Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231059

FundRecLLM: Fund Recommendation Based on Financial News and Research Analyst Report

Guang YANG [a,1], Peiyang HE [a] and Xuefeng LIU [a]
[a] *Amazon*

Abstract. Adopting AI in financial advisory is a challenging task as there exists multiple sources of information to digest and interpret. Such information consumption process are very lengthy for financial advisors, reducing the efficiency and timeliness for their advice and recommendation given to their clients. In this work, we introduce a multi-step framework that consumes and combines news and industry-focused fund research analyst report to assist in fund recommendation process using Large Language Models (LLMs). To quantitatively evaluate the efficacy of the approach, we track the weekly and monthly market performance of representative industry-focused fund after news and report released date, and compute a Normalized Discounted Cumulative Gain (NDCG) score between the rankings of the fund performance and recommendation rating scores. We find that utilizing analyst report and self consistency in the framework increase the NDCG score from 0.72 to 0.93 comparing to consuming news only without self consistency, based on the time frame of our experimental evaluation.

Keywords. large language model, generative AI, financial advisory, information extraction, text summarization

1. Introduction

Since its introduction, Large Language Models (LLMs) have demonstrated its power in handling complex tasks in various areas [1]. Its capability is continuously evolving and advancing to handle various challenges in different domains [2]. Although LLMs have been explored in finance area for various use cases, their strength exploitation remains on the topics of information extraction, summarization and synthesis. For the purpose of investment advisories, use cases are still constrained to certain types of classification such as sentiment analysis for specific stocks using headlines related to that stock or social media tweets or posts. In [3], ChatGPT is used to determine a sentiment score for a given news headlines. ChatGPT outperforms other basic models based on a numerical score that calculates correlation between "ChatGPT scores" and subsequent daily stock market returns. In [4], PaLM is used to generate financial sentiment labels (i.e. whether the stock price should go up, down or not sure.) for Reddit posts. In [5], FinBERT, a cus-

[1]Corresponding Author: Guang Yang, Amazon, Beijing Aerospace Building Tower 2, Shenzhen, China; E-mail: yaguan@amazon.com.

tomized LLM for financial domain, is used to generate sentiment measures for analyst report sentences, earnings conference call script, and also labels for ESG discussions. LLMs' capability in understanding and reasoning impact on broader financial markets (e.g. industries like travel and agricultural) has not been studied. In this work, we investigate LLMs' such competency over general news headlines and raw analyst report. On top of sentiment classification, we ask LLMs to provide specific recommendation on selected industry-focused funds and also give out specific reasons backed by the facts presented in the news and analyst report.

As making such financial advisories requires consolidated consumption of data from various sources, an intelligent agent needs to take multiple steps to reach to a conclusion. A number of studies have been conducted to examine the capability of LLMs to reason and execute complex tasks that involve multiple steps. CoT [6] is introduced to improve the multi-step reasoning ability of LLMs by explicitly instructing the model to generate intermediate reasoning steps. The reasoning path mimics the reasoning process a person might employ in solving a task. It has been observed that chain-of-thought prompting significantly improves model performance across a variety of multi-step reasoning tasks. As task gets more complicated, each reasoning step can rely on external tools to support computation beyond the core LLM capabilities. ToolLLM [7] is introduced to facilitate tool-use capabilities within open-source LLMs. To interleave the reasoning and acting (i.e. tool usage), ReAct framework [8] is introduced that uses LLMs to generate both reasoning traces and task-specific actions. On the other hand, instead of solving an overarching task in a single run, the concept of *Chaining LLM* [9] is introduced to divide a complex task into sub-tasks and use LLMs to solve sub-tasks sequentially. Each subtask can be completed by an independent run of an LLM, and the output of one or more sub-tasks is used as input for the next.

In this work, we proposed a framework named *FundRecLLM* that combines financial news and fund research report to recommend industry focused fund assets using LLMs. Our contributions include: (1) a chained tool that employs LLMs to generate automated answers that recommend industry focused funds based on news and analyst research report interpretations; (2) a general method to consume raw analyst research report, a typical finance material often released in pdf formats, from the perspective of financial advisories, and (3) a quantitative evaluation method of recommendation using fund market performance.

2. Methodology

Figure 1 shows the overall solution architecture and major components for *FundRecLLM*. A news interpreter is built to interpret the news through CoT prompting. And then important entities (such as industry, sentiment and reason) are extracted from the answers for aggregation and target search. Summaries are stored in a docstore for later usage. A report analyzer is implemented to consume the analyst research report starting from the raw pdf format. Texts blocks are extracted and ordered using a document layout parser and an OCR agent. Since most analyst reports are over 10 pages, the text blocks are normally over the context window size of most LLMs (especially for the open sourced ones), these blocks are processed and interpreted through a Map & Reduce fashion. Similarly to the news interpreter, important entities are extracted from the answers and summaries

are stored in a docstore. Finally, a moderator which is a sequential chain to link entities, consume the docstores from the news and report, and give final answers for the recommended funds with market insights summarization. Entities are linked through the same mapped industries from the news and reports. The sentiments are converted to scores and the scores are summed for news and report separately and then averaged across news and report for the same industry. The summary insights are also concatenated for the same industry.

Let $n_i \in N$ represents a piece of news, $r_j \in R$ represents an analyst report, $d_k \in D$ stands for the collection of target industries in the financial market. $NI_{l,t_{NI}}$, $RA_{l,t_{RA_m},t_{RA_r}}$ stands for the news interpreter and the report analyzer respectively while l denotes the chosen LLM and t denotes the prompt template. A step-by-step breakdown of the methodology is as follows, note that the report analyzer has a map prompt t_{RA_m} and a reduce prompt t_{RA_r}:

(1) Iterate through news data n_i and pass each raw news content to the input variable of the LLM prompt template t_{NI}. Apply LLM to interpret multiple times (M) for self-consistency. The sentiments (s_i^{NI}) and reasoning summaries (u_i^{NI}) on target industries are obtained $s_i^{NI}, u_i^{NI} = \cap_M NI_{l,t_{NI}}(n_i)$. Then the sentiments ($S_{d_k}^{NI}$) and summaries ($U_{d_k}^{NI}$) are aggregated for each target industry $S_{d_k}^{NI} = \sum_N (s_i^{NI}|d_k)$, $U_{d_k}^{NI} = \cup_N (u_i^{NI}|d_k)$.

(2) Iterate through analyst reports r_j. Send each report to a layout parser lp to get the text regions $tr_{j,p} \in TR_{j,p}$ for each page ($p_j \in P_j$). Crop the text regions ($TR_{j,p}$) for each page and pass them through an OCR engine o to get the text strings ($ts_{j,p}$) for each page. The whole process can be described as $ts_j = \cup_{P_j} \cup_{TR_{j,p}} o(lp(r_j))$.

(3) Apply LLM to summarize the texts for each page of the analyst report $U_{m,j} = RA_{l,t_{RA_m}}(ts_j)$. And then pass the summaries of all pages to the report analyzer to get the sentiments (s_j^{RA}) and reasoning summaries (u_j^{RA}) on target industries. This is also repeated multiple times (M) for self consistency $s_j^{RA}, u_j^{RA} = \cap_M RA_{l,t_{RA_r}}(\cup U_{m,j})$. And similarly to the news interpreter, the sentiments and summaries are aggregated $S_{d_k}^{RA} = \sum_R (s_j^{RA}|d_k)$, $U_{d_k}^{RA} = \cup_R (u_j^{RA}|d_k)$.

(4) Aggregate the sentiments and the reasoning summaries across the news source and analyst reports $S_{d_k} = avg(S_{d_k}^{NI}, S_{d_k}^{RA})$, $U_{d_k} = U_{d_k}^{NI} \oplus U_{d_k}^{RA}$.

More details of the key components are discussed in the following sections.

2.1. Sequential chain and docstore

We use a sequential chain to find relevant insights for the target industry from the docstores that are artifacts of the news interpreter and the report analyzer. The sequential chain is implemented with LangChain Sequential Chain and the docstores contain aggregated summaries for target industries. The relevant summaries for the recommended industries are retrieved through embedding space similarity search.

2.2. Layout detection and text OCR for industry research report

Layout Parser library [10] is used to parse the tables, texts and figures out from the pdf file. We use `mask-rcnn-X-101-32x8d-FPN-3x` as the layout detection model which is trained on PubLayNet dataset.

For the OCR engine, we use Tesseract engine which is natively supported within *Layout Parser*. Tesseract supports multiple languages, including simplified Chinese in our experiments.

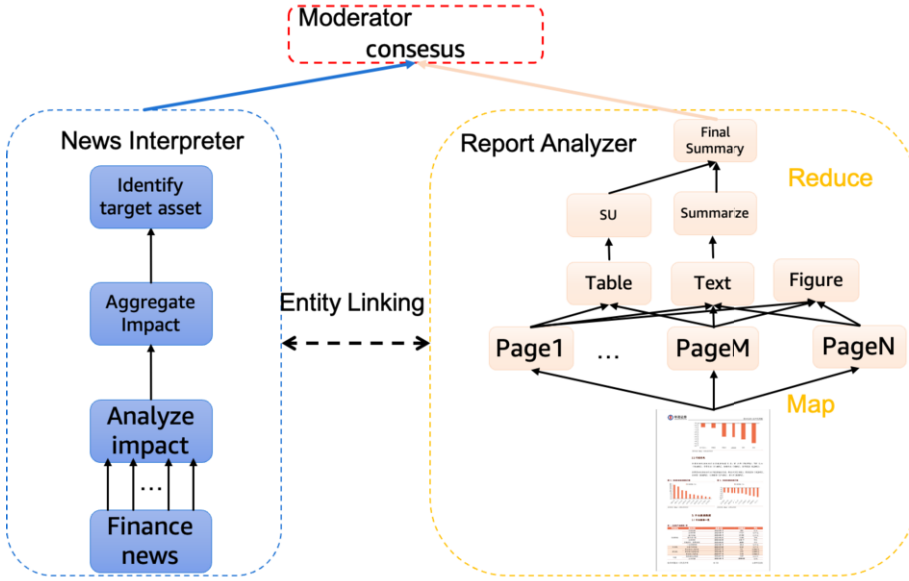


Figure 1. Components of *FundRecLLM*. The model includes a news interpreter to interpret the news, a report analyzer to consume the financial research report, and a moderator to extract, link and aggregate structured and unstructured information from the news interpreter and the report analyzer to reach to a consensus.

2.3. CoT prompting and entity extraction

We adopt CoT to make the LLM answer sub questions, thus implicitly forcing it to think in steps before drawing a conclusion. We instruct the LLM to give concrete answers based on the facts outlined in the source texts and output in json format for easier parsing.

2.4. Mitigation of LLM output stochasticity

One typical challenge of LLMs is model output stochasticity. Self-consistency marginalization [11] is proposed to mitigating the stochasticity of a single sampled generation while avoiding the repetitiveness and local-optimality that plague greedy decoding produces. It first samples a diverse set of reasoning paths instead of only taking the greedy one, and then selects the most consistent answer by marginalizing out the sampled reasoning paths. In this work, we generate multiple interpretations over the news and research report and marginalize the answers through majority voting.

3. Experiment

3.1. Datasets

To examine the proposed solution, we scrape and manually filter out a small dataset containing 251 financial news that are not specific to individual stocks and 141 industry research analyst reports from various brokers from Chinese public websites (e.g. eastmoney). We filter information related to individual stocks as we are interested in

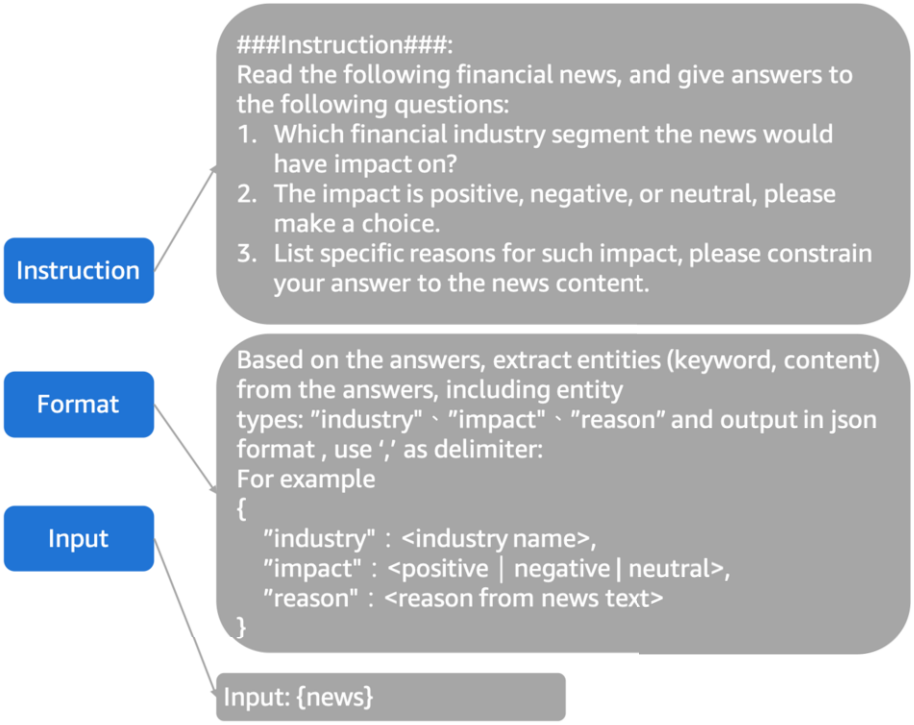


Figure 2. Prompt for News Interpreter.

industry-focused funds for this study. The news and the analyst reports are released on 07/17/2023. We follow the Shenwan(SW) standard for the industry categorization hierarchy standards (three levels) and identify 5 themed-funds for each level I industry on average. The daily price for each fund after 07/17/2023 is obtained using *xalpha*, and the weekly and monthly fund performance is averaged for each level I industry.

3.2. Large language model and prompts

Because the public dataset we obtain are in Chinese, we choose ChatGLM as the LLM for our experiment as ChatGLM is optimized for Chinese QA and dialogue [12]. To enable self consistency, we set the temperature to be 0.8 and generate 5 samples for each same prompt for both news interpreter and report analyzer. The detailed news interpretation prompt is shown in Figure 2. The Map & Reduce instruction prompting for analyst report is not included because of space limit.

4. Experimental results

Based on the experimental results, we find that ChatGLM can logically reason the impacts on financial markets through implicit factors that is not specifically disclosed in the raw news content. Through Map and Reduce operation for analyst reports, the proposed approach can well cover multiple aspects included in the report while overcoming the

limit of the context window size of ChatGLM. However, the target industry and overall recommendation/rating are diverged from the analyst at times upon close examination, since the conclusion from the analyst are often summarized in one page (usually between page 1 to 3).

An example of the extracted structured information from LLM's interpretation on news and analyst report table is shown in Table 1. An interesting observation is made on the real estate sector, for the LLM's interpretation difference between news and analyst report. The sentiment on the news for the housing market is pretty negative and therefore the score rating for real estate is pretty low. On the other hand, for the analyst report, analysts acknowledge that recent downward trend of the real estate market. However, analysts recognize there are great values to the valuable assets that good companies with a long history hold. In addition, central bureaus have provided incentives for housing markets to rebound. Note that the original content is in Chinese, here content is translated into English for readers.

Table 1. Examples of output from the news interpreter and the report analyzer.

Source	Industry	Sentiment	Reason
News	Real estate	Neutral	The National Bureau of Statistics said that in the next stage, it is necessary to adhere to the positioning that "houses are for living in, not for speculation", accelerate the establishment of a housing system with multi-subject supply, multi-channel security, and rent-purchase parallel...
News	Real estate	Negative	The real estate market is shifting from high-speed development in the past to stable development. This is determined by the change in the development stage of the real estate market and the adjustment of market supply and demand relations.
News	Real estate	Negative	Nationwide real estate development investment decreased by 7.9% year-on-year, of which residential investment decreased by 7.3%.
News	Chemical	Neutral	The company is committed to creating a second growth curve for high value-added additives, but did not mention the specific impact on the chemical industry.
News	Leisure services	Positive	The Asian Games is conducive to the development of the industry. Many companies are actively participating in tourism-related work. Hotel booking has reached to the level pre-pandemic.
Report	Real estate	Positive	The report's assessment of real estate investment is positive. Although the real estate market shows a downward trend, overall, the real estate industry still has strong investment value. In addition, the adjustment of credit policies, such as the RRR cut by the People's Bank of China, has a positive impact on the real estate industry.
Report	Chemical	Positive	- Market demand: With the recovery of the global economy, market demand in the chemical industry has gradually recovered. Especially after the COVID-19 epidemic, domestic and foreign market demand has gradually recovered, driving up prices. - Fundamental turning point: The profitability of some chemical companies has turned upward from the profit turning point in Q1 to Q2, which means that the market may continue to rebound.- Technical turning point: Chemical product prices have stopped falling and rebounded recently, suggesting that the market may strengthen.

As there is no ground truth for the textual answers given by *FundRecLLM*, to quantitatively evaluate the recommendation answers rendered by the solution, we design an evaluation schema based on the ratings for the target industry converted from the sentiments and the corresponding industry-focused fund performance. Steps of the calculation are as followings:

(1) Firstly, we parse the sentiment part of the answer for the recommended fund and convert it to a score. The conversion from sentiment to a score rating is shown in Table 2. The conversion for report is different as we find analysts usually write reports to recommend to buy. A neutral sentiment is just a weaker statement to buy in.

Table 2. Conversion from sentiment to recommendation score rating.

Source of information	Sentiment	Score
News	Positive	1
News	Neutral	0
News	Negative	-1
Report	Positive	2
Report	Neutral	1
Report	Negative	-1

(2) Secondly, we aggregate the scores for each industry and rank the scores. While aggregating the industry rating scores, we observe that LLM's interpretation on the industry can be on different levels according to SW standard, and the industry concept from the output are not exact match with the standards. To facilitate the analysis, we use *Sentence Transformer* to find the closet industry concept within SW standards at all levels based on word similarity and map all level III and level II industry to level I industry.

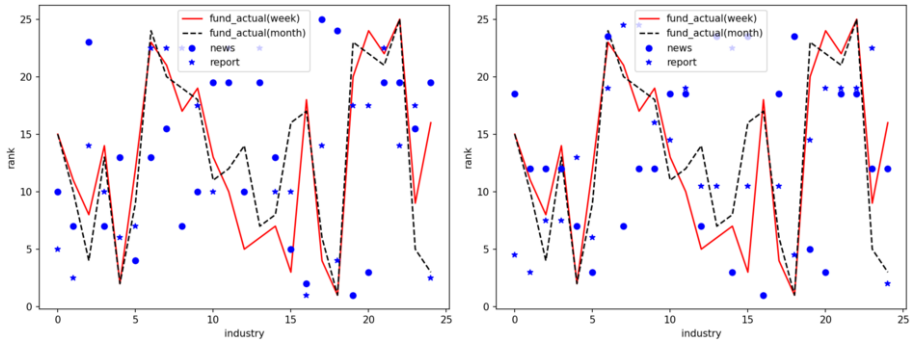
(3) We then calculate market performances for the following week and month of all the related level I industry sectors after 07/17/2023, using representative funds for each industry. Their market performance is calculated as the relative week-on-week and month-on-month ending price change. Then their performances are averaged for each level I industry sector and ranked accordingly. Market performances and the scores for selected industries are shown in Table 3. The rankings for different industry market performance and the sentiment ratings aggregated from both news and analyst report are shown in Figure 3.

(4) Lastly we calculate the Normalized Discounted Cumulative Gain (NDCG) between the ranks. NDCG is a common evaluation metric for comparing rankings.

Table 4 lists the NDCG score under different scenarios. To combine news and report ratings, a simple average is taken before ranking. It is observed that applying self consistency improves the NDCG score for news ratings but does not do so for report. Taking a closer look, by majority voting through more samples, news interpretations converges to a more frequent and reasonable interpretation. On the other hand, report content are much longer and covers multiple aspects. After map and reduce operation, model's interpretation are more diverse. Only strong sector signals such as public sector, financial services and real estate are dominating the rankings for report. In the previous discussion, we mention that the model's interpretation on analyst report towards real estate sector is much more positive compared to that of news. And indeed, market price for real estate sector has rebounded from its bottom since 07/17/2023. This helps to explain that why

Table 3. Fund market performances (shown as fund price change percentages) and sentiment scores with rankings for selected industries.

Industry	Weekly change(rank)	Monthly change(rank)	News score(rank)	Report score(rank)
Real estate	2.86%(1)	11.70%(1)	-1(23)	9(4)
Building materials	1.75%(2)	6.67%(2)	2(7)	5(13)
Leisure services	-0.21%(7)	1.40%(8)	4(3)	1(22)
Chemical	-0.79%(11)	-1.10%(10)	1(17)	12(3)
Auto manufacturing	-2.81%(19)	-3.34%(18)	1(12)	3(16)
Electrical equipment	-4.37%(23)	-9.04%(24)	-1(23)	2(19)



(a) (b)

Figure 3. Fund performance rank against news and report sentiment scoring rank: (a) w/o self consistency (b) w/ self consistency.

the ranking from the report ratings has a higher NDCG score compared with that from the news ratings.

Table 4. NDCG score between fund recommendation rating and fund performance ranking under different experiment setting.

Source of information	w/ self consistency	NDCG@weekly	NDCG@monthly
News Only	No	0.73	0.72
News Only	Yes	0.77	0.82
Report Only	No	0.89	0.90
Report Only	Yes	0.89	0.90
News & Report	No	0.78	0.79
News & Report	Yes	**0.91**	**0.93**

5. Conclusion

In this study, we investigate utilizing LLMs in financial advisories setting. We model our problem as a recommendation and Q&A problem and then design a solution framework of LLMs using a sequential chain of LLMs for interpreting news and analyst reports

respectively. The main challenge we face in our solution design method is on integrating multi-modal data sources and overcoming the context window size limit of the chosen LLM. We develop two parallel pipelines to consume and interpret news and raw analyst reports, and an intelligent agent as a moderator to synthesize the information and make the final conclusion.

The proposed solution and evaluation framework has certain limitations which direct us for future investigation. Firstly, the NDCG score is a noisy evaluation of recommendation efficacy as such measure captures both the accuracy of sentiment classification and how the market reacts to the sentiments. Given the time constraint, we are only able to process limited amount of information released at a single date that is quite near-term. Therefore, only short-term market performance is tracked for the evaluation. More backtests and more evaluation methods are needed to corroborate the effectiveness of the proposed solution. Secondly, we use all pages of the analyst report for the report analyzer. However, since most key contents are in earlier pages (i.e. page 1-2), an ablation study is needed to examine the effect of specifying page range for analysis. Lastly, we only use ChatGLM as the chosen LLM for our experiments. As there are multiple instruction LLMs and customized LLMs specifically for finance domain, it will be interesting study to compare with other alternatives.

References

[1] Kaddour, J, Harris, J, Mozes, M, Bradley, H, Raileanu, R, and McHardy, R. Challenges and Applications of Large Language Models. ArXiv abs/2307.10169 (2023).

[2] Naveed, H, Khan, AU, Qiu, S, Saqib, M, Anwar, S, Usman, M, Barnes, N, and Mian, AS. A Comprehensive Overview of Large Language Models. ArXiv abs/2307.06435 (2023).

[3] Lopez-Lira, A, and Tang, Y. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. ArXiv abs/2304.07619 (2023).

[4] Deng, X, Bashlovkina, V, Han, F, Baumgartner, S, and Bendersky, M. What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. Companion Proceedings of the ACM Web Conference 2023; 2023 30 April - 4 May; Austin, TX. Association for Computing Machinery; c2023. p. 107–110.

[5] Huang, AH, Wang, H and Yang, Y. FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemp Account Res, 2023; 40: 806-841. https://doi.org/10.1111/1911-3846.12832

[6] Wei, J, Wang, X, Schuurmans, D, Bosma, M, Chi, EH, Xia, F, Le, Q, and Zhou, D. Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv abs/2201.11903 (2022).

[7] Qin, Y, Liang, S, Ye, Y, Zhu, K, Yan, L, Lu, Y, Lin, Y, Cong, X, Tang, X, Qian, B, Zhao, S, Tian, R, Xie, R, Zhou, J, Gerstein, MH, Li, D, Liu, Z, and Sun, M. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. ArXiv abs/2307.16789 (2023).

[8] Wu, TS, Terry, M, and Cai, CJ. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; 2022 29 April - 5 May; New Orleans, LA. Association for Computing Machinery; c2022. p. 1–22.

[9] Shen, Z, Zhang, R, Dell, M, Lee, B, Carlson, J, and Li, W. LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. In: Lladós, J, Lopresti, D, Uchida, S, editors. Document Analysis and Recognition – ICDAR 2021. 2021 September 5–10; Lausanne, Switzerland. p. 131–146.

[10] Wang, X, Wei, J, Schuurmans, D, Le, Q, Chi, EH, and Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ArXiv, abs/2203.11171 (2022).

[11] Du, Z, Qian, Y, Liu, X, Ding, M, Qiu, J, Yang, Z, and Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022 May 22-27; Dublin, Ireland. Association for Computational Linguistics. p. 320–335.