MORNINGSTAR®

Scaling AI-Assisted Research

A framework for responsible adoption of large language models.

Morningstar Quantitative Research
December 2023
Version 1.0

Contents

Sabeeh Ashhar
Director of Quant Research
sabeeh.ashhar@morningstar.com

Prashant Srivastava
Lead Quant Analyst
prashant.srivastava@morningstar.com

Abhinav Chhabra
Data Scientist
abhinav.chhabra@morningstar.com

Executive Summary

Over the past few years, the field of investment research has experienced a profound revolution, fueled by the emergence of artificial intelligence algorithms and large language models, or LLMs. These advanced technologies have facilitated the analysis of vast quantities of financial data, allowing investment researchers to uncover valuable insights given specific instructions. By processing diverse streams of information, such as company reports, news, fundamental data, and market trends, LLMs have become indispensable tools that streamline research tasks and provide investors with comprehensive perspectives on market trends and potential opportunities. Fine-tuning LLMs for investment research tasks is challenging, and Retrieval Augmented Generation, or RAG, systems have gained prominence by mitigating issues such as hallucinations with access to external databases for context-specific information.

To this end, our study goal is twofold. Initially, we discuss challenges with machine-generated information using RAGs within investment research on various tasks, emphasizing the crucial need for human evaluations. Subsequently, considering the difficulties in scaling human evaluation, we explore automated metrics for scalable evaluation of machine-generated content. The key takeaways from this study will aid stakeholders in identifying optimal usage of LLMs for investment research automation.

Key Takeaways

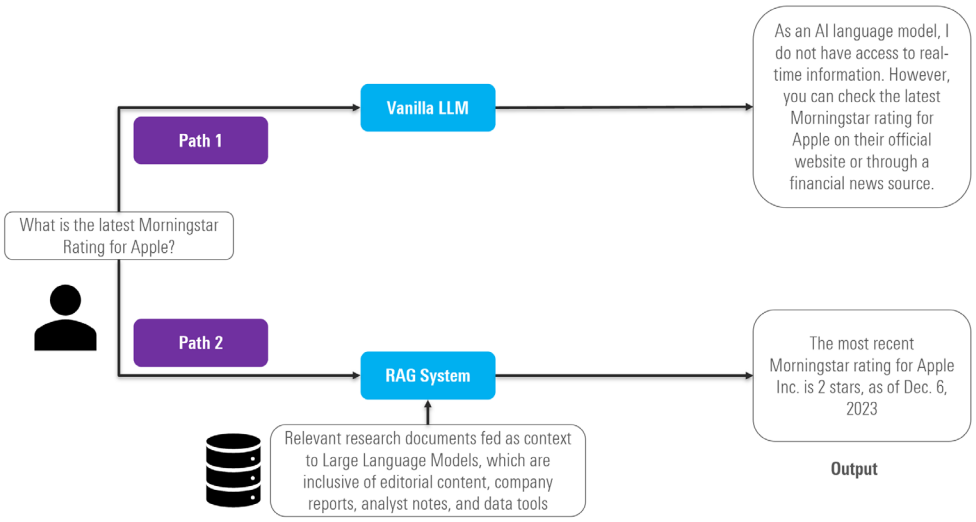
- Machine-generated text displays higher efficacy on simpler information retrieval and text summarization tasks, holding promise to augment the efficiency of analysts.
- Complex arithmetic calculations and logical-reasoning intensive research tasks remain challenging for LLMs today, with the need for continued expert human oversight due to factual knowledge gaps.
- Automated evaluation of machine-generated text using LLMs themselves yield a scalable and cost-efficient approach, aiding adoption of this technology. Based on our experiments there is 80% alignment between LLM-aided evaluations and human assessments.
- Based on experiments, GPT-4 model come out on top for text generation and evaluation of investment research tasks.
- Investors can adopt these emerging technologies responsibly to maximize return on investments by starting with well-defined use cases that have proven benefits and gauge the learnings before expanding scope to more complex tasks. Tracking the rapid incremental innovations in this space should still be a priority.

Introduction

Over the last decade the field of Natural Language Processing, or NLP, has rapidly evolved, leading to groundbreaking research such as large language models, or LLMs. In November 2023, ChatGPT, an AI assistant, marked its inaugural anniversary, demonstrating evolving abilities to solve intricate tasks swiftly through few-shot examples (Zhao and others 2023). The year 2023 also saw the release of new age foundational models like GPT-4-Turbo, Claude-v2.1, and LLAMA (with 65 billion parameters), finding applications in tasks ranging from reading comprehension, open-ended question/answering, named entity recognition, and code generation. While these models were made on general purpose datasets, the recent research efforts have explored the adaption of these models for domain-specific tasks (Suzuki and others 2023).

However, there is limited research on the application of LLMs for the investment research domain. Current research has mostly focused on evaluating LLMs for passing financial analyst exams (Ethan and others 2023). Other research focused on siloed tasks such as numerical reasoning (Chen and others 2022). There are opportunities for exploration across general investment research tasks encompassing the day-to-day work of analysts. There are practical challenges for LLM adoption here, ranging from diverse data formats and types, unique linguistic styles, domain-specific intent, and entity identification alongside evolving datasets. Hallucinations characterized by factually incorrect information is also a commonly detected issue. In an era marked by cost constraints, asset managers and advisors will find it challenging to allocate resources for provisioning computational resources and managing ongoing model maintenance costs, making the fine-tuning of models quite difficult. To address some of these challenges, retrieval-augmented generation, or RAG, systems have emerged, which fetch up-to-date or context-specific data from an external research database and make it available to an LLM during the text-generation process.

Exhibit 1: Overview of RAG Systems



Source: Morningstar.

Exhibit 1 above provides a quick overview of RAG systems. These systems can cite their sources, thereby improving auditability and transparency, which is a key requirement for regulated investment research entities.

As part of our research, we hope to uncover the challenges with the usage of RAG systems for practical investment research tasks ranging from information mining, text condensation, code generation, general research question and answering, numerical reasoning, and drafting narratives. These specific tasks consume most of the analyst's time during the content-generation process. Evaluation of machine outcome is first performed by humans to uncover potential challenges with machine-generated content. Given the challenges with scaling human evaluations, there is a need to develop algorithmically derived metrics. Additionally, we also test various LLMs like GPT-4, GPT-3.5, and Mistral-7b for these evaluations. The conclusions will serve as a guide for effective adoption of this emerging technology within the investment research domain.

Experiment Setup

Our evaluation begins with an exhaustive exploration of the diverse array of practical investment research tasks. These tasks span increasing levels of complexity and are described below:

Information Mining: This is the process of extracting vital financial information from unstructured textual sources like financial reports, news articles, or social media posts. Compared with other tasks described below, this task involves the least level of analysis and investment knowledge. Some examples involve fetching expense ratio of funds from fund fact sheets or financial data points from stock SEC filings.

Text Condensation: This involves condensing large volumes of information, such as financial reports, news articles, or research papers, into concise, easily digestible summaries, tear sheets, and so on. Compared with other downstream tasks, here the machine works by condensing quality, curated human text. Some examples involve condensing analyst narratives, unstructured text (risk factors, management discussion, and business outlook) in corporate filings, and so on.

General Research Question & Answering: This task involves generating information that is intended to help investors understand investments and associated services. These range from addressing questions about investment terminologies and methodologies, understanding different assets categories, and so on. Some examples here involve understanding risk/return trade-off of various asset classes or understanding terminologies like value versus growth investing.

SQL Code Generation: Investment analysts are frequently required to write SQL codes over data lakes to fetch data points, screen universes, draw inferences, and so on. We tested the task of writing SQL code to screen universes, fetching data points, and performing aggregation calculations over various universes. We limited this exercise to the most popular data points available in Morningstar Direct that pertain to equities and funds. Compared with information mining, the task involves analysis on vast universes of equities and funds. Some examples here include fetching returns, flows, and category information of funds and equities.

Numerical Reasoning: This task involves drawing quick calculations from structured data assets. Compared with other tasks, this involves a high level of financial domain knowledge and the ability to perform analysis by crunching numbers. Some examples here involve calculating growth figures over past time periods, returns, and flows.

Drafting Narrations: This task involves drafting comprehensive and in-depth analysis of investment opportunities, studying, and analyzing the performance of various financial instruments to provide the view of potential returns, risk factors, and other relevant factors. It goes beyond just looking at numbers and incorporating domain knowledge and logical inferences to answer queries. Some examples here involve comparing analyst narratives for two funds or defining structure fitment of risky assets such as cryptos in client portfolios.

For our analysis, we have curated a list of over 1,250 real-world questions and answers that are representative of the above tasks. The evaluation data set has been curated on live financial data using analyst notes, research papers, and filing documents available on Morningstar Direct. Also, for providing LLMs with context-specific data, we use RAG chunking configuration as chunk size of 500, chunk overlap of zero, and top-four stuffing strategy. We acknowledge that the above list of tasks may not be comprehensive and future research should include more, such as machine translation, financial model building, predictive analytics, and so on. Exhibit 2 describes the structure of the data and some characteristics, such as average prompt length for LLMs, alongside the percentage of numerical data and requirements for logical reasoning to solve the task.

Exhibit 2: Data Distribution of Investment Research Tasks

Investment Research Tasks	Average Prompt Length	% of Numerical Data	% Logical Reasoning
Information Mining	1796	48	12
Text Condensation	1878	24	16
General Research Question & Answering	1483	24	28
SQL Code Generation	1335	76	72
Numerical Reasoning	3386	84	88
Drafting Narrations	1643	48	92

Source: Morningstar.

Next, as part of machine-generated content evaluation for various tasks, we use different flavors of closed- and open-source LLMs such as GPT-4, Claude-v2, and Mistral-7b, which have shown impressive performance on various benchmarks. We further evaluated these models in zero-shot and few-shot prompt settings to check if the results improved.

Human Evaluations

The human evaluation is first conducted on machine-generated content to uncover potential gaps. Outcomes are evaluated on multiple dimensions such as relevancy, groundedness, and conciseness. Relevancy score ensures that generated responses are directly relatable to queries at hand. A groundedness score identifies links between generated output and context-specific data, thereby ensuing factuality or detecting hallucinations. A conciseness score relates to the analyst's writing style, conveying complex pieces of information in a concise and coherent way, enabling a better understanding and avoiding information overkill. Based on our experiments, the three metrics are mutually exclusive and least correlated with each other. For the purpose of scoring, we use a five-point grading scale between 1 to 5. We did not intend to use a higher grading scale (1-10) since it becomes difficult to distinguish between all scores. Also, we did not plan to use a binary grading (0-1) scale as this would lead to very trivial evaluation, and LLMs would not be able to distinguish between the complexity of content. Based on final score values, >=4 was marked correct and anything less than 4 was graded as unacceptable.

Automated Evaluation Process Defined

During our analysis we also identified that human evaluations are costly and not scalable, so we propose LLMs themselves to grade the response of machine-generated content. Recent research (Zheng and others 2023) points to the usage of LLMs for evaluation of other LLM outcomes. These metrics are described below. We evaluated metrics that cannot only grade responses but can also provide clear reasoning. The metrics are defined through a grading rubric provided to LLM in the form of a chain-of-thought prompt. Exhibit 3 below describes the rubric used for grading responses for relevancy metric.

For similar reasons described in the human evaluation section, we have also kept scores in the range of 1 to 5.

Exhibit 3: Example of Relevancy Score Prompt

Relevancy Score: Degree to which an answer directly addresses and is appropriate for a given question.

- Score 1: The answer is entirely irrelevant to the question and does not contain any information related to the investment topic or query.

- Score 2: The answer provides minimal relevance to the question, addressing only a single aspect of the query accurately, but it lacks comprehensive coverage.

- Score 3: The answer is somewhat relevant, addressing most aspects of the question correctly. However, it may miss or include some extraneous information that is not directly related to the query.

- Score 4: The answer is highly relevant, correctly addressing the main aspects of the investment question without missing any major component. It provides a well-rounded response.

- Score 5: The answer is not only highly relevant and accurate but also goes above and beyond by offering valuable additional insights, context, or supporting evidence that enhances the understanding of the investment topic.,

Source: Morningstar.

Exhibit 4 shows the LLM scores for the relevancy score metric. As seen, the evaluations can detect the aberrations in responses effectively.

Exhibit 4: Example Grades Generated by LLM

Question	Answer	Relevancy Score	Relevancy Score Reasoning
Why do YTD returns not show in my portfolio?	Direct indexing allows investors to personalize their portfolios by holding stocks directly instead of through a fund. This allows for customization based on specific orders or general requirements. Direct-indexing services simplify the process of assembling and maintaining a portfolio. It also provides portfolio tools that display overall asset allocation, investment-style breakdown, fees, and more.	1	The answer does not address the question about why YTD returns are not showing in the portfolio. Instead, it explains what direct indexing is and how it works, which is not relevant to the question asked.
How should I plan for retirement?	To plan for retirement, the first step is to run some detailed numbers to estimate whether you will have enough money to retire based on your current assets, expected spending, and estimated rate of return. It is also important to consider your retirement date and the potential benefits of working longer, such as delayed portfolio withdrawals, additional retirement plan contributions, tax-deferred compounding, and a larger Social Security benefit. You can seek guidance from a fiduciary advisor, such as a fee-only planner, to assess your retirement readiness and explore potential strategies.	4	The answer is highly relevant and accurate. It addresses all the main aspects of the question and provides additional insights and context. It also provides valuable advice on how to plan for retirement, such as considering your retirement date, evaluating anticipated lifestyle changes, and seeking guidance from retirement calculators and financial advisors.

Source: Morningstar.

Like human evaluations, we can grade machine-generated text on multiple metrics like relevancy, groundedness, and conciseness. As for the relevancy metric, we have prompts for groundedness and conciseness metrics. We finally take an average of the three metrics to arrive at consilience scores that are indicative of the efficacy of machine-generated content.

Next, we run each of the 1,250 questions in our sample set to generate metric scores mentioned above. We further count the proportion of the questions with good scores (4 and 5) to arrive at a final score.

Analysis of Results

Uncovering Challenges With Machine-Generated Text

Based on the experiment setup we discussed earlier, we have human evaluations for machine-generated outcomes across the model flavors. Exhibit 5 below demonstrates the efficacy scores across various research tasks.

Exhibit 5: Human Evaluations

Investment Research Tasks	GPT-4	Claude-v2	Mistral-7b
Information Mining	84.8%	76.8%	72.4%
Text Condensation	82.0%	70.0%	62.4%
General Research Question & Answering	74.8%	60.0%	51.6%
SQL Code Generation	64.0%	54.0%	42.4%
Numerical Reasoning	44.8%	32.8%	50.0%
Drafting Narrations	30.0%	22.8%	12.8%
Overall Efficacy	66.8%	52.7%	48.6%

Source: Morningstar.

Our human evaluations reveal challenges across all three model outputs in generating effective responses. This is described below for each of the task types:

Information Mining: The generated text fared well for information-mining tasks. This can be attributed to their inherent capacity to store and contextualize large databases. They are also aided by vector databases, which are effective at retrieving relevant text information and feeding to LLMs based on the question at hand. These models are also trained on diverse data formats, and with aid from the self-attention mechanism, they are better able to understand entities and relationships, incorporate contextual layouts, and extract pertinent details. Further, the nature of the task involves almost no arithmetic or logical reasoning, which also helped boost the accuracy. In line with the relevancy, groundedness and consciousness scores were above 80%. However, there are failures with multimodal data characterized by scanned images inside files. Failures were also detected in non-U.S. filings due to differing formats of files and nonstandard layouts. Finally, there were challenges with retrievers. With fixed-chunk sizes used, there were some chunks that demonstrated missing context, leading to incorrect outcomes.

Text Condensation: We also observed good outcomes on the text-condensation task. This can be attributed to well-curated human text used for the condensation tasks. Again, there was little to no logical inference required as most of information is clearly available within the input text. Context length was small here and well within the allowed context size of the models used. There are still challenges with summarizing large documents, and this is actively being researched (Sengjie Liu and others 2023). Some of the errors found here can also be attributed to a bias toward simpler text-generation lacking

strong reasoning and repetition bias due to the lack of domain understanding. Also, the writing styles were sometimes inconsistent, and garbage text was being generated at times. Due to this, while the relevancy and groundedness scores were high, the conciseness scores were lower.

General Research Question & Answering: The machine outcomes also fared well in the general research queries. However, as this task also involved higher domain knowledge, there were incorrect outcomes generated at times. This training data sparsity of foundational models limits exposure to investment domain terminologies during inference in a closed-domain setup. Like text condensation, there were issues with inconsistent writing styles and redundant information being generated.

SQL Code Generation: The machine outcomes started to deteriorate as we extended to code-generation tasks. Here there were frequent failures with intent and entity detection. Within our data there were over 5,000 companies and fund tickers. While foundational models were trained on data easily available for large entities, such as large-cap companies, there were issues with detecting lesser-known entities such as small-cap companies or lesser-known funds. There were also gaps detecting intents based on similar-sounding data points such as returns over various time periods. The engine also found it challenging detecting default intents when such information is missing in queries. Hence, answering such questions requires domain knowledge. These issues led to high levels of hallucinations in outcomes, although answers were relevant and concise at times.

Numerical Reasoning: The machine-generated outcomes lacked significantly here and did not have the ability to perform arithmetic calculations and draw reasoning. Some examples were failure to identify whether an analyst-rating grade of 5 is better versus 1. Similarly, for ESG risk scores, it was difficult to identify whether higher or lower scores are better. This limitation can be attributed to the inherent incapability of LLMs in arithmetic-reasoning tasks (Shima Imani and others 2023), contributing to observed lower efficacy. Unlike humans, LLMs lack an intuitive understanding of investment domain and the ability to internally represent and manipulate numbers, relying solely on surface-level semantic associations from their text-training data.

Drafting Narrations: The machine-generated narrations lacked coherent storytelling and contextual grounding found in analyst outcomes. Some examples where failed responses were generated included answering structure of thematic in investor portfolios or retirement planning. Further, the tasks displayed profound knowledge gaps and a misunderstanding of crucial terminology, the inability to map relationships between entities, and often having to rely on false assumptions while reasoning. Together, these limitations paint the picture of why today's language models fall short on delivering logically reasoned analysis expected from analysts who have years of lived experience for intuitively avoiding these pitfalls. Additionally, the uneven mixing of formal and informal styles raises concerns regarding suitability and consistency for end users when utilizing applications that demand accuracy and precision.

We also found iterations that refine chain-of-thought, or CoT, prompts through the description of associated data points and logical inferences resulted in reduced hallucinations and factual errors

compared with zero-shot prompts. Overall, we also noted variability in hedged (engine does not know anything) responses where engine was trying to play it safe with responses. Further, among all model versions, GPT-4 demonstrated superior efficacy, with Claude-2 following closely behind, and Mistral-7b lagging significantly. GPT-4 did not show remarkable improvement with few-shot prompts, while Claude-2 and Mistral-7b benefited with some examples. However, for complex tasks involving narration and deep research, notable enhancements were not seen across the board with even few-shot examples.

To overcome the above challenges, we propose fine-tuning a custom LLM that can align to specific writing style and tasks in finance (Chu and others 2023). The process involves the training of models on domain-specific data, allowing for customization of the model to generate responses that pull together relevant knowledge from documents or a knowledge base, while framing it appropriately for the brand's voice. Also, perfecting the RAG retriever parameters can also prove helpful. Research (Lewis and others 2020) has shown promise in improving the quality of automatically generated text output by perfecting RAG architectures through techniques and parameters such as chunk sizes, semantic similarity metrics, and reranking retrieved passages. Query expansion (Wang, Liang, and others 2023) also helps the retriever find more focused, salient passages to inform the generator. Recent research has also unveiled the potential of augmenting RAG systems with external tools, thereby significantly enhancing their problem-solving capacities and efficiencies (Yao and others 2023). These techniques help provide superior context-specific data as input to the generator, leading to higher quality of generated text.

Given the case discussed above, the machine outcomes cannot be relied on, leading to suboptimal user experiences, regulatory implications, and impacts on firm brands. To conduct human evaluations, we need significant analyst intervention, which may be challenging at times. Given the results, there is a necessity to scale the human evaluation process, which we discuss in the next sections.

Automated Evaluation of Results

We now discuss the results of automated LLM evaluations for machine-generated research. Exhibit 6 shows the LLM-driven evaluations for GPT-4 outcomes (as it was the best model) using the three model flavors with a focus on the metrics described above.

Exhibit 6: Automated Evaluations

Metric Name	GPT-4 Output	Claude-v2	Mistral-7b Output
Answer Relevancy	77.60%	76.00%	77.76%
Groundedness	60.00%	52.00%	50.00%
Conciseness	72.00%	70.00%	56.00%
Concilense Score	69.87%	66.00%	61.25%

Source: Morningstar.

Reviewing the reasoning reveals that automated evaluations detect similar issues in machine-generated outcomes as those detected in human evaluations. While answer relevancy scores are higher, there are issues with groundedness scores, which indicate hallucinations. Additionally, there are issues with lower conciseness scores. The overall consilience score is slightly in line with human evaluations. These

metrics can also be explained, with associated reasoning indicating favorability in a regulatory ecosystem. Aligning scores against human evaluations indicates 80% support with automated evaluations. Hence, these metrics provide a practical and promising choice, which can be extended to calculate derived metrics on the responsible related topics such as misogyny, toxicity, and sentiment of final outcomes through chain-of-thought rubrics. Similarly, regulatory metrics can also be calculated, such as promotion of financial products and adherence to local laws, to name a few.

The above explanation relates to usage of GPT-4 as an LLM evaluator. We also tested other LLMs, such as Claude-v2 and Mistral-7b acting as the LLM evaluator. Across all metrics, GPT-4 demonstrated exceptional alignment with human scores. The superior evaluation performance of GPT-4 can be attributed to its higher parameter counts, higher training data usage, and improved reasoning skill set.

Limitations of Algorithmic Evaluation

While algorithmic evaluation provides several benefits, below are some challenges as well. Consistency of outcome bias was observed when an LLM exhibited differing behavior over multiple runs. This bias is not unique to our problem at hand and has been seen in human scoring systems. Based on our experiments, there is less than a 2 to 3 percent variation across various metrics for multiple runs when LLM evaluators are used. Further, there were latency issues in LLM evaluations, and they cannot be run in real time. Our experiments for automated metrics on sample data ran under 60 minutes. Lastly, the evaluations do not come cheap and can have cost implications for evaluation on top of text generation. As a result, open-source models excel well but quality of outcomes could be compromised.

Talking About Efficiency Gains

Researching the applicability and limitations of LLMs for investment research automation made us also keen to understand the efficiency gains. To this end, we conducted a detailed cost-benefit analysis on integrating LLMs into analyst workflows using real-world data on task allocation from research teams. Results indicate automation enables substantial time savings. To estimate the potential cost savings, we first analyzed research analysts' workflows, providing insight into the time allocation across different tasks. The analysis is based on data collated by our research and customer support groups, who spent significant time understanding client workflows.

We then applied estimated savings percentages based on structured interviews with research analysts spanning the equity, manager, and investment research teams. These savings are then weighted according to the proportion of time spent in each task to arrive at the total estimated savings. The breakdown of analyst tasks and associated cost savings is described below:

Research: Analysts spend approximately 43% of their time on research-related activities. By leveraging LLM analysis capabilities as information mining and data extraction, we estimate a conservative 20% reduction in time spent on research tasks. This translates to a weighted savings of 8.6% across the team of analysts.

Note/Outline: Creating notes and outlines for investment strategies and reports typically accounts for 18% of an analyst's time. LLM ability to generate structured summaries and outlines can lead to a 5% time savings in this area, contributing to a weighted savings of 0.9%.

Writing: Writing reports, investment proposals, and other documentation consumes around 26% of an analyst's time. With the LLM narrative drafting ability, which can generate draft content and suggest language, we expect a substantial 50% reduction in writing time. This leads to a weighted savings of 13%.

Editing: Analysts spend about 13% of their time on editing tasks. Paraphrasing, grammar, and style-checking algorithms can help reduce errors and enhance the efficiency of the editing process by 65%. This results in a weighted savings of 8.45%.

After calculating the weighted savings for each task, we find that LLMs can potentially save a total of 30.95% of the analysts' time. For a large hypothetical investment research team of over 100 analysts with an average salary of $100,000, this would translate to an annual cost savings of close to USD 3 million. These significant savings can be strategically allocated to high-value generation areas, such as exploring new investment opportunities, conducting in-depth research on complex assets, and developing innovative strategies to enhance portfolio performance. To reap these productivity dividends, however, firms must budget for upfront LLM integration costs alongside ongoing maintenance. With responsible implementation guardrails, automation presents a compelling path to sustaining a long-term competitive advantage. The threshold efficiencies uncovered here offer a strategic roadmap for stakeholders who are eyeing a measured adoption of these strategies.

Conclusion

In this paper, we uncover potential challenges of RAG-based systems in generating machine text on real-world, time-consuming investment research tasks from information mining, text condensation, code generation, general research question and answering, numerical reasoning, and drafting narratives along multiple dimensions such as relevancy, groundedness, and conciseness. Based on human evaluations, LLMs can perform simple tasks like information mining and text condensation effectively. However, as the complexity of tasks grows, requiring arithmetic computations and logical-reasoning skills, they start to falter. Insufficient understanding of domain also leads to hallucinations. In addition, there are challenges detected with inconsistent writing styles. Given this, there is a greater need to review machine outcomes in investment research. As human evaluations are costly, we also developed automated metrics driven by LLMs themselves to evaluate other LLM outcomes. Based on our analysis, we see that automated metrics have better alignment when human evaluation scores are in range of 80%. The automated evaluation metrics can also be extended to check if generated text has regulatory and compliance issues. GPT-4 models come out on top for evaluation purposes here.

In conclusion, investors cannot solely rely on machine-generated text for end-user tasks. These tools should be treated as means of bringing in efficiency gains. Investors may start by running controlled experiments on simple use cases such as earnings-call summarizations or financial statement information extractions. Learnings from initial experiments can add in the incremental expansion of the scope of LLMs to more-advanced research workflows. Also, budgeting for the costs of a robust human-in-the-loop model (which requires human interaction) for efficacy testing would be beneficial here. Realizing the full productivity here would require improving LLMs' analytical reasoning prowess through techniques like fine-tuning or perfecting RAG retrievers. The evaluation process discussed above can aid investors in understanding shortcomings of LLMs in applications across daily work and then using these tools judiciously. Tracking emerging innovations in the field of LLMs for investment research should remain on investors' radar for a long-term competitive advantage. Though current limitations exist, rapid incremental enhancements provide a compelling case for firms to actively start their AI readiness journeys today within careful parameters.

References

Zhao, Wayne Xin; Zhou, Kun; Li, Junyi; Tang, Tianyi; Wang, Xiaolei; Hou, Yupeng; Min, Yingqian; Zhang, Beichen; Zhang, Junjie; Dong, Zican; Du, Yifan; Yang, Chen; Chen, Yushuo; Chen, Zhipeng; Jiang, Jinhao; Ren, Ruiyang; Li, Yifan; Tang, Xinyu; Liu, Zikang; Liu, Peiyu; Nie, Jian-Yun; Wen, Ji-Rong. 2023. "A Survey of Large Language Models." Cornell University. https://arxiv.org/pdf/2303.18223.pdf

Suzuki, Masahiro; Sakahi, Hiroki; Hirano, Masanori; Izumi, Kiyoshi. 2023. "Constructing and analyzing domain-specific language model for financial text mining." *ScienceDirect*. Volume 60, Issue 2, https://www.sciencedirect.com/science/article/abs/pii/S0306457322002953

Zheng, Lianmin; Chiang, Wsei-Lin; Sheng, Ying; Zhuang, Siyuan; Wu, Zhanghao; Zhuang, Yonghao; Lin, Zi; Li, Zhuohan; Li, Dacheng; Xing, Eric. P.; Zhang, Hao; Gonzalez, Joseph E.; Stoica, Ion. 2023. "Judging LLM-as-a-Judge With MT-Bench and Chatbot Arena." Cornell University. https://arxiv.org/abs/2306.05685

Manakul, Potsawee; Liusie, Adian; Gales, Mark J. F. 2023 "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models." Cornell University. https://arxiv.org/abs/2303.08896

Imani, Shima; Du, Liang; Shrivastava, Harsh. 2023. "MathPrompter: Mathematical Reasoning Using Large Language Models." Cornell University. https://arxiv.org/abs/2303.05398

Wang, Liang; Yang, Nan; Wei, Furu. 2023. "Query2doc: Query Expansion with Large Language Models" https://www.microsoft.com/en-us/research/publication/query2doc-query-expansion-with-large-language-models/

Chu, Zhixuan; Guo, Huaiyu; Zhou, Xinyuan; Wang, Yijia; Yu, Fei; Chen, Hong; Xu, Wanqing; Lu, Xin; Cui, Qing; Li, Longfei; Zhou, Jun; Li, Sheng. 2023. "Data-Centric Financial Large Language Models." Cornell University. https://arxiv.org/abs/2310.17784

Lewis, Patrick; Perez, Ethan; Piktus, Aleksandra; Petroni, Fabio; Karpukhin, Vladimir; Goyal, Naman; Küttler, Heinrich; Lewis, Mike; Yih, Wen-tau; Rocktäschel, Tim; Riedel, Sebastian; Kiela, Douwe. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Cornell University. https://arxiv.org/abs/2005.11401

Liang, Yaobo; Wu, Chenfei; Song, Ting; Wu, Wenshan; Xia, Yan; Liu, Yu; Ou, Yang; Lu, Shuai; Ji, Lei; Mao, Shaoguang; Wang, Yu; Shou, Linjin; Gong, Ming; Duan, Nan. 2023. "TaskMatrix.AI: Completing

Tasks by Connecting Foundation Models With Millions of APIs." Cornell University. https://arxiv.org/abs/2303.16434

Jiang, Albert Q.; Sablayrolles, Alexandre; Mensch, Arthur; Bamford, Chris; Chaplot, Devendra Singh; de las Casas, Diego; Bressand, Florian; Lengyel, Gianna; Lample, Guillaume; Saulnier, Lucile; Renard Lavaud, Lélio; Lachaux, Marie-Anne; Stock, Pierre; Le Scao, Teven; Lavril, Thibaut; Wang, Thomas; Lacroix, Timothée; El Sayed, William. 2023. "Mistral 7B." Cornell University. https://arxiv.org/abs/2310.06825

Chen, Zhiyu; Chen, Wenhu; Smiley, Charese; Shah, Sameena; Borova, Iana; Langdon, Dylan; Moussa, Reema; Beane, Matt; Huang, Ting-Hao; Routledge, Bryan; Yang Wang, William. 2022. "FinQA: A Dataset of Numerical Reasoning Over Financial Data." Cornell University. https://arxiv.org/abs/2109.00122

Liu, Sengjie; Healey, Christopher G. 2023. "Abstractive Summarization of Large Document Collections Using GPT." Cornell University. https://arxiv.org/abs/2310.05690

About Morningstar Quantitative Research

Morningstar Quantitative Research is dedicated to developing innovative statistical models and data. points, including the Morningstar Quantitative Rating, the Quantitative Equity Ratings, and the Global Risk Model.

For More Information
+1 312 244-7541
lee.davidson@morningstar.com

22 West Washington Street
Chicago, IL 60602 USA