RACCOON: Real-world Advanced financCial analysis through COmprehensive Natural language dataset

Seonghyun Kim[1], Kanghee Lee[2], Minsu Jeong[3], Junghan Yoon[3]

[1]Korea University, [2]Konkuk University, [3]Sungkyunkwan University,
[1]qksksk657@korea.ac.kr, [2]khlee91@konkuk.ac.kr, [3]pisces03@skku.edu, [4]yjh4037@g.skku.edu

Abstract

Our research introduces Raccoon [1], a benchmark dataset aimed at evaluating the cognitive capabilities of large language models (LLMs) in the complex domain of financial analysis. Traditional NLP benchmarks primarily focus on assessing the correctness of model outputs without examining the underlying cognitive processes. In contrast, Raccoon shows the simulation of human-like reasoning by integrating planning and reasoning tasks that mimic complicated human thought processes. Our study analyzes the extent to which LLMs understand the implicit meanings behind questions within the financial domain, and how these meanings are interpreted from various perspectives. To identify the differences, we compared the planning and reasoning processes of LLMs with those of human analysts. Our findings suggest that LLMs adopt more detailed approaches to problem-solving, which can sometimes limit their ability to effectively reach conclusions through reasoning. This comprehensive evaluation not only enhances our understanding of the cognitive limitations of current LLM architectures but also informs future development directions aimed at bridging the gap between artificial and human cognitive abilities in financial analysis.

1 Introduction

The emergence of LLMs in the field of computational linguistics has made considerable progress in natural language processing (NLP) tasks (Brown, Mann et al. 2020; Rosoł, Gąsior et al. 2023). These models have not only demonstrated capabilities at or near human expert levels in specialized domains such as legal (Cui, Li et al. 2023) and clinical (Kwon, Ong et al. 2024). Despite these advances, a major gap remains in our understanding of how LLMs simulate human-like thought processes and reach conclusions (Huang, Chen et al. 2023). This gap is highlighted by evaluation methodologies and datasets that focus primarily on the model's ability to identify 'correct' answers, rather than clarifying the underlying cognitive processes involved (Yang, Qi et al. 2018; Liang, Bommasani et al. 2022). In contrast, human problem-solving requires clear and logical progression: understanding the problem, preparing necessary knowledge, and systematically connecting this knowledge to derive solutions (Phogat, Harsha et al. 2023; Song, Xiong et al. 2023). This core process also necessitates what is called a step-by-step agent-based approach (Wang, Wei et al. 2022; Zhang, Zhang et al. 2022; Sun, Zheng et al. 2023).

Our study aims to introduce a new benchmark dataset within the financial domain—an area characterized by complex thought processes encompassing political, economic, historical, and sociological considerations. This dataset is not only a challenge for LLMs to demonstrate human-like reasoning but also serves as a robust benchmark for assessing cognitive abilities in one of the most sophisticated contexts of human thought.

To address the complexities inherent in evaluating the cognitive capabilities of LLMs in the financial domain, our study concentrates on two pivotal aspects: planning and reasoning. These cognitive functions are essential for mimicking the complicated human thought process, which is

* All authors contributed equally to this research.

[1] The raccoon dataset is available for download at https://github.com/MrBananaHuman/finance_raccoon

Proceedings of the Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen), Jeju, South Korea, August 3, 2024

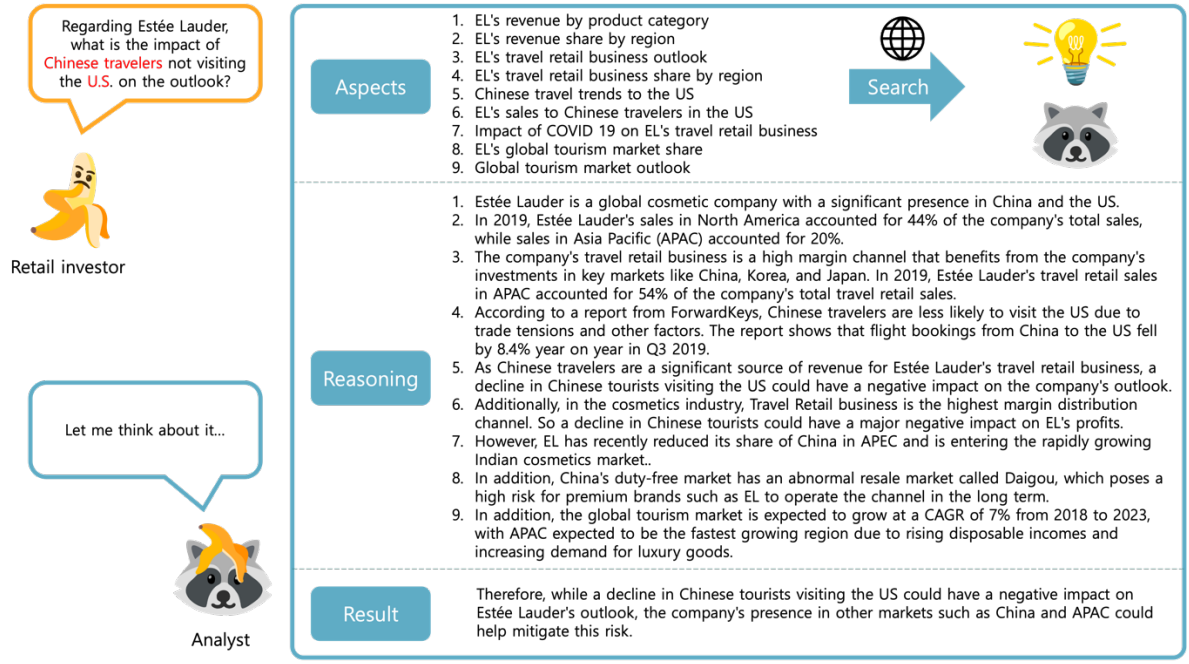

Figure 1: Overview of Raccoon dataset.

necessary for addressing complex and multifaceted issues.

Planning involves the ability to predict and clearly describe the steps necessary for problem-solving before beginning the process. This requires a strategic overview and sequence of actions aligned with achieving the intended outcomes. In the context of financial domain, this may include identifying relevant economic indicators, potential regulatory impacts, or market trends to consider.

On the other hand, reasoning is the process of connecting the dots between the information gathered during the planning stage and the specific question at hand. This includes deriving logical conclusions from a series of premises or known facts. This stage is crucial for navigating the complexities of financial data and interpreting it to make informed decisions or predictions.

Our study merges these two cognitive processes into a single framework for challenging LLMs with finance-related questions. (1) Aspect: Upon receiving a finance-related question, the model defines the aspects necessary to resolve the question. This involves sequentially outlining the key elements or considerations related to the query. (2) Reasoning: Once these aspects are planned and the necessary information is provided, the LLM generates a series of reasoning steps to construct a consistent and logical response.

We applied this two-step task to both human experts, LLMs and compared scenarios that did not include a planning stage. This comparative analysis not only highlights the importance of planning in complex problem-solving but also allows for evaluating the depth of understanding and cognitive similarities between humans and LLMs in handling complicated financial issues.

2 Related works

Inducing LLMs to generate reasoning steps can significantly assist them in identifying correct solutions in complex problems. Wei, Wang et al. 2022 implemented Chain of Thought (CoT) prompting in LLMs to trace reasoning pathways during problem-solving, improving both interpretability and accuracy (Chen, Ma et al. 2022; Mavi, Saparov et al. 2023; Lu, Peng et al. 2024). Despite these results, benchmarks often prioritize outcome correctness over reasoning process.

In financial analysis, the role of AI has traditionally been confined to predictive modeling. Jin, Tang et al. 2024 demonstrated the use of LLMs in forecasting stock market trends from historical data, yet their exploration into the reasoning processes of models was limited. This gap is starting to close with recent contributions like those from Chen, Chen et al. 2021; Son, Jung et al. 2023

who introduced datasets demanding semantic understanding and logical reasoning in economic contexts, though not fully replicating human-like cognitive processes.

Our research extends these efforts by offering a structured framework that assesses the capability of LLMs to perform step-by-step reasoning analogous with planning to that of a human financial analyst. This approach aligns with Chang, Wang et al. 2024, who discuss the operational paradigms of LLMs that emulate human cognitive processes, highlighting the need for frameworks that assess ethical fidelity alongside cognitive capabilities. Similarly, Momennejad, Hasanbeig et al. 2024 emphasize the necessity for integrating complex relational structures and functionalities like the human cognition to enhance the performance of LLMs in real-world tasks.

Furthermore, Li, Xu et al. 2024 underscore the growing role of LLMs in natural language generation (NLG) evaluation, focusing on their adaptability to produce coherent and contextually relevant assessments. Their work, employing techniques like Reinforcement Learning with Human Feedback (RLHF), strives to refine the generative capabilities of LLMs to enhance human-like reasoning processes, particularly in domain-specific tasks such as medical and financial text analysis.

Here, we address these issues by constructing a benchmark dataset for the financial domain that includes planning and reasoning tasks, with the goal of studying the differences in cognitive thinking between LLMs and humans.

3 Methods

3.1 Data Collection

Our study utilizes earnings call transcripts as the primary source of data. These transcripts were collected from Seeking Alpha, targeting companies listed in the S&P 500 index, spanning from the fourth quarter of 2019 to the second quarter of 2023. We extracted the necessary details such as ticker symbol, quarter, date, and participants using HTML tags. Additionally, supplementary information such as industry, sector, region, capitalization, and size were obtained via scraping the Nasdaq website. Each transcript was divided into sections, typically 'Presentation' and 'Q&A', using HTML tags. Within the Q&A sessions, statements made by each speaker were sequentially recorded, ensuring the data preserved the flow of dialogue and interaction.

3.2 Data Preprocessing

We focused on the dialogues involving key corporate figures such as CFOs, Presidents, Chairmen, and CEOs. When an exchange pattern such as BOS-Analyst → President & CEO → JPMorgan-Analyst → Deutsche Bank-Analyst was identified, only the highlighted interactions were retained. All other non-sequential data were excluded to create concise, single-QA dialogues. Only questions pertaining to business conditions, forecasts, and economic outlooks were retained. Total 72 keywords were selected to filter questions related to future projections and market conditions, which included terms such as 'Outlook', 'Projection', 'Market conditions', 'Economic climate', and so on. This method ensured that only dialogues concerning strategic business outlooks and financial forecasts were processed for analysis.

3.3 Raccoon Dataset Construction

In this section, we detail the structure and procedures of our proposed benchmark dataset, as depicted in Figure 1. Our dataset is composed of several components, each tailored to reflect the decision-making process inherent in financial analysis.

The question transformation process involves converting the key content of each transcript into a concise query format. This aims to capture the essence of the transcript, ensuring that the questions generated encapsulate significant themes or central insights. For instance, if a transcript discusses a notable merger between two companies, the question derived from this discussion could be, "What are the potential financial impacts of the merger between Company A and Company B?"

We have identified 'aspect' as crucial elements in planning to these questions. Treated as search keywords, these aspects guide financial analysts in adequately addressing the queries. The aspects are organized sequentially to facilitate logical navigation through the search process. For instance, in responding to the merger question, aspects might include 'Market share implications', 'Regulatory hurdles', and 'Synergy realization timelines.'

For each aspect, we associated virtual knowledge entries that provide the necessary information to address the aspect effectively. These entries are designed to emulate the type of data an

analyst might encounter when investigating a particular aspect in real-world scenarios. An example of such a knowledge entry for 'synergistic savings' might state, "Historically, similar mergers have reported an average of 15% synergy savings within the first two years."

Furthermore, we have formalized a step-by-step reasoning process that links the question, aspect, and corresponding knowledge to generate coherent responses. This process mirrors the analytical thinking employed by financial analysts. The number of aspects and reasoning steps varies depending on the content of the original transcript.

Finally, we constructed a dataset comprising 50 such instances, each recorded with details such as the speaker, year, source data, and quarter. This structured approach not only enables the systematic simulation of financial analysis tasks but also serves as a robust framework for training machine learning models to emulate and generate human-like reasoning in the financial domain.

3.4 Categorization of Aspect

We categorized 'aspect' that is key part within financial transcripts, with a focus on real-world business scenarios. Each category is designed to highlight specific areas that are routinely evaluated by financial analysts. Below, we outline these seven principal categories, each accompanied by a revised example demonstrating diverse corporate perspectives.

Sales Portfolio Proportion: This category addresses the analysis of how sales are distributed among different products or services. It is crucial for understanding which segments are most lucrative or need strategic attention. For example, an aspect for Apple might be, "Proportion of total revenue derived from iPhone sales compared to other products."

Customer List and Proportion: This focuses on identifying key customers and their sales contribution, which is vital for assessing risks associated with customer concentration. An aspect for Microsoft might be, "Percentage of total revenue contributed by enterprise clients in the cloud sector."

Business Outlook: This category evaluates the prospects based on current conditions and planned strategies. An aspect for Tesla could be, "Expected growth in electric vehicle sales following the introduction of new model lines."

Business Growth Strategy: We examine strategic initiatives aimed at business expansion. An example aspect for Amazon might be, "Strategies for market expansion in Asia through AWS services."

Impact of Specific Events on Business: This category assesses the effect of external events on business operations. An aspect relevant to Nvidia might be, "Impact of global chip shortages on GPU production."

Determine Economic Conditions Relevant to Your Business: This involves understanding macroeconomic factors that could impact a company. An example aspect for Goldman Sachs might be, "Effects of current interest rate trends on investment banking profitability."

Sales and Operating Profit Guidance: This includes forecasts and expectations regarding sales and profitability, crucial for investor relations and strategic planning. An aspect for Coca-Cola might be, "Guidance on operating margins in light of fluctuating commodity prices."

3.5 Response Generation

To compare response between human analyst and LLM, we constructed a dataset using responses generated by the Azure GPT API. This dataset is designed to investigate differences in reasoning processes when specific informational aspects are provided or not.

- Human: Financial analyst who have at least three years of experience at securities firm research centers.
- GPT-3.5: This model was prompted to generate reasoning without any prior provision of specific aspects or contextual knowledge; however, it was provided with 5-shot reasoning examples.
- GPT-3.5 with Raccoon: The model received both the aspects and the reasoning demonstrated in the human 5-shot scenarios, thereby aligning its generation process more closely with the analyst.

4 Evaluations

4.1 Human evaluation

In this section, we outline the qualitative evaluation methodology utilized to assess the effectiveness of our dataset, which includes both

qualitative and quantitative evaluation methods. The qualitative analysis focuses on the aspects and reasoning processes that the models generate in response to the transcripts.

The evaluation of aspects was approached from two distinct perspectives. First, the criterion of implicitness was used to determine if an aspect directly addresses the question's content or requires implicit, expert-level background knowledge pertinent to the discussed company and industry. An aspect is labeled as implicit if it draws upon knowledge not explicitly stated in the question but necessary to fully grasp the context. For example, a question regarding Estée Lauder that necessitates consideration of the US-China trade conflict would lead to an aspect deemed implicit, as it involves significant external economic factors impacting the business scenario.

Second, we assessed the relevance of each aspect. An aspect is considered relevant if it directly aids in answering the question or clarifying the topic discussed. Conversely, an aspect is marked as irrelevant if it does not align with the theme of the question. For instance, an aspect that discusses financial strategies would be relevant to a query about a company's future growth projections but would be irrelevant to a question focusing on the environmental impact of the company's operations. Following the evaluation of aspects, we also examined the reasoning generated by the models. The consistency metric checks if the reasoning steps maintain thematic and logical coherence throughout the response. A reasoning process is deemed consistent if each step logically follows the preceding one, without any abrupt deviations or shifts in logic. For example, reasoning that begins with a discussion on financial growth due to market expansion and then abruptly shifts to product quality without a logical link is considered inconsistent.

Lastly, the specificity of each reasoning step is evaluated based on its grounding in specific, verifiable data or detailed logical argumentation. Reasoning is classified as specific if it includes concrete data, references, or clearly defined logic. It is deemed nonspecific if it largely relies on vague statements or assumptions without significant support. For instance, a statement like, "The company will likely see a 10% increase in sales due to the new product launch, as indicated by early market tests," exemplifies specific reasoning. Conversely, a generalized statement such as, "The company will do better because it has good products," lacks specificity due to its reliance on broad, unsubstantiated claims.

4.2 Token Overlap

In the process of our quantitative evaluation, we assess the token overlap ratio to determine the lexical similarity between the generated aspects and the corresponding question. We compare the tokens from the question with those of the aspects to investigate whether the generated aspects are directly extracted from the question content or whether they introduce novel yet related concepts. The initial step involves tokenizing the question sentences and their corresponding aspects using a GPT tokenizer. Let T denote the tokenizer function, Q represents the question sentence, and A_n signifies the n-th aspect derived from Q. The token overlap is computed as follows:

$$\text{Overlap}(Q, A_n) = \frac{|T(Q) \cap T(A_n)|}{|T(Q) \cup T(A_n)|}$$

The token overlap value ranges from 0 to 1, where 0 denotes no overlap and 1 indicates complete duplication. This methodology facilitates the quantitative evaluation of the shared lexical content between two sequences. Such a metric proves especially valuable in tasks necessitating the measurement of lexical similarity, including paraphrase detection and text entailment.

4.3 Perplexity

The second metric for quantitative assessment is the comparison of perplexity across pre-trained large-scale language models. We measure the perplexity of the aspect generated by the language model when given a question as input. Perplexity serves as a metric to evaluate the likelihood of the sentences produced by the language model. Given a sequence of output tokens $Y = \{y_0, \dots, y_m\}$, we calculate the perplexity as follows:

$$\text{PPL}(Y) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{p(y_i|y_0, \dots, y_{i-1})}}$$

Therefore, to verify the generative plausibility of the aspect corresponding to the question, we compare the aspects generated by humans and those produced by the GPT for the same query.

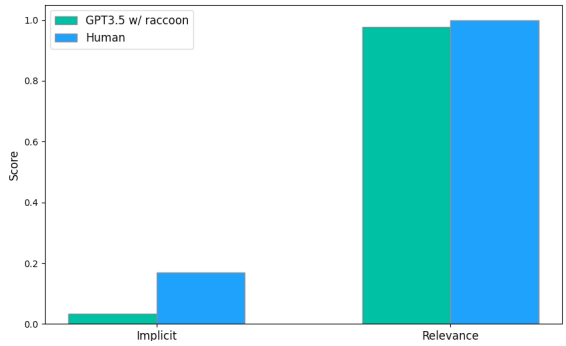

Figure 2: The performance of Human and GPT on aspect generation

5 Results

5.1 Aspect Analysis

To evaluate the performance of GPT in generating aspects, we conducted a human evaluation, as depicted in Figure 2. We found that GPT-generated responses contained a considerably lower proportion of implicit aspects compared to those generated by humans. The scores are normalized to a maximum of 1 point.

We classified the generated aspects into distinct categories and examined their distribution. According to the data presented in Figure 3, it is evident that human participants distributed their responses evenly across seven categories, with the distribution resembling an approximately uniform spread. In contrast, the aspects generated by GPT were disproportionately concentrated in certain categories, with more than half of the responses falling into specific ones. This analysis underscores a significant difference in the approach to problem-solving between humans and the GPT model. Humans tend to employ a diverse range of perspectives when addressing a question, which is reflected in the even distribution of response categories. On the other hand, GPT shows a tendency to focus narrowly on fewer categories, indicating a limitation in the model's ability to diversify its approach and consider multiple aspects of a problem. This pattern suggests that while GPT can effectively generate responses, its capacity to mimic the multifaceted approach typical of human reasoning is still constrained.

To substantiate this hypothesis, we examined the proportion of token overlap between the questions and the generated aspects. According to the data presented in Figure 4, the token overlap in aspects generated by humans was statistically significantly lower compared to those generated by GPT. This suggests that human participants tend to generate more varied and conceptually distinct aspects that do not merely repeat the tokens present in the questions.

On the other hand, GPT demonstrated a higher tendency to reuse tokens from the questions in its generated aspects. This behavior indicates a more literal or direct interpretation and utilization of the input text, which may limit the model's ability to generate responses that introduce new or diverse perspectives independent of the explicit content of the questions. This pattern provides quantitative support for the earlier observation that GPT, while capable of generating relevant aspects, tends to do so in a less diverse and more question-bound manner compared to human responses.

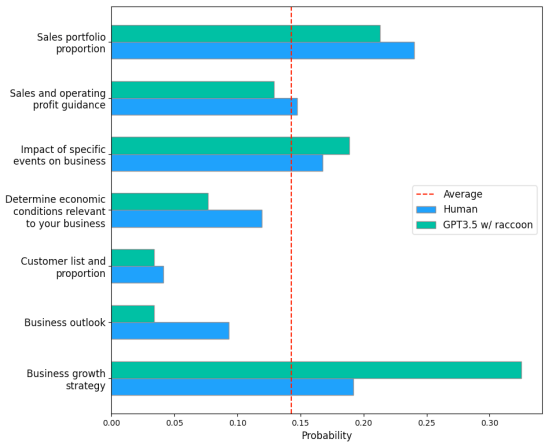

Figure 3: Categorical comparison of aspects

5.2 Reasoning Analysis

To compare reasoning steps, we discuss the methods used to compare reasoning steps by evaluating consistency, specificity, and answer validation across responses generated by human analysts and GPT models in Figure 5.

Consistency in reasoning was analyzed to determine how often GPT models repeated concepts from previous steps in subsequent reasoning processes, revealing a strong tendency towards redundancy. In contrast, human respondents frequently introduced logical leaps between steps, indicative of a more dynamic and less linear reasoning approach.

Specificity was assessed in scenarios where both GPT-3.5 with Raccoon and human participants were provided with specific aspects and knowledge. We observed that each reasoning step effectively referenced the necessary

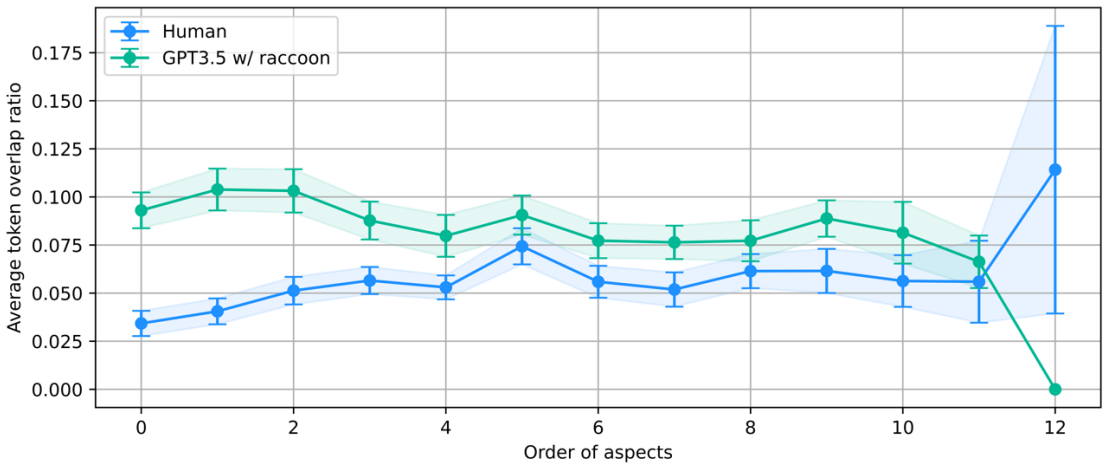

Figure 4: Averaged token overlap ratio for each number of aspects

information. This observation suggests that supplying well-defined aspects and relevant knowledge can reduce instances of 'hallucination' in GPT responses, where the model generates irrelevant or fictitious content.

Regarding answer validation, it was noted that GPT-3.5 often avoided definitive conclusions. Particularly with the GPT-3.5 Raccoon configuration, as the reasoning progressed, there was a noticeable tendency to generate conclusions that were not pertinent to the initial question posed. This pattern underscores a challenge in maintaining the relevance of the responses as the model attempts to integrate and reason with the provided knowledge and aspects.

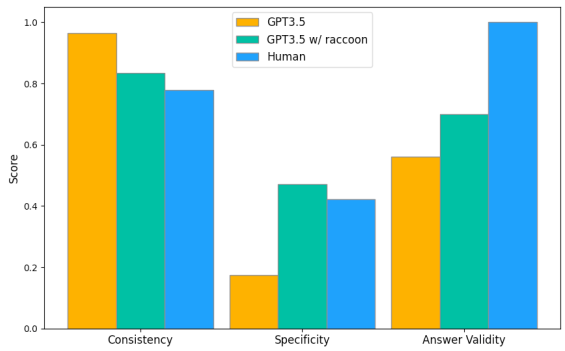

Figure 5: The performance of Human and GPT on reasoning generation

These evaluations underscore significant differences in reasoning quality between human analysts and language models, especially in maintaining consistency, utilizing relevant knowledge effectively, and producing valid conclusions. The insights from this comparative analysis are crucial for understanding and enhancing the reasoning capabilities of AI models in complex analytical tasks.

5.3 Perplexity Analysis

In furthering our examination of the generative differences between GPT and human responses, we explored whether the observed patterns were specific to Azure's GPT-3.5 API by comparing the PPL of aspects generated by GPT (Figure 6). For this purpose, we used the following prompt template for the perplexity evaluation and ensured that the same prompt was used for all models:

> You are a financial domain expert analyst. Please create search queries to answer questions related to the given ticker's company.
> Ticker: {*ticker_id*}
> Question: {*question*}
> Aspects: {*aspect_list*}

In the template, {*ticker_id*} represents the target company's stock symbol, {*question*} is the specific inquiry, and {*aspect_list*} contains either aspects generated by GPT or those created by human analysts. We calculated the PPL for a total of 50 examples.

Across the board, it was observed that the aspects generated by humans yielded relatively higher PPL values in all public models (Llama-2, Llama-3, Mistral, Phi-2, Falcon) compared to those generated by GPT. Notably, statistical significance in PPL differences was found within the outputs of Llama-2, Llama-3, and Phi-2 models. These

findings suggest that decoder-based models indeed generate in a manner that is distinctly different from human creation. The significant disparities in PPL values underscore the dissimilarity in the naturalness or predictability of language between the two sources. While GPT-generated aspects tend to align more closely with the language patterns the models have been trained on, human-generated aspects seem to reflect a broader and possibly more unpredictable range of language use within the financial domain.

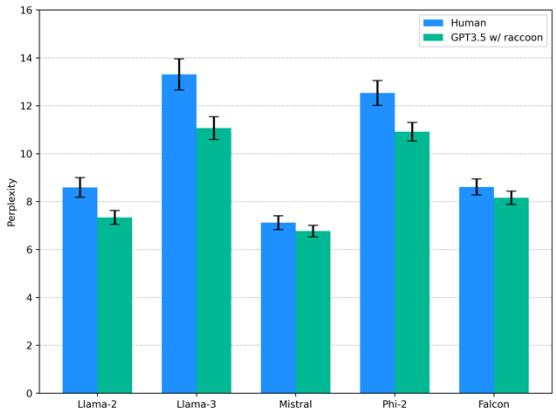

Figure 6: Comparative perplexity analysis of Human and GPT outputs across public LLMs.

6 Conclusion

In conclusion, our comprehensive investigation into the cognitive capabilities of LLMs within the financial domain has highlighted both the strengths and limitations inherent in these advanced AI systems. By examining the performance of LLMs in comparison to that of human experts in complex financial analysis tasks, our study has illuminated significant discrepancies in the depth and authenticity of the models' reasoning processes. While LLMs excel in producing relevant and logically coherent responses, they predominantly rely on explicit cues from input questions, which often limits their ability to generate diverse perspectives and understand implicit content.

Furthermore, despite enhancements through our 'Raccoon' configuration—which provides structured aspects and reasoning paths—challenges persist in ensuring consistency and relevance throughout the reasoning process. This configuration has indeed improved the performance of GPT models, but it also reveals that even advanced models struggle to match the nuanced understanding and broader information integration displayed by human analysts. This suggests that while LLMs can generate syntactically correct and contextually appropriate answers, they lack the human-like ability to seamlessly navigate and link multiple domains of knowledge, often resulting in a more constrained analytical scope.

Overall, the findings from this research underscore the critical need for continuous evolution in the design and development of LLMs, especially if they are to be effectively employed in complex, real-world tasks like financial analysis. By aligning model development with insights gained from rigorous comparative analyses with human cognitive processes, there is significant potential to enhance LLMs' capabilities. Such advancements could make these models not just supplementary tools but robust partners in augmenting human expertise, thereby ensuring their efficacy and reliability in practical applications across various domains.

References

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell. 2020. Language models are few-shot learners. Advances in neural information processing systems **33**: 1877-1901.

Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang and Y. Wang. 2024. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3): 1-45.

Chen, W., X. Ma, X. Wang and W. W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588.

Chen, Z., W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang and B. Routledge. 2021. Finqa: A dataset of numerical reasoning over financial data. arXiv preprint arXiv:2109.00122.

Cui, J., Z. Li, Y. Yan, B. Chen and L. Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092.

Huang, J., X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song and D. Zhou. 2023. Large

language models cannot self-correct reasoning yet. arXiv preprint arXiv:2310.01798.

Jin, M., H. Tang, C. Zhang, Q. Yu, C. Liu, S. Zhu, Y. Zhang and M. Du. 2024. Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities. arXiv preprint arXiv:2402.10835.

Kwon, T., K. T.-i. Ong, D. Kang, S. Moon, J. R. Lee, D. Hwang, B. Sohn, Y. Sim, D. Lee and J. Yeo. 2024. Large Language Models Are Clinical Reasoners: Reasoning-Aware Diagnosis Framework with Prompt-Generated Rationales. Proceedings of the AAAI Conference on Artificial Intelligence.

Li, Z., X. Xu, T. Shen, C. Xu, J.-C. Gu and C. Tao. 2024. Leveraging large language models for nlg evaluation: A survey. arXiv preprint arXiv:2401.07103.

Liang, P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu and A. Kumar. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

Lu, P., B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu and J. Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. Advances in Neural Information Processing Systems **36**.

Mavi, V., A. Saparov and C. Zhao. 2023. Retrieval-Augmented Chain-of-Thought in Semi-structured Domains. arXiv preprint arXiv:2310.14435.

Momennejad, I., H. Hasanbeig, F. Vieira Frujeri, H. Sharma, N. Jojic, H. Palangi, R. Ness and J. Larson. 2024. Evaluating cognitive maps and planning in large language models with CogEval. Advances in Neural Information Processing Systems **36**.

Phogat, K. S., C. Harsha, S. Dasaratha, S. Ramakrishna and S. A. Puranam. 2023. Zero-Shot Question Answering over Financial Documents using Large Language Models. arXiv preprint arXiv:2311.14722.

Rosoł, M., J. S. Gąsior, J. Łaba, K. Korzeniewski and M. Młyńczak. 2023. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Scientific Reports **13**(1): 20512.

Son, G., H. Jung, M. Hahm, K. Na and S. Jin. 2023. Beyond classification: Financial reasoning in state-of-the-art language models. arXiv preprint arXiv:2305.01505.

Song, Y., W. Xiong, D. Zhu, W. Wu, H. Qian, M. Song, H. Huang, C. Li, K. Wang and R. Yao. 2023. Restgpt: Connecting large language models with real-world restful apis. arXiv preprint arXiv:2306.06624.

Sun, J., C. Zheng, E. Xie, Z. Liu, R. Chu, J. Qiu, J. Xu, M. Ding, H. Li and M. Geng. 2023. A survey of reasoning with foundation models. arXiv preprint arXiv:2312.11562.

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery and D. Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le and D. Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**: 24824-24837.

Yang, Z., P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov and C. D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.

Zhang, Z., A. Zhang, M. Li and A. Smola. 2022. Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493.