Designing Heterogeneous LLM Agents for Financial Sentiment Analysis

FRANK XING, Department of Information Systems and Analytics, National University of Singapore, Singapore, Singapore

Large language models (LLMs) have drastically changed the possible ways to design intelligent systems, shifting the focus from massive data acquisition and new model training to human alignment and strategical elicitation of the full potential of existing pre-trained models. This paradigm shift, however, is not fully realized in financial sentiment analysis (FSA), due to the discriminative nature of this task and a lack of prescriptive knowledge of how to leverage existing generative models in such a context. This study investigates the effectiveness of the new paradigm, i.e., using LLMs without fine-tuning for FSA. Rooted in Minsky's theory of mind and emotions, a design framework with heterogeneous LLM agents is proposed and applied to FSA. The framework instantiates specialized agents using prior guiding knowledge from both linguistics and finance. Then, a summative agent reasons on the aggregated agent discussions. Comprehensive evaluations using six FSA datasets show that the framework yields better accuracies compared to many alternative multi-LLM agent settings, especially when the discussion contents are substantial. This study contributes to the design foundations and paves new avenues for LLMs-based FSA and potentially other tasks. Lastly, implications for business and management have also been discussed.

CCS Concepts: • **Information systems → Decision support systems**; • **Computing methodologies → Natural language processing**.

Additional Key Words and Phrases: large language model, financial sentiment analysis, agent discussion, theory of emotion, design science research

1 Introduction

Since OpenAI's ChatGPT went viral one year ago, large language models (LLMs) have gone through fast improvements, showing a variety of capabilities. The AI adaptation for many financial services is accelerating, and big data-supported financial decision-making is no exception. Financial sentiment analysis (FSA) is a prototypical task in that category and is becoming increasingly important as financial service processes and our social behavior digitalize: companies disclose electronic versions of their annual reports, earnings calls, and announcements, and investors join online communities, discussion forums, and social media to interact with others. The recent GameStop Saga [11] and the popularity of a spectrum of market sentiment indexes (e.g., MarketPsych [37]) have shown clear evidence that sentiment is a useful analytics tool for financial decision-making, forecasting short-term returns and volatilities [46], detecting fake news and fraud [14], and predicting risk [49]. The usefulness and the importance of accurate FSA are also underpinned by a long thread of research [3, 8, 12, 45]. Hendershott et al. [21] summarized that research on the application of AI on news, social media, and word-of-mouth data is a major category of leveraging AI in finance. Considering the wide usage in both academia and industry, accurate FSA is desired for multiple stakeholders.

The majority of FSA systems were developed in the past decade and their architecture and design ideas have gone through several iterations along with the advances in natural language processing. *Early systems* rely on

Authors' Contact Information: Frank Xing, Department of Information Systems and Analytics, National University of Singapore, Singapore, Singapore; e-mail: xing@nus.edu.sg.

ACM 2158-6578/2024/8-ART
https://doi.org/10.1145/3688399

sentiment word dictionaries and simple rules or statistics to derive sentence-level or message-level polarities. Efforts were made to discover words/phrases specific to the finance domain [31, 48]. A great amount of *learning-based systems* were later developed. Specifically, two benchmark tasks (SemEval 2017 Task 5 [9] and FiQA 2018 Task 1 [10]) were conducted, and the best results were achieved by regression ensemble (RE), convolutional neural network (CNN), and support vector regression (SVR) models based on combined features of sentiment lexica and dense word representations. The *following wave* of designs (approx. 2019 till now) was based on fine-tuning general-purpose pre-trained language models. For example, BERT[1] (Bidirectional Encoder Representations from Transformers) was fine-tuned as FinBERT [30] and achieved better FSA results. The state-of-the-art results using fine-tuning is from integrating multiple auxiliary knowledge sources to a BERT variant [15]. Although medium-sized LLMs (e.g., Pythia-1.4B and OPT-1.3B) can also be fine-tuned like BERT to achieve comparable performances to larger LLMs [24], fine-tuning the largest LLMs requires formidable time and monetary costs. Therefore, it is important to research how to leverage larger LLMs for FSA. In terms of leveraging LLMs for FSA, the current progress mainly employed the encoder type of transformer, e.g., BERT. However, the most powerful LLMs now are based on the decoder part of a transformer. The decoder architecture is natural for generative tasks such as discourse/chat completion and question answering, but can also be fitted for discriminative tasks and classification. This study is aware of the early stage and scant in-depth studies in this direction and thus explores ways of leveraging generative LLMs for FSA.

Different from many ad hoc designs developed from chain of thought (CoT) [13], tree of thoughts (ToT) [50], verification, self-consistency constraints, intermediate scratchpads, and multi-agent multi-role settings, the design framework presented here follows the design science guidelines by Hevner et al. [23] and contributes to the prescriptive knowledge as a "design theory" [20]. Based on Minsky's theory of mind and emotions, "emotional states" are our "Ways to Think" with a specific collection of resources turned on and others turned off given certain environment conditions [34]. Therefore, one FSA approach is to simulate the mental processes underlying the texts, requiring specialized LLM agents to play the roles of "resources", i.e., functional parts of our brain that make us react to the environment. In the context of financial analysis, the resources can either be linguistic knowledge or more advanced learned professional knowledge that is not innate parts of our brains. The design framework (**H**eterogeneous multi-**A**gent **D**iscussion) chooses to develop specialized LLM agents by prompting. The agents' main functions are either to pay attention to types of linguistic errors that LLMs are prone to make for the given FSA task or to think like an institutional/individual investor. The design artifact thus has seven (5+2) different agents. The final FSA result is based on a shared discussion considering outputs from all the agents. This design artifact is evaluated using multiple methods, and the results generally conclude the framework to be effective.

The major challenge in instantiating this design is the lack of design theory on what each agent's function should be. For this reason, many LLM multi-agent settings employ homogeneous agents. For example in the multi-agent debate framework, Du et al. [18] simply disseminate the same input to multiple LLM agents. Because of some randomness and perturbation, each agent's response will not be identical. Later each agent will take outputs from other agents (excluding its own output) as additional information to update its original response (Fig. 1 (a)). It may go through multiple rounds though empirical results show that consensus will be achieved fast. Another framework is to assign different roles to LLM agents. Sun et al. [42] described a negotiation procedure where a "discriminator LLM" is asked to judge whether it agrees with the output of a "generator LLM". The judgment statement is sent back to the generator if consensus is not made. The framework requires a third LLM to negotiate and vote for the final result if discrepancies persist (Fig. 1 (b)). Although the LLM agents in this

[1]There is no strict definition of "how large" a language model has to be to qualify for the name of LLM. It seems that LLMs are usually far larger than the word2vec models (around 1 million parameters). In this article, language models with > 100 M parameters are referred to as LLMs. This definition includes BERT (110-340 M parameters), Mistral (7 B parameters), GPT-3.5 (around 175 B parameters), GPT-4 (around 1760 B parameters), and more.

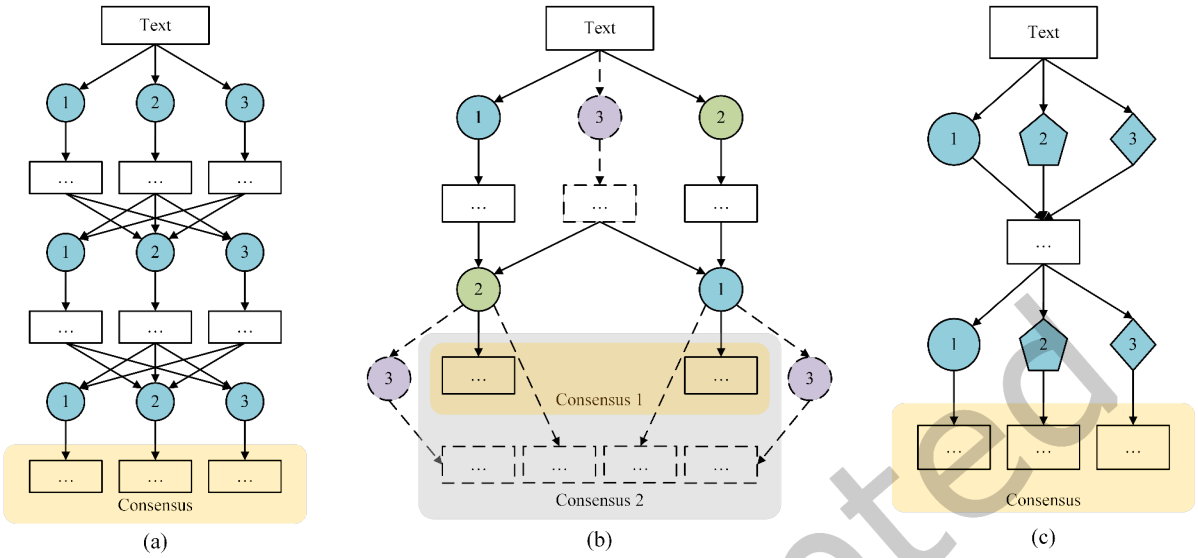

Fig. 1. Different multi-agent LLM frameworks for reaching a consensus: (a) homogeneous multi-agent debate [18], (b) multi-role multi-agent negotiation [42], (c) heterogeneous multi-agent discussion (HAD: the proposed framework). Colors denote different roles and shapes denote heterogeneous agents.

framework play different roles, their capability assumptions remain the same. In such a sense, these agents are still non-specialized and are homogeneous. In the proposed framework (Fig. 1 (c)), each agent has the same role and goes through a symmetric discussion workflow (unlike [42]), but is purposely designed to simulate the mental functions of different resources. Their responses are aggregated for FSA just like resources are activated to generate different emotional states in Minsky's theory.

Therefore, one objective of this study is to test whether linguistic error types for FSA [47, 54] and domain knowledge of investor types in finance [2] can be a useful guideline for developing heterogeneous agents. Specifically, The following research questions are investigated:

- RQ1: How effective is HAD compared to naive prompting and the fine-tuning paradigm?
- RQ2: How to prompt LLM agents to behave heterogeneously for sentiment analysis in finance?
- RQ3: What are the quantified contributions of each LLM agent and their relative importance?

To address these questions, HAD is evaluated using multiple methods including empirical analysis of performance metrics on six FSA datasets, ablation analysis with different sets of agents, and case studies of outputs and intermediary representations. The experimental results show that HAD can in general improve the FSA performance. The improvements are more significant than other benchmark methods (MD, MSV, HSV) and are more pronounced when more advanced base LLM agents are used. It has been observed that a simple template "please pay special attention to [error type]" can change LLM agents' attention and prompt them to behave differently. Mood, rhetoric, and reference agents as well as institutional/individual agents seem to be the main performance drivers and are more critical than other LLM agents, though the contributions are non-linear and have complicated interactions.

This study contributes to the design science literature by presenting an AI kernel theory-informed design artifact. Many kernel theories from the natural or social sciences were introduced to information system design,

whereas kernel theories from AI are comparatively rare. This study has implications for emotion theory, LLM collaboration research, and financial decision-making practices. Firstly, it supports the society of mind and emotion machines [34] to be actionable theories that explain how emotions emerge as an important type of human intelligence; Secondly, this study applies multi-agent LLMs in FSA. This framework has been used for factuality checking, arithmetic/mathematical reasoning, optimization, and general-purpose sentiment analysis, but not yet on FSA to the best of my knowledge. This study thus provides new materials for LLM collaboration, and also reinforces the design science-based approach to framework development; Lastly, the findings contribute to the prescriptive knowledge of FSA system design. Investors and traders may iterate and improve their own FSA systems based on the HAD framework or be more informed when they decide to select or purchase technical solutions of a similar kind.

2 Related Work and Design Process

In this section, related literature is organized into two lines: (any type of) use of LLMs for FSA, and ways of prompt design (not limited to FSA). After providing the literature, The theoretical foundations of employing heterogeneous agents for FSA are elaborated in Sec. 2.3 and 2.4.

2.1 Using LLMs for Financial Sentiment Analysis

Financial sentiment analysis (FSA) is a domain-specific business-oriented application closely related to the general natural language processing task of sentiment analysis [17]. Because of its heavy use of terminologies and other linguistic features [39, 47], general sentiment analysis performances are usually not representative and will drop in the finance domain. Due to its complexity and the requirement for intricate reasoning in the absence of domain-specific data, FSA has been incorporated to thoroughly evaluate LLM capabilities, along with tasks such as Named-Entity Recognition (NER), knowledge recall, question answering, and reading comprehension, among others [39, 44].

In terms of using a singular LLM for FSA, the task is sometimes formulated together with auxiliary tasks, such as target and aspect detection [15]. Target refers to an entity, and aspect refers to an attribute most directly associated with the sentiment expressed. This additional information (targets and aspects) may be used to improve FSA performances. For example, Lengkeek et al. [27] used the hierarchical structure of aspect systems to constrain FSA results, though this information is rarely available in real-world production environments. Zhang et al. [52] observed that financial news is often overly succinct. A model that retrieves additional context from reliable external sources to form a more detailed instruction is consequently developed. Deng et al. [13] found that forcing the LLM through several reasoning paths with CoT helps generate more stable and accurate labels. The LLM-generated labels are also useful and meet the quality requirement of complementing human annotations for conventional supervised learning methods. Similarly, Fei et al. [19] developed a three-hop reasoning framework inspired by CoT that infers firstly the implicit aspect, secondly the implicit opinion, and finally the sentiment polarity. However, it has been pointed out [42] that a singular LLM has difficulties in fully exploiting the potential of LLM knowledge. This is especially true for FSA as it involves multiple LLM capabilities, such as reasoning [16], fact-checking [18], syntactic/semantic parsing, and more. A similar phenomenon as reported in [53] is observed that LLM performances on more complicated tasks are not as satisfactory as on the binary classification task. Moreover, the aforementioned designs (storage retrieval and CoT) and more designs that are not yet applied to FSA, such as verification, self-consistency constraints, or intermediate scratchpads, are also largely heuristic, at most based on experiences, and lack a solid theoretical foundation.

The proposed framework, unlike those using a singular LLM, adopts in-context learning (ICL) and leverages multiple LLM instantiations (agents), which is also referred to as LLM collaboration. Strategies of collaboration include auxiliary tasks (e.g., verification) [7], debate [18], and various role-assignment [42] including generator,

discriminator, programmer, manager, meta-controller, etc. Again, the design of auxiliary tasks and roles appears arbitrary and lacks solid theoretical foundations. LLM collaboration is also more investigated on many general natural language processing tasks including sentiment analysis, but their applicability on FSA lacks direct evidence. One of the most comparable endeavors in terms of using LLM collaboration for an application (medical, financial) domain to the proposed HAD design framework is MedPrompt [36]. However, because it uses an ensemble of randomly shuffled CoT from homogeneous agents, the MedPrompt design is more computationally heavy and difficult to transfer to the finance domain as existing financial question-answering datasets are more sparse. Topologically, the proposed HAD design framework resembles Multi-agent Debate [18], despite the apparent difference that HAD uses theory-inspired heterogeneous agents.

2.2 Prompt Engineering

An important question to HAD is to decide how to (or whether it is possible to) create heterogeneous agents simply using different prompts. Before the emergence of generative LLMs, a well-accepted way of applying a language model to downstream tasks is through fine-tuning: remove the last neural network layer (referred to as the "head" layer) and let the training errors back-propagate with the bottom layers parameters fixed. Two major problems with it are: (1) a not-too-small training set and labels are still needed, and (2) the training process can be computationally intensive. With the observation that generative LLMs are very powerful, in-context learning (ICL) contends it is possible to get the desired output without fine-tuning and elicit the model capability with an appropriate "prompt". Typically, prompt engineering involves the development of task-specific prompt templates, which describe how a prompt should be formulated to enable the pre-trained model to perform the downstream task at hand. Liu et al. [28] provided a survey on recent advances in prompt engineering and systematically compared major prompt shapes like cloze prompts and prefix prompts. HAD uses prefix prompts because agents extract specific information as answers.

Prompt templates can be automatically searched for using stochastic optimization-based methods. Sorensen et al. [41], for example, discovered that a good template is one that maximizes the mutual information between input and the generated output. But more often, prompt templates in application domains are manually designed. For example, Liu and Chilton [29] studied text-to-image generative models and the prompt template "[SUBJECT] in the style of [STYLE]". They found the clarity and salience of keywords are important to the generation quality. Yu et al. [51] presented the idea of using domain knowledge to guide prompt design. It was reported that for the legal information entailment task, the best results are obtained when prompts are derived from specific legal reasoning techniques, such as Issue-Rule-Application-Conclusion (IRAC) as taught at law schools. For FSA, however, the design guidelines are unclear and most studies used naive prompts. For example, Chen and Xing [6] used "You are a helpful sentiment analysis assistant - [EXAMPLE MESSAGE]:[SENTIMENT]. User: [TEST MESSAGE]." and BloombergGPT's FSA template [44] is simply "[TEST MESSAGE] Question: what is the sentiment? Answer with negative/neutral/positive." For the proposed HAD agents, "Consider [MESSAGE], what is the sentiment [AGENT-SPECIFIC INSTRUCTION]" is used.

2.3 Kernel Theory: Emotions and the Society of Mind

Kernel theory is a key component of the information system design process according to Walls' information system design theory (ISDT). It explains how/why the anticipated system would work and sheds light on the meta-requirements. In the context of FSA, the theory has to be one that explains the formative mechanism of emotion. For this reason, Minsky's theory of mind and emotions is preferred over other descriptive/contrastive theories of emotions, such as Plutchik's wheel of emotions or Russell's circumplex model.

Society of mind is a reductionistic perspective of human intelligence that influenced AI greatly and argues no function directly produces intelligence. Instead, intelligence comes from the managed interaction of a variety of

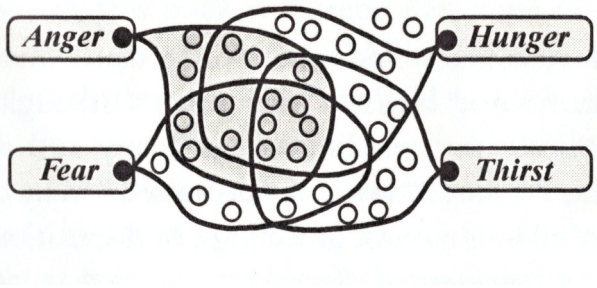

Fig. 2. Illustration of the generation of emotional states from activating a collection of resources, cf. pg. 4 in [34].

resourceful but simpler and non-intelligent agents. For example, when drinking a cup of tea, there activates a motor agent that grasps the cup, a balancer that keeps the tea from spreading, and a temperature sensor that confirms our throat will not be hurt. This theory sees emotional states as patterns of activation. For example, the state we call “angry" could be what happens when a cloud of resources that help you react with unusual speed and strength are activated — while some other resources that make you act prudently are suppressed (Fig. 2).

Minsky’s theory of emotion posits that you feel “angry" when your cake is stolen by other kids, because the IF-THEN-DO rules activate resources to help you take it back. The activation is adaptive as we learn and grow. For FSA, a crucial procedure is to decide what candidate resources need to be designed: it will not require the full set of resources in our brain which will be more challenging to build. In the remainder of this section, the design rationales will be described using a kernel theory-based design science framework (Table 1).

Table 1. Kernel Theory-Based Design: A Meta-Framework

Kernel theory	“Society of mind” and “Emotion machines”. The theories posit that emotions come from activation of different resources.
Meta-requirements	1. To simulate the resources, we should define heterogeneous agents and their functions. 2. To activate the agents, we should provide information about the subjectivity. 3. To achieve a well-informed decision, we should aggregate info. from different agents.
Meta-designs	1. Types of error are used as linguistic knowledge (domain language style) and types of investor used as finance domain knowledge to guide building heterogeneous capabilities. 2. The user message is distributed to each LLM agent. 3. Specialized agent outputs are concatenated to form the summative prompt.
Testable hypotheses	Evaluate the effectiveness of the metadesigns. Specific testable hypotheses are as follows: **H1**: The HAD framework can improve the accuracy of existing naive prompts for FSA. **H2**: The agents have different importance but all contribute positively to the analysis.

2.4 Meta-requirements, Meta-designs, and Hypotheses

Although the society of mind relies heavily on the conceptual construct of “resource", it is purposefully kept in a hazy way according to Minsky (pg. 25 in [34]), referring to all sorts of functional parts that range from perception and action to reflective thinking. Therefore, it seems appropriate to simulate the resources using LLM agents with

polymathic capabilities, and specialize their functions via prompts. This way of simulating resources also enables 'activation', in a sense that specialized agents generate meaningful responses (i.e., activate) only when the input text contains relevant information. It is thus designed such that all the LLM agents will receive the original user message just like resources react to the same stimuli. To aggregate information, a widely used technique is to concatenate them into a longer prompt [18, 22, 28, 42]. By translating the meta-requirements into more detailed meta-designs, the HAD framework can be formally represented as:

(1) Define *heterogeneous* agents and their prompt templates A_1, A_2, ..., A_k.
(2) Obtain intermediary analysis $O_i = A_i(User\_Message)$
(3) Obtain summative analysis $Result = A(User\_Message, O_1, ..., O_k)$

where function $A_i()$ is an acronym for "Agent". The agents are intended to produce natural text answers.

The second step can be carried out for multi-rounds before sending for a summary depending on the consensus situation, though in the evaluation sections the results are reported on single-round only. An illustration of the workflow is presented in Fig. 3. This framework is universally applicable to not only FSA but any decision process. To instantiate the agents for a specific task (e.g., for FSA), one has to decide the guiding knowledge based on understanding of the task.

Noteworthy, Minsky's theories' most important influence on the design is the heterogeneity of agents. Further, they provide some hints/constraints on the agent design principles: those should not be any principle that engineers the system to work, but mimic functional parts of the human brain because the human brain is the very place that sentiment and emotions emerge from. Solely relying on the multi-agent design literature, one is more likely to end up with homogeneous multi-agents, and would not quest for the guiding knowledge.

For FSA, there are two main sources of guiding knowledge available, i.e., the linguistic knowledge of how information is communicated in finance and the domain (expert) knowledge of how investors tend/learned to think. In the following Section, how the guiding knowledge informed the design of specialized agents will be elaborated. To assess whether the proposed framework is effective, two testable hypotheses are developed. If the guiding knowledge and agent design are appropriate, we would expect the performance metrics to improve (**H1**). Because of the noted data imbalance issue in FSA, F-1 score should be more credible on top of accuracy when evaluating performance. Another observation is that the occurrences of each type of linguistic error vary across different language domains [47, 54], and the performances of different investors are not equal. It is thus hypothesized that the agents will have different importance but all contribute positively to the FSA task (**H2**).

3 Design Artifact: Heterogeneous Agent Discussion (HAD)

To instantiate a design artifact based on the HAD framework, the number of agents (k) has to be decided. We would prefer a not too large k for the sake of comprehensibility and computational efficiency.

For linguistic knowledge, Zimbra et al. [54] had investigated a comprehensive list of Twitter sentiment analysis methods and concluded three major challenges: (1) language brevity, (2) imbalanced classes, and (3) temporal dependency. Because of these challenges, 13 categories of commonly occurring classification errors were identified. The main categories that ground to linguistic features can be summarized as: (1) humor, (2) subtlety or a mixture of sentiment, (3) irrelevance (e.g., aspect mismatch), (4) marketing information mistaken as positive, and (5) atypical contextual usage. In paralell, Xing et al. [47] investigated the common errors in a slightly different scope: specifically for FSA and including text sources other than Twitter[2]. The 6 categories of errors identified, i.e., (1) irrealis mood, (2) rhetoric, (3) dependent opinion, (4) unspecified aspects, (5) unrecognized words, and (6) external reference, have significant overlap with those reported from [54].

With this background, five agents are designed based on [47] because (1) these categories are more directly FSA relevant and (2) these categories are less in number (6 compared to 13) and more operational. Since LLMs are

[2]Now has been re-branded as "X".

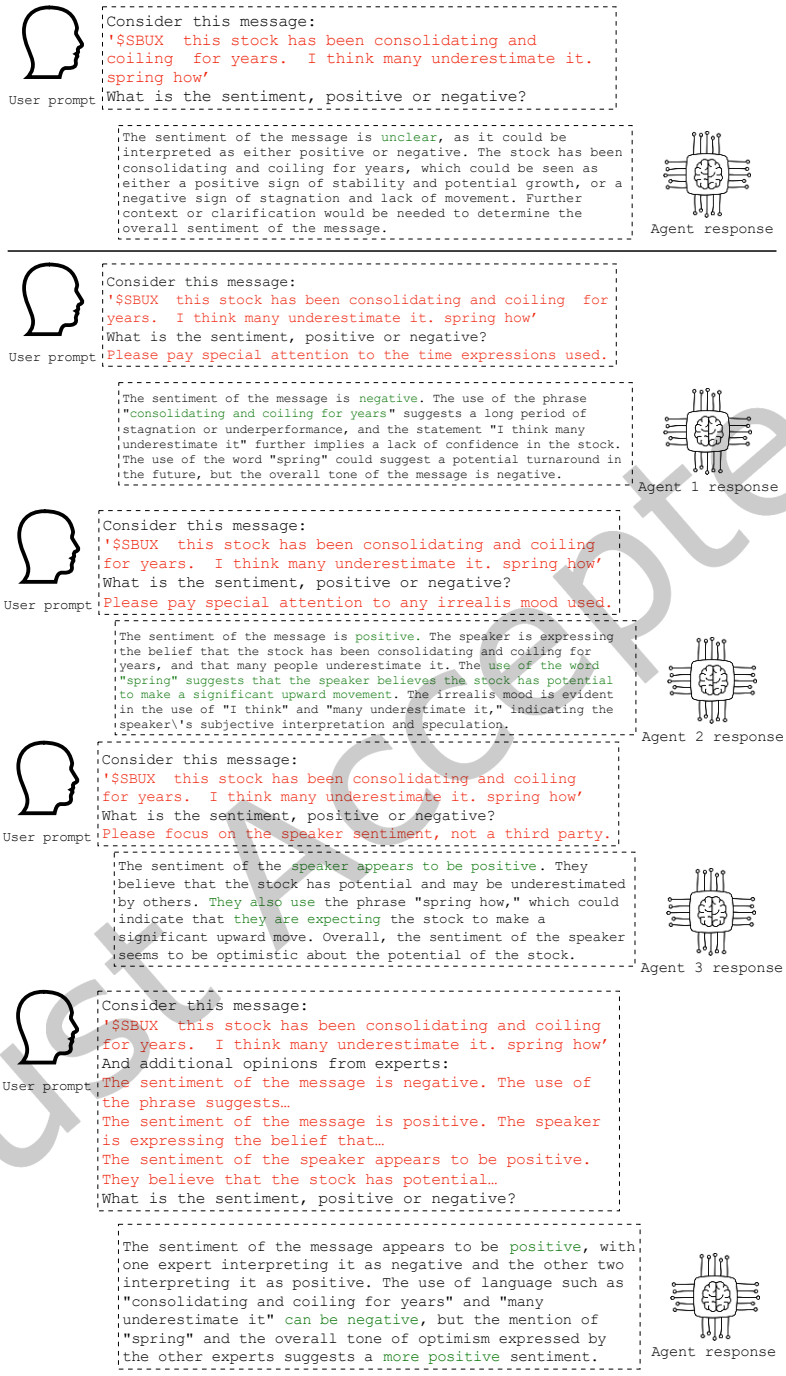

Fig. 3. An illustrative comparison between naive prompting (the upper example) and the proposed HAD framework (the lower example) with heterogeneous agents inspired by FSA error types. The illustration only shows 3 out of 7 specialized agents due to space limit.

observed to be robust to unrecognized words and spellings from the web, no special agent is designed according to this error. The five agents and their characteristic prompts are:

- A1 (mood agent): Please pay special attention to any irrealis mood used.
- A2 (rhetoric agent): Please pay special attention to any rhetorics (sarcasm, negative assertion, etc.) used.
- A3 (dependency agent): Please focus on the speaker sentiment, not a third party.
- A4 (aspect agent): Please focus on the stock ticker/tag/topic, not other entities.
- A5 (reference agent): Please pay special attention to the time expressions, prices, and other unsaid facts.

For domain knowledge, the most widely documented dichotomy is "institutional versus individual investors". Finance literature posits that both institutional/individual investors may act irrationally or exhibit behavioral biases, but their differences are quite evidential and robust. Barber and Odean [2] concluded that compared to institutional investors, average individual investors are poorer in terms of long-term performance, but are strong in short-term horizons. They trade more actively and have patterns to their wealth and lifecycle, but are less-informed, under-diversified, more influenced by media, have less energy/attention, and have stronger disposition effects. Therefore, two additional agents are designed and their characteristic prompts are:

- A6 (institutional agent): Consider like an institutional investor, focusing on long-term, fundamental effects.
- A7 (individual agent): Consider like an individual investor, focusing on price changes and technical indicators.

Finally, the summative prompt $A(\cdot)$ takes the form of "Considering this message from [SOURCE]: [TEST MESSAGE], and additional opinions from experts [OPINIONS], what is the sentiment, positive/negative/neutral?". Some nuances are adjusted according to whether the testbed classification is binary or ternary.

4 Evaluation

Hevner et al. [23] described five kinds of design evaluation methods, i.e., analytical, case study, experimental, field study, and simulation. Since experiments and field studies are not currently applicable, this study leverages the rest three out of them: (1) simulated empirical testing on existing datasets and the produced performance metrics, (2) ablation analysis with manipulated module components, and (3) observational evaluation based on case studies. The following Sections 4.2 to 4.4 present the evaluation results.

4.1 Data and Base Models

The proposed design framework is evaluated on six existing datasets spanning types from financial news to social media post and forum discussions, i.e., the Financial PhraseBank [33], StockSen [47], CMC [6], FiQA Task 1 [32], SEntFiN 1.0 [40], and FinEntity [43]. The last three are finer-grained financial sentiment analysis datasets with sentiment intensity scores or multiple targets/entities labels, though quantization and filtering have been applied to fit the evaluations into a consistent classification problem. For example, the original FiQA dataset [32] has 1173 messages with sentiment scores ranging from -1 to +1. By filtering those scores with an absolute value larger than 0.3, only 771 messages are left and mapped to the positive/negative classes. The detailed statistics of the processed datasets are reported in Table 2. In terms of text genre, Financial PhraseBank (FPB) is from news and SEntFiN is from news headlines. StockSen and CMC are from social media (StockTwits and CoinMarketCap.com respectively), the whole FiQA is consolidated from crawling a mix of StackExchange, Reddit, and StockTwits. Finally, FinEntity contains financial news from Refinitiv. These datasets do not cover formal documents and earning calls because those data are less applicable for FSA and are hardly directly labeled in high quality.

Currently, there are many base LLMs available include GPT, LlaMa, Claude, Mistral, Gemini, BLOOMZ, and they can be classified into two types, i.e., open-access and restricted-access. The HAD framework is tested on three

Table 2. Summary statistics of the six FSA datasets (post-processing)

Dataset	FPB	StockSen	CMC	FiQA	SEntFiN	FinEntity
Positive	570	4542	12022	507	2832	503
Negative	303	1676	1523	264	2373	498
Neutral	1391	–	–	–	2701	1130
Total Size	2264	6218	13545	771	7906	2131

instruction-finetuned language models: GPT[3] (-3.5 Turbo and -4o) as a commercial restrict-access representative; BLOOMZ[4] (the 560 M version [35]) and LlaMa3[5] (the 70 B version) as open-access representatives. This way, LLMs of different sizes are covered. BLOOMZ runs locally on a laptop with an 8-core Apple M1 chip and 16 GB memory. The LlaMa3 model results are obtained from two cloud host & inference services Deepinfra[6] and Groq[7]. GPT (-3.5 Turbo and -4o) results are obtained through OpenAI API. For this reason, BLOOMZ-560m inference was the slowest despite being the smallest LLM. Because of time constraint, some BLOOMZ performance metrics are also approximated or not available in Table 3.

For the performance metrics, one experiment (one LLM base model on one dataset) takes hours to days execution time. The performance metrics are reported in Table 3. To calculate the metrics, unclear or irrelavant final output such as "mixed sentiment", "...cannot determine...", "...provide more information..." are replaced with a random insertion from positive/negative/neutral. Those random insertions are around 0.1% - 1% of the total outputs and are unlikely to affect the major conclusions. For ternary classifications (FPB, SEntFiN, and FinEntity), macro F-1 scores are used. Some metrics (in grey color) of BloombergGPT [44] and (Fin-)BERT [6, 15, 39, 40, 47] are included to help roughly assess the gaps to fine-tuning based results. Noteworthy, these metrics are cited from other studies and BloombergGPT is a proprietary model, so the metrics may be obtained from different evaluation settings (e.g., 3/5-classes or data splits different from reported here) and are not precisely comparable.

4.2 Performance Analysis

The HAD framework is benchmarked to several other multi-agent settings include MSV, MD, and HSV.

(1) MSV (Multi-agent, homogeneous, Simple Voting): The agents cannot see each other's response and vote for the final output based on majority.
(2) MD (Multi-agent, homogeneous, Debate): Similar to the proposed model in [18], all models are given the same prompts and agents can see each other's response.
(3) HSV (multi-agent, Heterogeneous, Simple Voting): The agents are given different prompts but do not communicate. Instead of a summative agent, the final output are based on majority voting.

Formulaically, the agents for MSV are k instances of the same **g**eneral template A_1^g, A_2^g, ..., A_k^g. The MSV result is the majority in $\{A_1^g(User\_Message), A_2^g(User\_Message), ..., A_k^g(User\_Message)\}$, so the summative agent is not needed;

For MD, the intermediary analysis output $O_i^g = A_i^g(User\_Message, A_{\neq i}^g(User\_Message))$. The MD result is produced by the summative agent $A(User\_Message, O_1^g, O_2^g, ..., O_k^g)$;

[3]https://platform.openai.com/docs/models
[4]https://huggingface.co/bigscience/bloomz
[5]https://llama.meta.com/llama3
[6]https://deepinfra.com/Meta-Llama-3-70B-Instruct
[7]https://groq.com/Llama3-70b-8192

Table 3. Effects of instantiating the HAD design framework using different base LLMs (benchmarked)

Model\Dataset	FPB		StockSen		CMC		FiQA		SEntFiN		FinEntity	
	Acc.	F-1	Acc.	F-1	Acc.	F-1	Acc.	F-1	Acc.	F-1	Acc.	F-1
L&M Dictionary [31]	69.89	54.40	70.70	81.70	–	–	–	–	–	–	–	–
(Fin-)BERT	91.69	89.70	76.90	84.50	93.50	–	–	–	94.29	93.27	–	83.00
BloombergGPT	–	51.07	–	–	–	–	–	75.07	–	–	–	–
BLOOMZ-560m	34.63	32.90	63.65	72.47	87.16	92.62	78.33	83.64	**51.32**	**41.87**	32.43	26.90
BLOOMZ-560m (MSV)	29.57	30.13	65.46	74.41	–	–	77.87	82.83	–	–	32.57	27.98
BLOOMZ-560m (MD)	34.25	33.97	68.90	79.46	–	–	**78.80**	**84.21**	–	–	32.71	28.37
BLOOMZ-560m (HSV)	33.11	29.01	**69.09**	**83.47**	–	–	78.50	83.85	–	–	**46.36**	22.96
BLOOMZ-560m (HAD)	**34.89**	**38.40**	68.06	78.44	**87.67**	**92.95**	77.42	83.31	50.16	40.69	34.69	**32.93**
LlaMa3-70b	45.49	50.00	70.78	71.79	78.25	64.73	84.65	81.44	58.34	52.39	43.21	43.78
LlaMa3-70b (MSV)	42.23	46.87	75.26	**84.88**	**81.47**	**88.76**	85.73	89.38	57.70	50.36	42.75	42.78
LlaMa3-70b (MD)	46.27	51.03	76.35	73.61	77.05	63.00	84.96	84.55	55.36	47.77	45.83	44.69
LlaMa3-70b (HSV)	40.85	40.80	**77.67**	80.41	79.02	87.39	78.34	84.91	51.73	40.66	40.75	38.22
LlaMa3-70b (HAD)	**68.48**	**84.31**	74.25	76.12	71.91	82.06	**92.09**	**94.01**	**63.51**	**66.72**	**60.23**	**59.36**
GPT-3.5	78.58	81.06	67.64	73.93	85.31	91.05	90.53	92.41	67.99	63.21	55.84	56.00
GPT-3.5 (MSV)	72.22	74.68	68.62	76.08	86.31	92.25	87.81	90.29	74.06	64.62	**68.46**	**68.61**
GPT-3.5 (MD)	69.83	73.92	69.40	75.72	86.74	91.47	87.95	90.86	69.65	69.30	56.51	57.29
GPT-3.5 (HSV)	65.12	64.95	69.35	76.03	84.04	90.44	87.42	90.05	40.56	64.96	42.37	40.07
GPT-3.5 (HAD)	**81.25**	**87.10**	**70.01**	**77.97**	**90.91**	**92.69**	**95.07**	**96.20**	**78.16**	**77.72**	61.80	62.56

For HSV, the final output is the majority in $\{O_1, O_2, ..., O_k\}$. Noteworthy, these intermediary analysis O_i are without superscript $\cdot^g$, i.e., they are from specialized agents as in HAD.

The "single LLM + self-reflection" results are not reported because on FPB, StockSen, and FiQA, the results are only comparable or worse than the naive use of a single LLM. Therefore, it does not seem a meaningful benchmark and experiments are aborted early.

The first observation from Table 3 is on the perceivable effects of different base model choices. BLOOMZ was trained on a very large Open-science Open-collaboration Text Sources corpus [26], which is mainly crowd-sourced scientific datasets. Llama3 was trained on "over 15T tokens that were all collected from publicly available sources". GPT-3.5 was trained mainly on the Common Crawl corpus [4], which archives the web. The six testing datasets are all crawled from the web: which may be closer to the training language domain of Llama3 and GPT-3.5. Another possible source of general performance differences is the LLM sizes, where Llama3 (70 B) is between BLOOMZ (560 M) and GPT-3.5 (175 B). Behavior-wise, it is observed that GPT-3.5 is better instruction-tuned with its proprietary human feedback. In contrast, BLOOMZ inclines to the language completion task. An example is that the prompt "Translate to English: Je t'aime" without a full stop (.) at the end may result in the model trying to continue the French sentence instead of translating it. BLOOMZ also inclines to complete/answer with concise language. For the sentiment-related open-ended questions to heterogeneous agents, BLOOMZ often answers a final judgment of positive/negative without much justification, and is not good at predicting "neutral" messages. The behavior of Llama3 is between BLOOMZ and GPT-3.5, but much closer to GPT-3.5. Performance-wise, probably for the afore-discussed factors, there are steady increases as larger base models are used. For example, the ranges of accuracies and F-1 scores increase on FinEntity from (BLOOMZ: 32-46%, 23-33%) to (Llama3: 41-60%, 38-59%) and (GPT-3.5: 42-68%, 40-69%). The performance differences are more pronounced for FPB, SEntFiN, and FinEntity, which contain neutral classes.

The second observation is that HAD generally improves the accuracies and F-1 scores on the base models (Table 3). In the 18 experiments, HAD achieved the best performance 12 times, whereas MSV, HSV, MD, and naive prompting only achieved the best performance for 3, 1, 1, and 1 time respectively. If compared to naive prompting, MD is 12 times better, whereas MSV and HSV are 9 and 6 times better in 16 experiments. This means that MD is generally more effective than naive prompting; MSV is generally not effective; and HSV can make the performance worse (below 8 in 16). The author suspects this is because specialized agents are biased, judging only from one perspective. Thus they are worse classifiers to be used directly. It can be concluded that voting is not a good technique for integrating multi-agents' responses. Finally, HAD's improvements became more consistent on Llama3 and GPT-3.5, probably due to the richer intermediary analysis generated. HAD's improvement on BLOOMZ is also confident for naive prompting, despite the differences being minimal from other multi-agent benchmarks. Noteworthy, despite StockSen being the dataset on which the linguistic error types for agent design are derived, this knowledge generalizes well on other independently developed datasets.

The last observation is on assessing the significance of the improvements. Theoretically, fine-tuning the LLMs to a downstream task will perform better than the ICL/instruction-based/zero-shot setting just as in the differences of supervised/unsupervised learning. The cost of fine-tuning is bi-fold in the context of FSA: you have to ask experts to accumulate and label thousands of examples; and the performance will be fragile to data distribution shifts and dependent on the optimization techniques applied. Therefore, one cares about to what extent HAD closes the gap between ICL and fine-tuning. For example, if the gap between GPT-3.5 and (Fin)-BERT measured by F-1 score is 83.00% - 56.00% = 27.00%, and HAD's improvement is 62.56% - 56.00% = 6.56%, the fix is 6.56% ÷ 27% = 24.30%. By comparing the improvements to the overall differences between naive prompting and (Fin)-BERT on FPB, StockSen, CMC, SEntFiN, and FinEntity, a fair estimation is that the HAD framework can fix 20%–50% of the gap between ICL and fine-tuning.

4.2.1 GPT-3.5 versus GPT-4o. The landscape of LLMs is changing rapidly due to new models and version upgrades. For example, recently Bard is re-named Gemini from 2024 February, and Llama3 supersedes Llama2 as the latest version from 2024 April. A reasonable question to ask is whether the upgrades of the same base model will significantly impact the overall performance or the effectiveness of HAD. To address this question, the performances using gpt-4o-2024-05-13 and gpt-3.5-turbo-1106 are compared on two relatively small datasets, i.e., FPB and FiQA (Table 4). It can be observed that solely upgrading to a higher version LLM does not necessarily improve the performance. In fact, the naive prompting results all dropped for GPT-4o. However, the GPT-4 (HAD) results all improved from GPT-3.5 (HAD) results, making the effectiveness of HAD more pronounced on GPT-4o than on GPT-3.5. In sum, version upgrading is unlikely to impact the effectiveness of HAD.

4.3 Ablation Analysis

To test the importance of each LLM agent, their intermediary responses are removed singly and the performance increases/decreases benchmarked on GPT-3.5 (HAD-5), i.e., w/o (A6+A7), are reported in Table 5. Because of time constraints, only three relatively small datasets and their average results are used: FPB, FiQA, and SEntFiN.

It is observed that the institutional (A6) and individual agent (A7) are very important because HAD-7 results are constantly higher than HAD-5. In terms of linguistic knowledge, the mood agent (A1), the rhetoric agent (A2), and the aspect agent (A4) are the most important: removing any of them will generally have a negative impact on the performance. The reference agent (A5) is less important: the effect of removing it is uncertain across different datasets. The dependency agent (A3) seems ineffective: removing A3 will further improve the performance. The ineffectiveness of A3 may suggest considering this error type is unnecessary, or be attributed to an ineffective prompt design. Either way, the observed performances suggest that heterogeneous agents have complicated non-linear interactions, and the presented design can be further optimized with more empirical

Table 4. Performance analysis on base model upgrade: GPT-3.5 Turbo versus GPT-4o

Model\Dataset	FPB		FiQA	
	Acc.	F-1	Acc.	F-1
GPT-3.5	78.58	81.06	90.53	92.41
GPT-3.5 (MSV)	72.22	74.68	87.81	90.29
GPT-3.5 (MD)	69.83	73.92	87.95	90.86
GPT-3.5 (HSV)	65.12	64.95	87.42	90.05
GPT-3.5 (HAD)	**81.25**	**87.10**	**95.07**	**96.20**
GPT-4	74.57 (↓ 4.01)	77.99 (↓ 3.07)	87.68 (↓ 2.85)	86.52 (↓ 5.89)
GPT-4 (MSV)	72.84 (↑ 0.62)	76.20 (↑ 1.52)	88.33 (↑ 0.52)	91.01 (↑ 0.72)
GPT-4 (MD)	71.03 (↑ 1.20)	74.13 (↑ 0.21)	85.08 (↓ 2.87)	83.20 (↓ 7.66)
GPT-4 (HSV)	63.04 (↓ 2.08)	66.12 (↑ 1.17)	79.64 (↓ 7.78)	86.04 (↓ 4.01)
GPT-4 (HAD)	**88.34** (↑ 7.09)	**88.62** (↑ 1.52)	**95.73** (↑ 0.66)	**96.71** (↑ 0.51)

Table 5. Effects of removing specialized agents on performance metrics (using gpt-3.5-turbo-1106 as the base model)

Model\Dataset	FPB		FiQA		SEntFiN		Average	
	Acc.	F-1	Acc.	F-1	Acc.	F-1	Acc.	F-1
HAD-7	+0.77	+5.69	+1.16	+0.98	+0.71	+0.79	+0.88	+2.49
HAD-5 (w/o A6,7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HAD-4 (w/o A1,6,7)	-0.71	+0.64	-0.01	+0.02	-0.58	-0.61	-0.43	+0.02
HAD-4 (w/o A2,6,7)	-2.12	-0.39	+0.64	+0.52	-0.80	-0.99	-0.76	-0.29
HAD-4 (w/o A3,6,7)	+3.00	+3.56	+0.51	+0.42	+0.01	+0.03	+1.17	+1.34
HAD-4 (w/o A4,6,7)	+0.04	+0.97	+0.25	+0.22	-0.66	-0.69	-0.12	+0.16
HAD-4 (w/o A5,6,7)	+4.32	+4.29	-0.01	-0.00	-0.52	-0.43	+1.26	+1.28
GPT-3.5	-1.90	-0.35	-3.38	-2.81	-9.46	-13.72	-4.91	-5.63

evidence. Noteworthy, the relative agent importance is preliminary results on GPT-3.5 and the generalizability to other LLMs needs further investigation.

4.4 Case Study

Five cases are presented in Table 6 to illustrate the quality of HAD outputs and how those outputs predict a polarity different from naive prompting.

In Case 1, multiple companies are mentioned and naive prompting produces a negative prediction without much explanation. With HAD, A1 and A2 believe this message is neutral according to their perspectives. A1's argument is reasonable as the positivity is more directly associated to Wells Fargo than to Berkshire. With A3 to A7 all considering the message as positive, the framework finally summarizes a correct polarity as positive. Noteworthy, A6 uses background knowledge, i.e., "Berkshire is a renowned/reputable company" to support the confidence of its prediction.

Case 2 is challenging and can easily be mistaken as positive by naive prompting with key-phrase such as "drive ... higher" spotted. To correctly understand the context, one has to know that Taylor Wimpey and Ashtead are

home construction and construction equipment rental companies. So “driving the markets higher" may refer to the index or property markets and is setting an economic scenario. It has complicated implications for the two companies and is not as direct as “Barclays falls". A1, A4, A5, A7 are correct about the mixed sentiment. Interestingly, A6 disagrees with A7 because they are instructed to focus on long-term/short-term implications respectively. With diverse predictions ranging from positive, neutral, negative, and mixed, the framework finally summarizes a correct polarity as negative.

Case 3 is predicted as positive by naive prompting. Although as A1 explained, less smuggling is good for society, the message is apparently commenting on gold itself as a commodity. Despite the fact that no irrealis mood or any rhetorics are present, A1, A2, A5, A7 correctly predict the message as negative. With A3 and A4 seeing it as neutral and A6 seeing it as positive, the final decision is negative as per the majority.

Case 4 is difficult to understand even for professionals unfamiliar to India’s context and is predicted as positive by naive prompting. Jargon and external reference are the main challenges. From the responses of A1, A2, and A5, it can be observed that “Puravankara is a real estate company", “CIS means Collective Investment Scheme" and “Sebi is a security regulatory authority" are shared background knowledge. Surprisingly, A2 exhibits temporal and counterfactual reasoning, which is helpful in understanding this message.

The last Case 5 was wrongly predicted as positive by naive prompting, probably due to the slight positive color of phrasing “head to". A1 and A2 are critical and correctly identified the uncertainty associated with irrealis mood. In fact, Whirlpool’s price was in the range of 75-250 in the past 5 years and has never been or likely to reach 450-475. A3, A6, A7 detected the same positivity as naive prompting, while the other four agents all predict the message as neutral. With the dominant number of neutral predictions (4:3), the framework correctly summarized the polarity as neutral. This shows HAD’s capability to correct slight and uncertain sentiments with a discussion mechanism.

Table 6. Cases where HAD predicts a correct polarity label different from naive prompting (agent analysis abridged)

Case No. & Text	Agent Analysis	Agent Stmt.	HAD Stmt.	True Label
(**Case 1**) Berkshire applies to boost Wells Fargo stake above 10 pct	(**A1**) No indication of positive or negative sentiment. The use of "applies to" suggests an irrealis mood, means Berkshire is seeking to increase its stake but the action has not yet been completed.	Neutral	Positive	Positive
	(**A2**) No indications of sarcasm or negative assertion in the message.	Neutral		
	(**A3**) The sentiment of the speaker in this message is likely positive, as they are expressing their intention to increase their stake, which suggests confidence.	Positive		
	(**A4**) The sentiment is positive for "Wells Fargo" as Berkshire Hathaway's application suggests confidence and optimism in the company's potential.	Positive		
	(**A5**) Generally positive. Berkshire Hathaway's decision potentially have a positive impact on Wells Fargo's stock price and overall reputation. The increase timing suggests potential for long-term success despite any current challenges facing Wells Fargo.	Positive		
	(**A6**) Berkshire Hathaway, a renowned investment firm, applying to boost its stake in Wells Fargo. This could indicate that Berkshire believes in the fundamental strength and potential growth of Wells Fargo. This news could be viewed as a vote of confidence in the bank's leadership, business strategy, and overall performance over the long term.	Positive		
	(**A7**) This news may lead to increased demand for Wells Fargo's stock, potentially driving up its price and triggering positive technical indicators, a bullish signal.	Positive		
(**Case 2**) London open: Taylor Wimpey and Ashtead drive markets higher, Barclays falls	(**A1**) Drive markets higher suggests a positive sentiment, while Barclays falls indicates a negative sentiment. No use of irrealis mood.	Mixed	Negative	Negative
	(**A2**) It highlights the fact that Barclays falls, while giving less emphasis to the positive performance of Taylor Wimpey and Ashtead. No specific rhetorics used.	Negative		
	(**A3**) They are simply reporting on the market movements and the performance of certain companies.	Neutral		
	(**A4**) Positive for Taylor Wimpey and Ashtead as they are driving the markets higher, while the sentiment is negative for Barclays as it is falling.	Mixed		
	(**A5**) Some companies are performing well and driving the market higher, others, such as Barclays, are not faring as well.	Mixed		

Case No. & Text	Agent Analysis	Agent Stmt.	HAD Stmt.	True Label
	(**A6**) Positive towards Taylor Wimpey and Ashtead driving markets higher. This indicates strength and growth in the housing and construction sectors, in the long-term. On the other hand, the falling of Barclays may raise concerns as it indicates weak market stability in the financial sector. Overall, sentiment would be cautiously optimistic.	Positive		
	(**A7**) Positive for Taylor Wimpey and Ashtead, but negative for Barclays as it is falling. As an individual investor focusing on price changes and technical indicators, there are both buying and selling opportunities. Overall, the sentiment would be mixed.	Mixed		
(**Case 3**) Smuggling of gold sees a decline as its demand softens	(**A1**) The sentiment is negative. Word “decline" indicates a decrease in smuggling activity, which could be seen as positive, but the softening demand for gold implies a negative impact on the overall market for gold. There is no irrealis mood used in this message.	Negative	Negative	Negative
	(**A2**) The sentiment is negative. Decline and soften indicate a decrease in the demand for gold, which could have negative implications for those involved in the smuggling of gold. No specific rhetorical devices present.	Negative		
	(**A3**) Speaker is neutral.	Neutral		
	(**A4**) Neutral.	Neutral		
	(**A5**) The sentiment is negative. The decrease in the demand indicates a downturn in the market, leading to a decline in smuggling activities. This suggests a weakening gold market, which could impact prices and trade within the industry.	Negative		
	(**A6**) Cautious optimism. The decline in smuggling suggests a decrease in illegal market activity, which means a more stable and transparent gold market. However, the softening demand may also signal broader trends and shifts in investor sentiment towards alternative investments. Long-term implications need to be monitored.	Positive		
	(**A7**) Negative. The softening demand could potentially lead to a decrease in gold prices. This is concerning for an investor considering buying gold, especially if they rely on technical indicators. It may also signal broader economic concerns. Overall, the sentiment is cautious and potentially bearish.	Negative		

Case No. & Text	Agent Analysis	Agent Stmt.	HAD Stmt.	True Label
(**Case 4**) Puravankara's real estate scheme not a CIS: Sebi	(**A1**) The sentiment is negative. The use of the word "not" indicates the negation of a positive outcome, and the mention of Sebi suggests that there may be regulatory or legal issues associated with Puravankara's real estate scheme. There does not appear to be any irrealis mood used in this message.	Negative	Negative	Negative
	(**A2**) "not a CIS" may imply suspicion that Puravankara's real estate scheme was a CIS (Collective Investment Scheme), which is regulated by SEBI (Securities and Exchange Board of India). The use of the term "scheme" also carries a negative connotation, as it can imply a dubious or deceptive plan. Overall, the message communicates a negative view of Puravankara's real estate activities in relation to SEBI regulations.	Negative		
	(**A3**) Neutral. The statement just delivers factual information.	Neutral		
	(**A4**) Neutral.	Neutral		
	(**A5**) The sentiment appears to be negative, indicated by the mention of Sebi not considering Puravankara's real estate scheme as a Collective Investment Scheme (CIS). There may be some regulatory or compliance issues with the scheme, which could have negative implications for the company and its investors. Additionally, the fact that the scheme is being scrutinized by a regulatory authority implies a potential risk or concern. However, without further details about the specific nature of the scheme or the reasons for Sebi's decision, it is difficult to fully assess the sentiment.	Negative		
	(**A6**) Cautious and skeptical. The statement suggests uncertainty or controversy surrounding the real estate scheme and its classification. This could raise red flags for institutional investors who seek stability and transparency in their investments.	Negative		
	(**A7**) Negative. The statement suggests uncertainty or lack of clarity about Puravankaraś real estate scheme. This uncertainty could lead to negative price changes and could negatively impact technical indicators for the company.	Negative		
(**Case 5**) Whirlpool may head to around 450-475: Devang Visaria	(**A1**) The sentiment is neutral. The use of "may" indicates an irrealis mood, suggesting that the statement is hypothetical or uncertain.	Neutral	Neutral	Neutral
	(**A2**) There are no apparent rhetorics or emotional language. The statement simply presents a speculated range for the potential movement of Whirlpool.	Neutral		
	(**A3**) Sentiment of the speaker seems to be somewhat positive as they are mentioning a potential increase in the stock price of Whirlpool, indicating a sense of optimism or anticipation.	Positive		

Case No. & Text	Agent Analysis	Agent Stmt.	HAD Stmt.	True Label
	(**A4**) Neutral.	Neutral		
	(**A5**) The sentiment is neutral. It simply provides a prediction regarding the potential price movement without expressing positive or negative. It is based on analysis and does not convey any emotion or bias.	Neutral		
	(**A6**) Positive. The mention of Whirlpool potentially heading to around 450-475 suggests a favorable increase for long-term investment. The mention of "Devang Visaria" indicates that this forecast is coming from a credible source, further supporting a positive sentiment.	Positive		
	(**A7**) The message suggests that the Whirlpool stock may head to a price range of 450-475, indicating a upward trend. This indicates potential gains in the stock price.	Positive		

5 Discussion

Using the evaluations and analysis in Section 4, the answers to the research questions and hypotheses can be discussed. The remainder of this section provides more details on how to understand the contributions of this research in a broader IS context, implications for academics and practitioners, and notable limitations in interpreting the results.

In view of the research questions and hypotheses, it has been found that HAD effectively improves FSA accuracies across a number of existing datasets and base LLMs (**RQ1**). When more advanced LLMs are used (BLOOMZ-560m to LlaMa3-70b to GPT-3.5), it has been observed that the LLM agents produce more substantial and meaningful discussions [25], and the improvements in accuracies and F-1s become more consistent and stable. For example, Table 3 shows that applying HAD improves FSA results 4 out of 6 times when BLOOMZ-560m is used, and 6 out of 6 times when GPT-3.5 is used. As the size and capability of LLMs are fast advancing, the performance benefits of HAD are likely to become more noticeable and predictable. If the (Fin-)BERT and BloombergGPT results are considered as rough upper bounds, e.g., (91.69, 89.70) on FPB, the performance gap for GPT-3.5 (78.58, 81.06) is around ten percent. Therefore, HAD (81.25, 87.10) closes around half of the gap. In average, the proposed designed framework fixes ca. 20% – 50% of the performance gap between prompting and fine-tuning.

Knowledge-based prompting is proven to be a successful strategy (**RQ2**). With guiding knowledge from both linguistics and finance, the specialized LLM agents behave heterogeneously with different focuses, as evidenced in the case studies. Performance-wise, homogeneous agents-based framework (MSV and MD, where no knowledge is needed) ranked top only 3 and 1 time in 18 experiments, whereas HAD ranked top 12 times in 18 experiments. However, the summative agent is crucial. The specialized agents are biased, therefore worse in voting. The "collective wisdom" is only achieved when discussion happens between specialized agents.

In terms of quantified contributions of agents (**RQ3**), ablation analysis and Table 5 show that institutional and individual agents are very important, leading to constant improvements from HAD-5 to HAD-7; the mood, rhetoric, and aspect agents are more important than the reference agent. In summary, {A6, A7} > {A1, A2, A4} > {A3, A5}. A possible explanation may be that the reference capability has been well internalized in GPT-3.5. Noteworthy, the agents' relative importance is a purely GPT-based observation because of the limited experiment scale. In general, it seems that the agents are especially useful when the base LLM has the potential capability but does not emphasize or exhibit such capability in free-text generation. The evaluation results support Hypothesis 1 but reject Hypothesis 2 with the observation that the performance can be further optimized if the dependency agent is removed.

5.1 Research Contributions

This study's contribution is mainly technical and can be understood in relation to the Knowledge Contribution Framework (KCF) of deep learning [38] and the Text Analytics Information Systems Research (TAISR) Framework [1]. In terms of KCF, a prompt and agents based deep learning framework (HAD) has been formulated and executed in a new application domain of FSA, creating a new way of integrating deep learning in a broader, higher-impact system, where fine-tuning is still a dominant paradigm and LLM collaboration is rarely applied. The framework is zero-shot, training-free, and provides useful explanations [5]. Therefore, the performance improvement should be able to generalize to other FSA datasets. This article also instantiated the design into an AI artifact, such that the framework's effectiveness can be tested. In terms of TAISR, this study provides a new case of conducting text analytics using LLMs and shows the versatility of translating a research objective into a certain task type. Normally, FSA is considered as a classification task. This study shows that the generative capability of LLMs can be used to better serve the research objective when a different task type is chosen.

5.2 Implications for Research and Practice

Multi-agent LLMs represent a significant trend in leveraging state-of-the-art AI capabilities for decision-making tasks, offering new opportunities for both research and practical applications. In addition to measuring public opinion in financial news, social media, and product reviews [1] faster and more accurately, this study has research and practical implications in connection to the broader understanding of emotional theory, operationalization of LLM collaboration frameworks, and how practitioners could customize their in-house sentiment analysis tools, or extend their knowledge of possible system designs for FSA and similar tasks.

Firstly, the HAD framework supports Minsky's theories of mind and emotions and leads us to reflect on how sentiment in financial contexts emerges. Traditional sentiment analysis methods often rely on simplistic positive or negative classifications, which may overlook the complexity of investor reasoning and emotions. By utilizing multiple LLMs, each trained to recognize different emotional nuances, researchers can capture a more detailed and accurate picture of market sentiment. For example, one LLM might identify underlying anxiety in investor communications, while another detects cautious optimism. The multi-faceted emotional analysis enables the development of sophisticated models that can better predict market sentiment.

Secondly, the dynamics between multiple LLMs open new avenues for research in LLM collaboration. As different LLMs bring unique perspectives and strengths to sentiment analysis, studying their interactions can lead to further optimized performance. Research can focus on how LLMs can complement each other, manage conflicts in their interpretations, and integrate their insights into a coherent and comprehensive analysis. Such advancements in LLM collaboration not only enhance financial sentiment analysis but also contribute to AI and NLP research in general.

Finally and practically, financial advisors, traders, fund managers, and other types of investors could use this framework to build their own FSA tools with their expertise in the data and market conditions. For example, a financial advisor may create an agent that knows the wealth information and retirement plans of a customer to constrain interpretations of certain sentiments; a fund manager may know better the linguistic features in his/her theme and update the linguistic agents accordingly. The seven agents listed in this article can be a foundation list for practitioners to adapt.

5.3 Limitations and Future Work

This study has a few limitations, which may inspire future research. The first limitation is *scalability*. Predicting or Discussion with LLM agents is slower compared to statistical analysis and incurs costs. For this reason, a large system, i.e., with more agents, is possible, but not always plausible during design and evaluation. The second limitation is the *confidentiality* of evaluation datasets. StockSen, CMC, SEntFiN, and FinEntity are relatively new, but FPB and FiQA have been there for quite a few years. Because the training material for LLMs is usually not fully transparent and some LLMs keep updating using reinforcement learning and human feedback[8], the possibility that the evaluation datasets have been exposed to the LLMs before, causing some information leaks can not be excluded. Finally, the case studies show that the identified error types can almost be solved. It is therefore interesting to explore what are the reasons for the new errors made by LLMs and assess what are the human/expert-level performances on these FSA datasets. This will require a larger scale case study like conducted in [54] and [47].

6 Conclusion

A novel theory-informed LLM collaboration design for FSA, named HAD, is studied. Unlike many state-of-the-art LLM collaboration designs that instantiate homogeneous agents [18, 42], HAD involves heterogeneous LLM agents and specializes them with knowledge from both linguistics and finance. This knowledge, i.e., FSA error

[8]As of now (Aug 2024), there are several versions of GPT-4 available, using training data up to Sep 2021 or Dec 2023.

types and different cognitive patterns in investing, is discovered from the past literature [2, 16, 47]. This design is more computationally intensive than naive prompting, but has far less complexity compared to many other LLM collaboration designs and fine-tuning-based approaches.

Some of the unique challenges in FSA, e.g., external references to facts and world knowledge, were thought to be impossible to solve in the short-term future before the transformer architecture models came into existence in 2019. With the hope of artificial general intelligence (AGI) around the corner, this study exhibits the versatile capabilities of LLM that are useful for FSA, and calls for more research on this important task.

Acknowledgments

The author is thankful for the detailed comments and suggestions from the anonymous reviewers and the journal editorial team, and funding support from Singapore MOE's AcRF Grant 251RES2107.

References

[1] Benjamin Ampel, Chi-Heng Yang, James Hu, and Hsinchun Chen. 2024. Large Language Models for Conducting Advanced Text Analytics Information Systems Research. *ACM Transactions on Management Information Systems* (2024). https://doi.org/10.1145/3682069 Forthcoming.

[2] Brad M. Barber and Terrance Odean. 2013. *The Behavior of Individual Investors.* Elsevier, 1533–1570. https://doi.org/10.1016/b978-0-44-459406-8.00022-6

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of NeuIPS'20.* 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

[5] Mirko Cesarini, Lorenzo Malandri, Filippo Pallucchini, Andrea Seveso, and Frank Xing. 2024. Explainable AI for Text Classification: Lessons from a Comprehensive Evaluation of Post Hoc Methods. *Cognitive Computation* (2024), 1–19. https://doi.org/10.1007/s12559-024-10325-w

[6] Siyi Chen and Frank Xing. 2023. Understanding Emojis for Financial Sentiment Analysis. In *Proceedings of ICIS'23.* 1–16. https://aisel.aisnet.org/icis2023/socmedia_digcollab/socmedia_digcollab/3/

[7] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching Large Language Models to Self-Debug. https://doi.org/10.48550/ARXIV.2304.05128

[8] Liya Chu, Xue-Zhong He, Kai Li, and Jun Tu. 2022. Investor Sentiment and Paradigm Shifts in Equity Return Forecasting. *Management Science* 68, 6 (2022), 4301–4325. https://doi.org/10.1287/mnsc.2020.3834

[9] Keith Cortis, Andre Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *SemEval Workshop.* https://doi.org/10.18653/v1/S17-2089

[10] Dayan de França Costa and Nadia Felix Felipe da Silva. 2018. INF-UFG at FiQA 2018 Task 1: predicting sentiments and aspects on financial tweets and news headlines. In *Companion Proceedings of WWW'18.* 1967–1971. https://doi.org/10.1145/3184558.3191828

[11] Jiaying Deng, Mingwen Yang, Matthias Pelster, and Yong Tan. 2023. Social Trading, Communication, and Networks. *Information Systems Research* (2023). https://doi.org/10.1287/isre.2021.0143

[12] Shuyuan Deng, Zhijian (James) Huang, Atish P. Sinha, and Huimin Zhao. 2018. The Interaction Between Microblog Sentiment and Stock Returns: An Empirical Examination. *MIS Quarterly* 42, 3 (2018), 895–918. https://doi.org/10.25300/misq/2018/14268

[13] Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. In *Companion Proceedings of WWW'23.* https://doi.org/10.1145/3543873.3587324

[14] Wei Dong, Shaoyi Liao, and Zhongju Zhang. 2018. Leveraging Financial Social Media Data for Corporate Fraud Detection. *Journal of Management Information Systems* 35, 2 (2018), 461–487. https://doi.org/10.1080/07421222.2018.1451954

[15] Kelvin Du, Frank Xing, and Erik Cambria. 2023. Incorporating Multiple Knowledge Sources for Targeted Aspect-based Financial Sentiment Analysis. *ACM Transactions on Management Information Systems* 14, 3 (2023), 1–24. https://doi.org/10.1145/3580480

[16] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. An Evaluation of Reasoning Capabilities of Large Language Models in Financial Sentiment Analysis. In *Proceedings of IEEE CAI'24.* https://doi.org/10.1109/CAI59869.2024.00042

[17] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial Sentiment Analysis: Techniques and Applications. *Comput. Surveys* 56, 9 (2024). https://doi.org/10.1145/3649451

[18] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of ICML'24*. https://doi.org/10.48550/ARXIV.2305.14325

[19] Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In *Proceedings of ACL'23*. 1171–1182. https://doi.org/10.18653/v1/2023.acl-short.101

[20] Shirley Gregor and Alan R. Hevner. 2013. Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly* 37, 2 (2013), 337–355. https://doi.org/10.25300/misq/2013/37.2.01

[21] Terrence Hendershott, Xiaoquan (Michael) Zhang, J. Leon Zhao, and Zhiqiang (Eric) Zheng. 2021. FinTech as a Game Changer: Overview of Research Frontiers. *Information Systems Research* 32, 1 (2021), 1–17. https://doi.org/10.1287/isre.2021.0997

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of ICLR'21*. https://doi.org/10.48550/arXiv.2009.03300

[23] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design Science in Information Systems Research. *MIS Quarterly* 28, 1 (2004), 75–105. https://doi.org/10.2307/25148625

[24] Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. 2023. Large Language Model Adaptation for Financial Sentiment Analysis. In *Proceedings of FinNLP'23*. https://doi.org/10.18653/v1/2023.finnlp-2.1

[25] Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment Analysis in the Age of Generative AI. *Customer Needs and Solutions* 11, 1 (2024). https://doi.org/10.1007/s40547-024-00143-4

[26] Hugo Laurençon, Lucile Saulnier, Thomas Wang, and alia. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Proceedings of NeurIPS'22*. https://doi.org/10.48550/arXiv.2303.03915

[27] Matteo Lengkeek, Finn van der Knaap, and Flavius Frasincar. 2023. Leveraging hierarchical language models for aspect-based sentiment analysis on financial data. *Information Processing & Management* 60, 5 (2023), 103435. https://doi.org/10.1016/j.ipm.2023.103435

[28] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Comput. Surveys* 55, 9 (2023), 1–35. https://doi.org/10.1145/3560815

[29] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of CHI '22*. https://doi.org/10.1145/3491102.3501825

[30] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. In *Proceedings of IJCAI'20*. 4513–4519. https://doi.org/10.24963/ijcai.2020/622

[31] Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66, 1 (2011), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

[32] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: financial opinion mining and question answering. In *Proceedings of WWW'18*. 1941–1942. https://doi.org/10.1145/3184558.3192301

[33] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65, 4 (2014), 782–796. https://doi.org/10.1002/asi.23062

[34] Marvin Minsky. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster.

[35] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of ACL'23*. 15991–16111. https://doi.org/10.18653/v1/2023.acl-long.891

[36] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. https://doi.org/10.48550/ARXIV.2311.16452

[37] Richard L. Peterson. 2016. *Trading on Sentiment: The Power of Minds Over Markets*. Wiley. https://doi.org/10.1002/9781119219149

[38] Sagar Samtani, Hongyi Zhu, Balaji Padmanabhan, Yidong Chai, Hsinchun Chen, and Jay F. Nunamaker. 2023. Deep Learning for Information Systems Research. *Journal of Management Information Systems* 40, 1 (2023), 271–301. https://doi.org/10.1080/07421222.2023.2172772

[39] Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain. In *Proceedings of EMNLP'22*. https://doi.org/10.18653/v1/2022.emnlp-main.148

[40] Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology* 73, 9 (2022), 1314–1335. https://doi.org/10.1002/asi.24634
[41] Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In *Proceedings of ACL'22*. https://doi.org/10.18653/v1/2022.acl-long.60
[42] Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment Analysis through LLM Negotiations. https://doi.org/10.48550/ARXIV.2311.01876
[43] Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. FinEntity: Entity-level Sentiment Classification for Financial Texts. In *Proceedings of EMNLP'23*. 15465–15471. https://doi.org/10.18653/v1/2023.emnlp-main.956
[44] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. https://doi.org/10.48550/ARXIV.2303.17564
[45] Frank Xing, Erik Cambria, and Roy Welsch. 2018. Intelligent Asset Allocation via Market Sentiment Views. *IEEE Computational Intelligence Magazine* 13, 4 (2018), 25–34. https://doi.org/10.1109/mci.2018.2866727
[46] Frank Xing, Erik Cambria, and Yue Zhang. 2019. Sentiment-aware volatility forecasting. *Knowledge Based Systems* 176 (2019), 68–76. https://doi.org/10.1016/j.knosys.2019.03.029
[47] Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets. In *Proceedings of COLING'20*. 978–987. https://doi.org/10.18653/v1/2020.coling-main.85
[48] Frank Xing, Filippo Pallucchini, and Erik Cambria. 2019. Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management* 56, 3 (2019), 554–564. https://doi.org/10.1016/j.ipm.2018.11.002
[49] Yi Yang, Yu Qin, Yangyang Fan, and Zhongju Zhang. 2023. Unlocking the Power of Voice for Financial Risk Prediction: A Theory-Driven Deep Learning Design Approach. *MIS Quarterly* 47, 1 (2023), 63–96. https://doi.org/10.25300/misq/2022/17062
[50] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Proceedings of NeuIPS'23*. 1–14.
[51] Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks. In *Findings of the ACL'23*. https://doi.org/10.18653/v1/2023.findings-acl.858
[52] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In *Proceedings of ICAIF'23*. https://doi.org/10.1145/3604237.3626866
[53] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the NAACL'24*. https://doi.org/10.18653/v1/2024.findings-naacl.246
[54] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Transactions on Management Information Systems* 9, 2 (2018), 5:1–5:29. https://doi.org/10.1145/3185045

Received 23 December 2023; revised 6 August 2024; accepted 8 August 2024