changes in language from one year to the next.

These differences may indicate some potential corporate risk, according to the study. Active tax audits effectively deterred tax avoidance but led to increased stock volatility and reduced capital spending, the researchers find. Even after an audit had concluded, companies in their sample tended to continue to decrease tax avoidance strategies, capital investments, and new debt issuance.

The research suggests that an LLM can be applied to corporate disclosures to tease out a company's current audit status and help anticipate and avoid any potential related fallout.

9 LLMs can measure and identify corporate policies.

On earnings calls, investment policies often aren't stated simply, or even at all. For example, an executive may say, "We are investing in

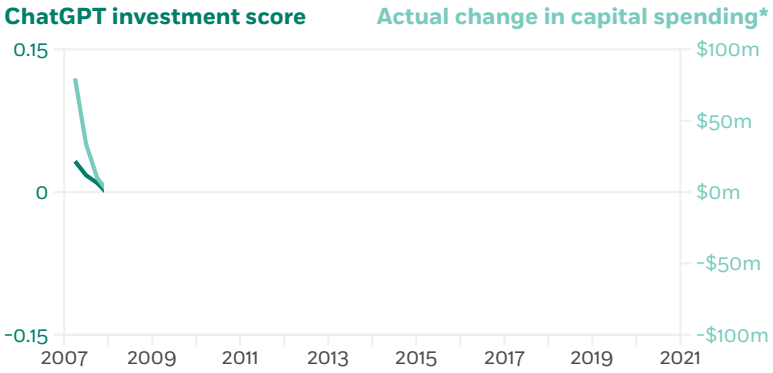
growth initiatives." While the line doesn't directly state as much, this might imply some large, upcoming capital expenditures that could affect near-term profitability.

Researchers from Georgia State, Manish Jha, PhD student Jialin Qian, and Baozhong Yang, along with Booth's Michael Weber, designed a method using ChatGPT that they suggest is capable of discovering sometimes-hidden policies. The underlying LLM can analyze call transcripts and predict future corporate policy changes—such as shifts in capital investments, dividend levels, or head count, their research finds.

The researchers used ChatGPT to generate a likelihood score for changes in corporate policies. The score was validated by its alignment with CFO survey responses about corporate investment plans. Their method's predictions for capital spending and the actual capital expenditures were highly correlated.

ChatGPT was able to decipher the corporate policy changes from the transcripts with a high degree of accuracy, the researchers write. The scoring system they devised could serve as a tool for investors by revealing potential corporate policy shifts not fully priced into the market. In the research, high investment scores were linked to notable negative abnormal returns over subsequent quarters, suggesting this tool can offer an advantage in portfolio management, especially in conjunction with other analyses such as Tobin's Q, which is used by investors to evaluate corporate policies. (For more, read "AI can discover corporate policy changes in earnings calls.")

ChatGPT investment score compared with realized investment from a sample of companies



*Difference in companies' average capital spending four quarters after and four quarters before the current quarter

Jha et al., 2023

CBR

10 LLMs can combine insights from text with numbers to produce more accurate predictions.

LLMs can incorporate the textual information in the MD&A section of a 10-K filing to enhance the value of the numerical information disclosed by a company, suggests research by Kim and Nikolaev. They used BERT to contextualize accounting numbers by incorporating textual information and find that this improved the accuracy of predictions about future earnings, cash flows, and stock returns.

Specifically, integrating textual information about demand trends and strategic plans for a company with the numerical data about profitability improved the model's performance compared with using solely numerical or

textual data, according to one of two related papers they wrote on the topic. Also, predictions of share prices and portfolio performance improved when the researchers included in their model a measure that they created, context-adjusted profitability.

The findings suggest that investors can improve their strategies by using LLMs to incorporate these textual data. (For more, read "Large language models can improve stock market forecasts.")

Kim and Nikolaev, 2023

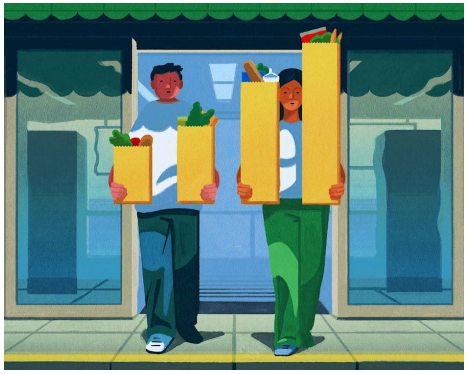
The takeaway: Make use of LLMs

AI is transforming practically every sector of the economy. The technological progress being made has implications for everything from job recruiting to medical diagnosis to filmmaking. (For more, read "AI is going to disrupt the labor market. It doesn't have to destroy it.") In finance, LLMs are mining public data to find varied and largely unexploited investment opportunities—and are evolving from being analytical

tools to capable decision-makers, paired with investors in the ongoing hunt for profit.

Works Cited

More from Chicago Booth Review



People Can Forecast Price Rises—If Asked the Right Questions

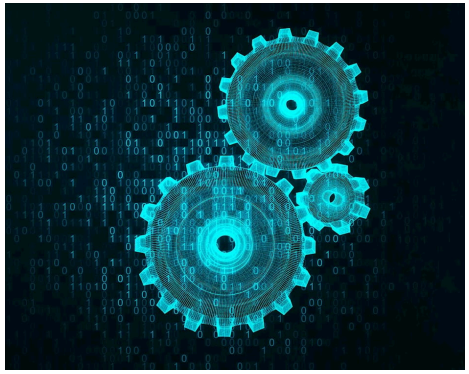
While consumers are a little hazy about overall inflation, asking them about prices for individual categories yields more realistic forecasts.

CBR - Economics

The Two Big Strategic Mistakes That Investors Make

Research finds a discrepancy between what people plan to do when trading—and what they actually do.

CBR - Finance



For A.I. Companies, Technology Is Inherent in Strategy

Business decisions have to be made with one eye on the algorithm.

CBR - Artificial Intelligence

Related Chicago Booth Review Topics

CBR - Technology CBR - Artificial Intelligence CBR - Fall 2024 CBR - Finance

More from Chicago Booth

A Legacy of Leadership and Mentorship in Healthcare

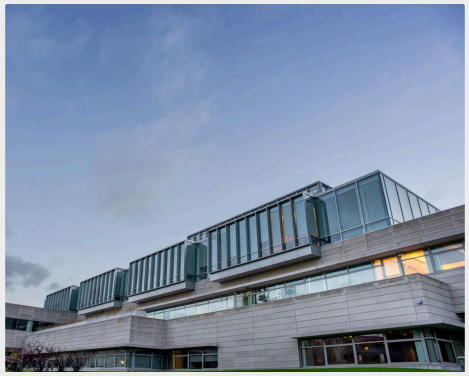
With over four decades of experience in the healthcare industry, Bob Atlas has held influential roles while dedicating himself to mentoring the next generation of professionals.

Healthcare Initiative

Expanding Diversity in Economics Program Moves to Booth

The Expanding EDE+ program, which aims to diversify the pool of students pursuing degrees and careers in economics, is undergoing an exciting transformation.

Economics



Healthcare Initiative Advisory Circle

The Advisory Circle for the Healthcare Initiative at the University of Chicago Booth School of Business plays a pivotal role in guiding and supporting our mission.

Healthcare Initiative

Related Chicago Booth Topics

Technology Finance Artificial Intelligence

Get more Chicago Booth Review

First Name*

Last Name*

Email*

* I agree to receive electronic communications from the University of Chicago. I understand I may unsubscribe at any time.

SUBMIT

Your Privacy

We want to demonstrate our commitment to your privacy. Please review Chicago Booth's privacy notice, which provides information explaining how and why we collect particular information when you visit our website.

Chicago Booth Review

Explore CBR

+ All CBR Topics | About Us | All Issues | Contact Us | Chicago Booth

Privacy Notice | Accessibility

Follow

Chicago Booth Review

Research driven insights on business, policy, and markets.