An Evaluation of Reasoning Capabilities of Large Language Models in Financial Sentiment Analysis

Kelvin Du♠, Frank Xing♣, Rui Mao♠ and Erik Cambria♠

♠*School of Computer Science and Engineering, Nanyang Technological University, Singapore*
♣*Department of Information Systems and Analytics, National University of Singapore, Singapore*
zidong001@e.ntu.edu.sg; xing@nus.edu.sg; rui.mao@ntu.edu.sg; cambria@ntu.edu.sg

***Abstract*—Large Language Models (LLMs) have garnered significant attention within the academic community due to their advanced capabilities in natural language understanding and generation. While empirical studies have shed light on LLMs' proficiency in complex task reasoning, a lingering question persists in the field of Financial Sentiment Analysis (FSA): the extent to which LLMs can effectively reason about various financial attributes for FSA. This study employs a prompting framework to investigate this topic, assessing multiple financial attribute reasoning capabilities of LLMs in the context of FSA. By studying relevant literature, we first identified six key financial attributes related to semantic, numerical, temporal, comparative, causal, and risk factors. Our experimental results uncover a deficiency in the financial attribute reasoning capabilities of LLMs for FSA. For example, the examined LLMs such as PaLM-2 and GPT-3.5 display weaknesses in reasoning numerical and comparative attributes within financial texts. On the other hand, explicit prompts related to other financial attributes showcase varied utilities, contributing to LLMs' proficiency in discerning financial sentiment.**

***Index Terms*—financial sentiment analysis, large language models, prompt engineering**

I. Introduction

The substantial research focus on Large Language Models (LLMs) stems from their robust capabilities in natural language understanding and generation. The human-like language proficiency of LLMs motivates researchers to investigate their reasoning capabilities across various tasks, aiming to unveil decision-making skills that extend beyond linguistic abilities. The significance of studying the reasoning capabilities of LLMs lies in unraveling the depths of their cognitive processes [1] and intention awareness [2]. This exploration not only enhances our comprehension of artificial intelligence but also guides the refinement of future models for more sophisticated problem-solving capabilities. Previous works studied the reasoning capabilities of LLMs by conducting evaluation on different tasks, e.g., logical reasoning [3], commonsense reasoning [3], causal reasoning [4] and a wide range of scientific knowledge [5]. However, these empirical studies focused on the reasoning capabilities of LLMs in the general domain or the specific domains that are irrelevant to Financial Sentiment Analysis (FSA). The significance of FSA resides in its practical applications across various domains, including but not limited to investment decision-making, financial forecasting, risk management, corporate strategy, and regulatory compliance [6]–[8].

FSA exhibits notable distinctions from general sentiment analysis across several dimensions. Firstly, it frequently encounters the use of metaphorical expressions within financial communications, wherein figurative language is deployed to convey emotions or portray market conditions [9], [10]. For example, a ubiquitous metaphor such as "The market is riding a bull" serves as a symbolic representation of a robust and upward market trend, introducing a layer of intricacy to the sentiment analysis of financial texts. Besides, FSA frequently hinges on the direction of events or changes, underscoring the importance of contextual consideration [11]. For instance, the term "profit" can carry both positive and negative sentiments contingent upon the direction. An upsurge in profit typically conveys positivity, while a decline is generally regarded as negative. Lastly, unlike conventional sentiment analysis, which predominantly deals with textual content, financial texts often amalgamate qualitative discourse with quantitative data [12]. For example, given "in the four weeks that followed its release, the standard iPhone 15 sold 130.6 percent more than the standard iPhone 14 did in the same time period last year"[1], the sentiment inference for the sell of iPhone 15 is upon the understanding of the "130.6 percent" increases and the comparative relationship between iPhone 15 and iPhone 14. This requires FSA to not only interpret the language employed within financial documents but also process and evaluate numerical information in conjunction with its surrounding textual context, thereby fostering a holistic understanding of sentiment. The distinctive features of FSA underscore the need for specialized reasoning skills in LLMs. This requirement arises from the nuanced language and specific knowledge demands inherent in FSA, setting it apart from sentiment analysis in a broader context. Furthermore, although several empirical studies in have tested LLMs in general affective computing tasks [13], [14], they did not test the reasoning capabilities of LLMs regarding different financial attributes. Conducting sentiment analysis on financial texts using LLMs necessitates an examination that extends beyond the mere accuracy of sentiment polarity detection. It is crucial to investigate whether LLMs possess the capability to comprehend financial attributes during the process of reasoning sentiment polarities.

[1]https://koreajoongangdaily.joins.com/news/2023-11-21/business/industry/iPhone-15-sells-record-units-doubling-its-predecessor/1918158

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00042

Thus, we aim to answer the following research questions (RQs) in this work:

1) What distinctive attributes within financial texts have the potential to convey sentiment, whether explicitly or implicitly?
2) To what extent can LLMs comprehend the financial attributes when deducing financial sentiment?

We conducted a comprehensive literature review in FSA, identifying six key financial attributes believed to convey sentiment for RQ-1. We examined the effectiveness of prompts associated with these attributes, aiming to answer RQ-2. The subsequent exploration of related literature and experimental outcomes yielded the following key findings: **a)** Considerable research efforts have been dedicated to FSA, particularly through the modeling and analysis of essential financial attributes. These attributes encompass semantic, numerical, temporal, comparative, causal, and risk factors, all of which have been identified as pivotal elements for FSA. **b)** We observe that certain LLMs exhibit a lack of financial attribute reasoning capabilities to perform FSA. Among the identified six financial attributes, the examined LLMs, e.g., PaLM-2 [15] and GPT-3.5 [16] are particularly weak in reasoning numerical and comparative attributes in financial text. The explicit prompts on other financial attributes can bring different utilities to LLMs in deducing financial sentiment. Our contributions can be summarized as twofold: Firstly, we identified six specific financial attributes with the potential to influence financial sentiment. Secondly, we devised a prompting framework to assess the reasoning capabilities of LLMs concerning these six financial attributes.

II. Related Works

A. Reasoning Capabilities of LLMs

The work of [5] summarized several reasoning tasks that were tested with LLMs. [3] developed 60 questions to test LLMs in inductive, deductive, and abductive reasoning, demonstrating satisfying logical reasoning skills of LLMs in a general domain. [3] tested ChatGPT using three commonsense datasets, revealing that 80 out of 90 predictions made by ChatGPT were accurate. The model demonstrated the ability to provide comprehensive explanations for its reasoning process. However, the studies performed by [17], [18] indicated that ChatGPT's performance in commonsense reasoning lagged behind fine-tuned baseline models. [19] highlighted instances of commonsense and specifically physical reasoning failures in ChatGPT, along with pointing out other identified shortcomings. The evaluation conducted by [4] systematically assessed event causality identification, causal discovery, and causal explanation generation. In comparison to state-of-the-art models, both ChatGPT and GPT-4 exhibited lower scores in causality identification. [20]–[24] tested LLMs' reasoning in science domains, e.g., mathematics, computer science, physics, chemistry, and medicine demonstrating the knowledge that has been grasped by LLMs. However, the aforementioned evaluations of LLMs' reasoning capabilities were not studied in the financial domain, which depends on domain-specific reasoning skills. There are other task-specific evaluations of LLMs for affective computing [13], [14], whereas those evaluations focused on assessing the accuracy of LLMs under traditional task setups, neglecting an examination of LLMs' understanding of financial attributes.

B. Financial Attributes for FSA

Concurrently, existing literature underscores that the multifaceted nature of financial texts makes them challenging in FSA tasks. Financial texts are characterized by distinct attributes that span a variety of dimensions. **Semantically**, financial texts are rich in specialized vocabulary, reflecting the nuances of economics and finance [9], [25]. **Numerically**, numbers are a fundamental aspect of financial texts, as they are replete with figures, percentages, and quantitative details that provide essential financial metrics, which are pivotal in financial analyses and decision-making processes within the domain of finance [12], [26], [27]. **Temporally**, financial texts frequently denote specific time frames, such as quarterly or annual periods, thereby underscoring the significance of temporal trends and shifts [28]–[30]. In **a comparative context**, these texts routinely juxtapose current financial performance against previous periods or competitor benchmarks, thus offering a relative evaluation of performance, as noted by [31]–[33]. **Causally**, financial texts establish connections between financial outcomes and specific business decisions or market events, explaining the reasons behind financial results [34]–[36]. Lastly, addressing **risk and uncertainty**, financial texts often discuss prospective uncertainties and challenges that could affect financial stability, thereby underscoring the critical role of risk management in financial planning and decision-making [37]–[39]. While earlier research primarily focused on exploring the adaptability of LLMs in FSA task [13], [14], [40], [41], we take a step forward by assessing the financial attribute reasoning capabilities of LLMs, focusing on the aforementioned financial attributes that have the potential to convey sentiment.

III. Proposed Approach

Based on the review in Section II-B, we observe that the following financial attributes have demonstrated great significance in previous theoretical and empirical research in FSA, namely semantic, numerical, temporal, comparative, causal, and risk attributes. Thus, the following examination focuses on analyzing the reasoning capabilities of LLMs from these perspectives. The reasoning capabilities of LLMs are evaluated based on the performance variant upon explicitly prompting LLMs with the identified financial attributes. The hypothesis is that if an LLM has a strong capability to infer the impact of a financial attribute in FSA, the explicit prompting of the financial attribute to the LLM should yield weak utility, e.g., lower accuracy gains in FSA. Conversely, if a prompt related to the financial attribute leads to substantial accuracy gains, it signifies a limited awareness of the connotations associated with that specific financial attribute. We develop a Financial Attribute Prompting (FAP) framework that ensemble the six financial attributes to evaluate the FSA

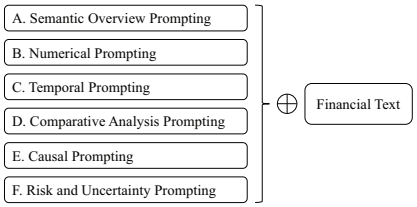

Fig. 1. The proposed financial attribute prompting framework.

performance of LLMs. The framework is shown in Figure 1, which provides explicit step-by-step instructions to LLMs on how to conduct sentiment analysis of financial text upon the financial attributes.

A. Semantic Overview Prompting

The semantic overview serves to assess general sentiment indicators and the overall tone of the financial text, which encompasses semantics understanding and overarching sentiment indicators. The prompt is designed as:

```
Semantic Overview: Start by assessing general
sentiment indicators and the overall tone is
positive, neutral or negative.
```

This step, inspired by cognition, mirrors the human cognitive process, which involves contextual comprehension before task execution. Semantic understanding seeks to decipher the underlying meaning of word associations and develop a general understanding of the context. We guide an LLM with overarching sentiment indicators to establish connections between contextual meanings and the specific task at hand.

B. Numerical Prompting

Financial statements, forecasts, and financial reports are inherently quantitative. The quantitative nature of financial texts brings an abundance of numbers, ratios, percentages, and quantitative measures that can convey sentiment, especially concerning expectations. For instance, a "10% growth" might be positive, but if the expected growth was 15%, the sentiment could be negative. Thus, we develop a prompt to explicitly remind LLMs to notice the numbers or metrics by:

```
Numerical Context: Next, focus on any specific
numbers or metrics mentioned and their sentiment
implications. Figures without context could be
considered neutral.
```

The numerical prompting aims to analyze numbers, percentages, ratios, and other quantitative metrics to gauge sentiment based on performance, achievements, or projections. It involves scrutinizing specific numerical values or metrics mentioned in the text to ascertain their sentiment.

C. Temporal Prompting

The time value of money is a fundamental concept in finance. Assessing performance requires looking at past, present, and future metrics and frequent time expressions like "YoY" (Year-over-Year), "QoQ" (Quarter-over-Quarter), "forecasted", "by the end of the fiscal year", etc. Future outlook and past performance expressions can be sentiment-laden. "Expect a stronger next quarter" or "had a challenging past year" are examples. Thus, the prompt is developed as

```
Temporal Context: Delve into any timeframes or
specific contexts that might influence the sentiment.
If there is recent and past sentiment, the final
sentiment should be based on the most recent
sentiment.
```

The temporal context prompt focuses on sentiment cues tied to specific timeframes, such as past performance, present conditions, or future expectations. It involves investigating any timeframes or specific contextual information that may influence the sentiment expressed in the text.

D. Comparative Analysis Prompting

To assess business performance, companies often compare their results with competitors, past performance, or industry benchmarks, which is a practice that stems from both competitive analysis and fundamental financial analysis techniques. For instance, a decrease in profitability typically signifies negative sentiment, while growth in sales, an increase in share price, or a reduction in losses are indicative of positive sentiment. Comparative phrases and benchmarks, such as "better than competitors", "outperformed the industry average", "highest since", etc. are commonly used in financial texts. Comparative analysis with competitors, past performance, or industry benchmarks can also carry the sentiment. Phrases like "outperformed competitors" or "lagged behind industry benchmarks" are indicative. Thus, the comparative analysis prompt is formed as

```
Comparative Analysis: If the statement compares
performance with another entity or timeframe,
derive sentiment from this relative performance.
For example, a decrease in profitability represents
negative sentiment, a growth in sales, increase in
share price or reduction in loss is positive. When
there is mixed sentiment, you should follow the rule
that improvement stands for positive in finance.
```

The comparative analysis prompt is to evaluate the financial text in relation to benchmarks, past performance, or competitors to determine relative sentiment. It entails evaluating the performance relative to another entity or timeframe and deriving sentiment based on this relative performance.

E. Causal Prompting

The framework also includes the consideration of causality. In financial contexts, major events, milestones and strategic moves can impact investor sentiment. For instance, if a leading tech company announces a merger with a smaller innovative firm, investors might view it positively, anticipating market expansion and technological synergies, leading to a stock price surge. Thus, the prompt is defined as

```
Causal Attribution: Identify and assess any causal
factors or strategic moves mentioned that carry
sentiment.
```

The causal prompt is to evaluate the financial text in pinpointing specific factors or events that contributed to observed outcomes, providing context to the sentiment. It is essential to weigh the impact of these factors on the overall sentiment expressed.

F. Risk and Uncertainty Prompting

Risks and uncertainty are also associated with sentiment reasoning. Various risk and uncertainties can sway financial sentiment. Political instability or international conflicts can introduce uncertainty, affecting global markets and investment climates. The threat of an economic downturn can lead to cautious investment strategies, impacting market sentiment. Thus, we prompt risks for LLMs via

```
Risk and Uncertainty Analysis: Evaluate potential
risks, threats, or uncertainties that carry
sentiment.
```

The focus of the risk and uncertainty prompt is to highlight and evaluate potential risks, threats, or uncertainties mentioned in the financial text that may have sentiment implications.

G. Prefix and Suffix Prompts

Empirical studies, e.g., Chain-of-Thoughts [42] has demonstrated that prompting LLMs with "thinking step-by-step" can enhance performance on downstream tasks. Thus, we add a widely-adopted prefix before the financial attribute prompts by

```
Analyze the sentiment of the provided financial text
through a structured approach below:
```

Besides, to guarantee that LLMs can produce the intended sentiment labels, e.g., "positive", "neutral", or "negative", we append a suffix following the financial attribute prompts

```
Please proceed through these steps, assess
holistically and provide a final sentiment
classification as either 'Positive', 'Neutral', or
'Negative'.
```

IV. Experimental Setup

A. Datasets

We conduct sentiment analysis experiments on two well-received datasets i.e., PhraseBank [43] and Twitter Financial News dataset [44]. The PhraseBank dataset consists of 4,846 pieces of news that have been categorized into positive, neutral, and negative sentiments by 16 individuals possessing expertise in financial markets from an investor perspective. The dataset includes four reference datasets, each based on the level of agreement among annotators, namely 100%, 75%, 66%, and 50% agreement. In this study, the 100% and 50% agreement datasets are adopted as the benchmark. The Twitter Financial News dataset consists of an annotated collection of tweets in English, focusing on financial topics. The dataset holds 11,932 finance-oriented tweets which are categorized into bearish, bullish, and neutral sentiment.

B. Baseline Models

We compare the performance of LLMs against several baselines in FSA. Lexicon-based methods include LM [45], SMSL [46] and FinSenticNet [47], which leveraged financial knowledge bases to predict sentiment polarities in unsupervised approach. We also include several supervised-learning-based methods in our baselines, namely Linearized Phrase-Structure (LPS) model [43], Hierarchical Sentiment Classifier (HSC) [48], FinSSLx [49], ULMFit [50] and FinBERT [51], [52]. Since there are two versions of FinBERT, we denote the earlier version [51] as FinBERT^a and the later version [52] as FinBERT^b. We test the reasoning capability of LLMs upon GPT-3.5 from Open AI and PaLM-2 from Google. We also report the improvements of our prompting method upon the state-of-the-art LLM, GPT-4.

V. Result and Analysis

A. Effectiveness of the Financial Attribute Prompting

We use accuracy and macro-averaged F1-Score as the primary metrics for FSA, with the results presented in Table I. A general observation suggests that the incorporation of FAP enhances the performance of all the scrutinized LLMs, particularly PaLM-2 and GPT-3.5. This implies that the reasoning capabilities of LLMs regarding financial attributes are inadequate, and these models lack complete awareness of financial attributes without explicit prompting. Furthermore, the performance disparity between LLMs with and without the ensemble of prompts underscores the absence of a developed structural thinking framework of LLMs, comparable to human cognitive processes in the FSA domain. Zero-shot-based LLMs with FAP exceed all unsupervised lexicon-based methods, while the vanilla PaLM-2 and GPT-3.5 fall behind the state-of-the-art lexicon-based method, FinSenticNet [47]. GPT-4 with FAP also exceeds the latest FinBERT^b [52]. Such an observation demonstrates the effectiveness of the proposed FAP framework and our defined financial attributes (RQ1).

B. Financial Attributes Reasoning Capabilities of LLMs

Next, we conduct an ablation study with PaLM-2 and GPT-3.5 to assess the utility of different financial attribute prompts and to answer RQ2: To what extent can LLMs comprehend the financial attributes when deducing financial sentiment? Relevant results can be viewed in Table II. **1. Overall semantic understanding:** It is interesting to notice that PaLM-2 (FAP w/o Overview) exceeds the full model PaLM-2 (FAP), showing that the explicit semantic overview prompting does not bring accuracy gains for PaLM-2. This implies that PaLM-2 has had the intention of grasping the overall semantics of the input. An extra prompt may introduce noise for PaLM-2 inferring FSA. In contrast, without the explicit semantic overview prompting, GPT-3.5 achieves lower accuracy, indicating the necessity of the step that builds a general understanding of the context. **2. Numerical**

TABLE I
COMPARISON WITH BASELINE METHODS ON FSA BENCHMARK DATASETS. BOLDFACE INDICATED THE TOP 3 RESULT. "-" MEANS NOT REPORTED.

Method	Model	PhraseBank-100%		PhraseBank-50%		Twitter Fin. News		Average	
		Accuracy	*Macro-F1*	*Accuracy*	*Macro-F1*	*Accuracy*	*Macro-F1*	*Accuracy*	*Macro-F1*
Unsupervised lexicon-based methods	SMSL [46]	0.2800	0.2890	0.3082	0.2927	0.3027	0.3016	0.2969	0.2944
	LM [45]	0.6444	0.3688	0.6244	0.5020	0.5971	0.4604	0.6219	0.4437
	FinSenticNet [47]	0.7619	0.7216	0.6624	0.6215	0.6000	0.5269	0.6747	0.6233
Supervised learning-based methods	LPS [43]	0.7900	0.8000	0.7100	0.7100	-	-	-	-
	HSC [48]	0.8300	0.8600	0.7100	0.7600	-	-	-	-
	FinSSLx [49]	0.9090	0.8770	-	-	-	-	-	-
	ULMFit [51]	0.9300	0.9100	0.8300	0.7900	-	-	-	-
	FinBERT^a [51]	**0.9700**	**0.9500**	**0.8600**	**0.8400**	-	-	-	-
	FinBERT^b [52]	0.9169	0.8970	0.7926	0.7514	**0.7483**	0.6612	**0.8192**	0.7698
Zero-shot LLM-based methods	PaLM-2	0.5631	0.6245	0.5006	0.5446	0.4367	0.4589	0.5001	0.5426
	PaLM-2 (w/ FAP)	0.8361	0.8511	0.6964	0.7274	0.5640	0.5780	0.6988	0.7188
	GPT-3.5	0.7906	0.8140	0.6597	0.6989	0.5967	0.6077	0.6823	0.7068
	GPT-3.5 (w/ FAP)	0.9187	0.9174	0.7783	0.7718	**0.7324**	**0.7057**	0.8098	**0.7983**
	GPT-4	**0.9602**	**0.9557**	**0.8383**	**0.8290**	**0.7510**	**0.7362**	**0.8498**	**0.8403**
	GPT-4 (w/ FAP)	**0.9639**	**0.9593**	**0.8232**	**0.8234**	0.7280	**0.7150**	**0.8383**	**0.8325**

TABLE II
ABLATION STUDY ON FSA BENCHMARK DATASETS

Model	PhraseBank-100%		PhraseBank-50%		Twitter Fin. News		Average	
	Accuracy	*Macro-F1*	*Accuracy*	*Macro-F1*	*Accuracy*	*Macro-F1*	*Accuracy*	*Macro-F1*
PaLM-2	0.5631	0.6245	0.5006	0.5446	0.4367	0.4589	0.5001	0.5427
+ FAP	0.8361	0.8511	0.6964	0.7274	0.5640	0.5780	0.6988	0.7188
+ FAP w/o Overview	0.8458	0.8583	0.7067	0.7362	0.5812	0.5934	0.7112	0.7293
+ FAP w/o Numerical	0.7539	0.7852	0.6297	0.6719	0.5067	0.5271	0.6301	0.6614
+ FAP w/o Temporal	0.8255	0.8428	0.6884	0.7217	0.5699	0.5827	0.6946	0.7157
+ FAP w/o Comparative	0.7751	0.8019	0.6423	0.6822	0.5016	0.5224	0.6396	0.6688
+ FAP w/o Causal	0.8268	0.8432	0.6869	0.7194	0.5615	0.5758	0.6917	0.7128
+ FAP w/o Risk and Uncertainty	0.8383	0.8519	0.6976	0.7287	0.5745	0.5876	0.7034	0.7227
GPT-3.5	0.7906	0.8140	0.6597	0.6989	0.5967	0.6077	0.6823	0.7068
+ FAP	0.9187	0.9174	0.7783	0.7718	0.7324	0.7057	0.8098	0.7983
+ FAP w/o Overview	0.8873	0.8913	0.7505	0.7452	0.7600	0.7249	0.7992	0.7871
+ FAP w/o Numerical	0.8922	0.8957	0.7484	0.7482	0.6754	0.6651	0.7720	0.7696
+ FAP w/o Temporal	0.9218	0.9214	0.7907	0.7912	0.7257	0.7023	0.8127	0.8049
+ FAP w/o Comparative	0.8745	0.8787	0.7577	0.7652	0.6637	0.6591	0.7653	0.7676
+ FAP w/o Causal	0.9416	0.9375	0.7989	0.7827	0.7378	0.7092	0.8261	0.8098
+ FAP w/o Risk and Uncertainty	0.9160	0.9174	0.7865	0.7904	0.7156	0.6981	0.8060	0.8019

attribute reasoning: Removing the numerical prompting results in sharp losses in PaLM-2 and GPT-3.5. It indicates the weaknesses of the two LLMs in reasoning quantitative information in the financial text. **3. Temporal attribute reasoning:** Excluding the temporal prompt results in the accuracy decreases in PaLM-2 in the PhraseBank with 100% and 50% agreements, while the accuracy loss can be obsessed in evaluating GPT-3.5 with the Twitter Financial News. For other datasets, the absence of the temporal prompt can improve the accuracy of the LLMs. The inconsistency may arise from the subjective nature of human sentiment annotation and variations in annotation criteria across different datasets. The indeterminate outcomes do not conclusively establish the temporal attribute reasoning capabilities of LLMs. Although there might be some degree of temporal knowledge acquisition by the models, the ambiguity in the results precludes a definitive assertion. **4. Comparative attribute reasoning:** The absence of the comparative analysis prompt leads to a discernible decline in accuracy for both PaLM-2 and GPT-3.5 across all datasets. This underscores the inherent limitation of these LLMs in comprehending comparative relationships among distinct entities or temporal frames. **5. Causal attribute reasoning:** While PaLM-2 benefits from the causal prompt in terms of accuracy gains, GPT-3.5 exhibits higher accuracy even in the absence of the causal prompt. This suggests that GPT-3.5 has a strong capability for causal reasoning, while PaLM-2 is comparatively weaker in this aspect. **6. Risk and uncertainty attribute reasoning:** The influence of the risk and uncertainty prompt is modest, as its effects vary for PaLM-2 and GPT-3.5 across different datasets, producing both positive and negative impacts. This suggests that the reasoning capabilities of these LLMs regarding risk and uncertainty are present to some extent.

VI. CONCLUSION

This study has assessed the reasoning capabilities of LLMs in performing FSA. By studying relevant literature, we define six critical financial attributes to test the reasoning capabilities, related to semantic, numerical, temporal, comparative, causal, and risk attributes. Our experiment results demonstrate the effectiveness of our developed FAP framework covering the six financial attributes. We further observe that LLMs are particularly weak in numerical and comparative reasoning. The evaluation of reasoning capabilities for other financial attributes yields varied conclusions, as the prompts can have both positive and negative impacts on the performance of LLMs across different datasets.

REFERENCES

[1] R. Mao, K. Du, Y. Ma, L. Zhu, and E. Cambria, "Discovering the cognition behind language: Financial metaphor analysis with MetaPro," in *IEEE ICDM*, 2023, pp. 1211–1216.

[2] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[3] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung *et al.*, "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," *arXiv:2302.04023*, 2023.

[4] J. Gao, X. Ding, B. Qin, and T. Liu, "Is ChatGPT a good causal reasoner? a comprehensive evaluation," *arXiv:2305.07375*, 2023.

[5] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A survey on assessments of ChatGPT and GPT-4," in *Proceedings of LREC-COLING*, 2024.

[6] W. J. Yeo, W. van der Heever, R. Mao, E. Cambria, R. Satapathy, and G. Mengaldo, "A comprehensive review on financial explainable AI," *arXiv:2309.11960*, 2023.

[7] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, 2024.

[8] K. Du, R. Mao, F. Xing, and E. Cambria, "A dynamic dual-graph neural network for stock price movement prediction," in *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, 2024.

[9] R. Mao, X. Li, M. Ge, and E. Cambria, "MetaPro: A computational metaphor processing model for text pre-processing," *Information Fusion*, vol. 86-87, pp. 30–43, 2022.

[10] R. Mao, X. Li, K. He, M. Ge, and E. Cambria, "MetaPro Online: A computational metaphor processing online system," in *ACL*, 2023, pp. 127–135.

[11] J. Park, H. J. Lee, and S. Cho, "Automatic construction of context-aware sentiment lexicon in the financial domain using direction-dependent words," *arXiv:2106.05723*, 2021.

[12] F. Siano and P. D. Wysocki, "The primacy of numbers in financial and accounting disclosures: Implications for textual analysis research," *Available at SSRN 3223757*, 2018.

[13] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, "A wide evaluation of ChatGPT on affective computing tasks," *arXiv:2308.13911*, 2023.

[14] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," *arXiv:2305.15005*, 2023.

[15] Google, "PaLM 2," 2023, https://ai.google/discover/palm2/ [Accessed: (07-Jan-2024)].

[16] OpenAI, "GPT-3.5," 2023, https://platform.openai.com/docs/models [Accessed: (07-Jan-2024)].

[17] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a general-purpose natural language processing task solver?" *arXiv:2302.06476*, 2023.

[18] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. X. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," *arXiv:2305.18486*, 2023.

[19] E. Davis, "Benchmarks for automated commonsense reasoning: A survey," *arXiv:2302.04752*, 2023.

[20] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner, "Mathematical capabilities of ChatGPT," *arXiv:2301.13867*, 2023.

[21] S. Bordt and U. von Luxburg, "ChatGPT participates in a computer science exam," *arXiv:2303.09461*, 2023.

[22] G. Kortemeyer, "Could an artificial-intelligence agent pass an introductory physics course?" *Physical Review Physics Education Research*, vol. 19, no. 1, p. 010132, 2023.

[23] T. M. Clark, "Investigating the use of an artificial intelligence chatbot with general chemistry exam questions," *Journal of Chemical Education*, 2023.

[24] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash *et al.*, "How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, no. 1, p. e45312, 2023.

[25] F. Xing, L. Malandri, Y. Zhang, and E. Cambria, "Financial sentiment analysis: an investigation into common mistakes and silver bullets," in *COLING*, 2020, pp. 978–987.

[26] Y. Ma, R. Mao, Q. Lin, P. Wu, and E. Cambria, "Multi-source aggregated classification for stock price movement prediction," *Information Fusion*, vol. 91, pp. 515–528, 2023.

[27] ——, "Quantitative stock portfolio optimization by multi-task learning risk and return," *Information Fusion*, vol. 104, p. 102165, 2024.

[28] M. Baxter and R. G. King, "Measuring business cycles: approximate band-pass filters for economic time series," *Review of Economics and Statistics*, vol. 81, no. 4, pp. 575–593, 1999.

[29] T. Loughran and B. McDonald, "Measuring readability in financial disclosures," *the Journal of Finance*, vol. 69, pp. 1643–1671, 2014.

[30] Y. Guo, C. Hu, and Y. Yang, "Predict the future from the past? on the temporal data distribution shift in financial sentiment classifications," *arXiv:2310.12620*, 2023.

[31] I. Maignan, "Consumers' perceptions of corporate social responsibilities: A cross-cultural comparison," *Journal of Business Ethics*, vol. 30, pp. 57–72, 2001.

[32] M. M. Parast and S. G. Adams, "Corporate social responsibility, benchmarking, and organizational performance in the petroleum industry: A quality management perspective," *International Journal of Production Economics*, vol. 139, no. 2, pp. 447–458, 2012.

[33] K. G. Palepu, P. M. Healy, S. Wright, M. Bradbury, and J. Coulton, *Business analysis and valuation: Using financial statements*. Cengage AU, 2020.

[34] M. Powell and D. Ansic, "Gender differences in risk behaviour in financial decision-making: An experimental analysis," *Journal of Economic Psychology*, vol. 18, no. 6, pp. 605–628, 1997.

[35] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *IJCAI*, 2015.

[36] Q. Li, J. Tan, J. Wang, and H. Chen, "A multimodal event-driven lstm model for stock prediction using online news," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, pp. 3323–3337, 2020.

[37] P. MacKay and S. B. Moeller, "The value of corporate risk management," *The Journal of Finance*, vol. 62, no. 3, pp. 1379–1419, 2007.

[38] L. Rigotti and C. Shannon, "Uncertainty and risk in financial markets," *Econometrica*, vol. 73, no. 1, pp. 203–243, 2005.

[39] P. Bromiley, "Testing a causal model of corporate risk taking and performance," *Academy of Management Journal*, vol. 34, p. 37, 1991.

[40] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia, "Is ChatGPT a good sentiment analyzer? A preliminary study," *arXiv:2304.04339*, 2023.

[41] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[43] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.

[44] HuggingFace, "Twitter financial news sentiment," 2023, https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment [Accessed: (07-Jan-2024)].

[45] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[46] H. Tatsat, S. Puri, and B. Lookabaugh, *Machine Learning and Data Science Blueprints for Finance*. O'Reilly Media, 2020.

[47] K. Du, F. Xing, R. Mao, and E. Cambria, "FinSenticNet: A concept-level lexicon for financial sentiment analysis," in *IEEE SSCI*, 2023, pp. 109–114.

[48] S. Krishnamoorthy, "Sentiment analysis of financial news articles using performance indicators," *Knowledge and Information Systems*, vol. 56, no. 2, pp. 373–394, 2018.

[49] M. Maia, A. Freitas, and S. Handschuh, "FinSSLx: A sentiment analysis model for the financial domain using text simplification," in *IEEE ICSC*, 2018, pp. 318–319.

[50] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *ACL*, 2018, pp. 328–339.

[51] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," *arXiv:1908.10063*, 2019.

[52] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, 2023.