Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI, IoT, Robotic Engineering & Algorithm
DOI: 10.54254/2755-2721/2026.LD35648

A Survey of Zero-Shot and Few-Shot Learning with Large Language Models for Financial Sentiment Analysis

Liyang Gao

School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China
gaoliyang0318@Gmail.com

Abstract. Financial sentiment analysis has long relied on labeled data to fine-tune models like FinBERT, a process that is both costly and time-consuming. The arrival of large language models (LLMs) has changed the landscape: with zero-shot and few-shot prompting, one can now extract sentiment from financial texts using few or no annotated examples. This survey takes stock of how LLMs are being applied to this task. It begins by clarifying the core ideas behind in-context learning and chain-of-thought prompting. It then examines a range of prompt designs that have been developed to cope with the peculiarities of financial writing, such as numerical expressions, implicit sentiment and long documents. A comparison of LLM performance on standard benchmarks against fine-tuned domain models shows that general-purpose LLMs are often competitive, especially when prompts are carefully crafted. Yet three problems remain unresolved: numerical reasoning errors, hallucination, and the practical hurdles of cost, latency, and privacy. These challenges are discussed in detail, and retrieval-augmented generation, trustworthiness frameworks, and efficient open-source models are pointed out as the most promising paths forward. Overall, LLMs offer a flexible and annotation-light alternative to traditional fine-tuning, but their successful deployment in finance will depend on robust prompt engineering and solid factual grounding.

Keywords: large language models, financial sentiment analysis, zero-shot learning, few-shot learning, prompt engineering

1. Introduction

Financial sentiment analysis is about gauging the tone of earnings reports, news articles, and social media posts, information that feeds directly into trading and risk management. Historically, practitioners have turned to lexicon-based methods and, more recently, to fine-tuned models like FinBERT [1]. These approaches perform well but need large amounts of labeled financial text, which is expensive to create. The arrival of large language models (LLMs) has opened a different door. Thanks to their zero-shot and few-shot capabilities, it is now possible to get decent sentiment signals without any task-specific training data [2, 3]. This survey takes a close look at how that potential is being realized and where the remaining bottlenecks lie.

DOI: 10.54254/2755-2721/2026.LD35648

Recent studies have already demonstrated that LLMs such as GPT-4 can hold their own on financial sentiment benchmarks without being explicitly fine-tuned [4]. Researchers have also developed increasingly sophisticated prompting techniques that push accuracy higher. Nevertheless, three persistent challenges remain unresolved. First, LLMs often struggle with numerical reasoning [5], misreading the implications of profit margins and growth figures. Second, they are prone to hallucination [6], which means they may confidently produce fake numbers or quotes that would be disastrous in a financial context. Third, using commercial LLMs at scale raises serious questions about cost, latency, and data privacy [3]. These challenges form the backbone of this review.

The focus of this survey is on general-purpose LLMs doing financial sentiment analysis with zero-shot and few-shot learning. The rest of the paper is organized as follows. Section 2 lays out the basic concepts. Section 3 reviews the art of prompt engineering for financial texts. Section 4 presents a quantitative comparison on standard benchmarks. Section 5 digs into the main challenges. Section 6 suggests where the field might go from here and wraps up the discussion.

2. Foundations of zero-shot and few-shot learning

In a zero-shot setup, the model receives only a natural language instruction with no labeled examples. Few-shot, or in-context, learning adds a handful of demonstrations right before the query [7]. Chain-of-thought (CoT) prompting goes a step further by including reasoning steps inside those demonstrations [8]. For financial sentiment, a zero-shot prompt simply asks the model to classify a text as positive, negative, or neutral. A few-shot variant prepends one to three labeled headlines. A CoT variant walks the model through identifying key numbers before inferring the overall mood. One surprising finding is that even a generic instruction like "Let's think step by step" can trigger surprisingly good zero-shot reasoning in large models [9], which suggests that a lot of reasoning capability is already embedded in these models and just needs the right trigger.

The clearest reason is that the pre-training corpora already contain substantial financial text, exposing the model to domain-specific vocabulary and sentiment patterns [7]. This is somewhat similar to how a well-read person can understand a new domain without formal training. Beyond passive exposure, large-scale models also exhibit emergent reasoning abilities that allow them to interpret nuanced language—for instance, recognizing that revenue declined but beat expectations conveys mixed rather than purely negative sentiment [10], a distinction that simpler models often miss. Instruction tuning further reinforces this by teaching models to follow task descriptions faithfully, effectively converting a generic completion engine into a sentiment analyzer without any financial supervision.

Empirically, Lopez-Lira and Tang [11] showed that ChatGPT's zero-shot sentiment scores derived from news headlines can predict subsequent stock returns, which is strong evidence that general LLMs can pick up economically meaningful sentiment. Projects like FinGPT [12] have taken this a step further by using instruction tuning to adapt open-source LLMs specifically for finance, demonstrating that the gap between general and financial models can be narrowed substantially with relatively lightweight adaptation.

3. Prompt engineering for financial sentiment

3.1. Prompt design strategies

Getting sentiment right from an LLM often comes down to the wording of the prompt. Common strategies include plain classification instructions, role-playing (e.g., acting as a senior financial

DOI: 10.54254/2755-2721/2026.LD35648

analyst), CoT prompting combined with self-consistency voting to improve robustness [13], and structured outputs such as JSON. Empirical studies show that when using few-shot examples, picking the right demonstrations matters more than how many are included. Retrieving examples by semantic similarity typically beats random selection [14], which makes sense because similar examples give the model a better template to work from. FinGPT [12] provides a collection of prompt templates that blend several of these techniques and can lift zero-shot accuracy by a noticeable margin. The choice of strategy often depends on the specific task and the model being used, so practitioners usually need to experiment to find what works best.

3.2. Handling financial text challenges

Financial texts are tricky. They are dense with numbers, often rely on context from specific companies, and can run very long. Prompt design can help, but it cannot eliminate all errors. For number-heavy passages, explicit cues like reminding the model to distinguish between profit increased by 30% and profit increased to 30% can reduce errors. When sentiment hinges on entity context—rising costs might be bad for a car manufacturer but good for a steel supplier—injecting the target company's name and industry into the prompt makes a difference. Long documents, such as transcripts of earnings calls, benefit from a two-stage approach: first, extract sentiment from each section, then aggregate the results. Even so, the FinQA benchmark [5] reminds us that current prompting cannot fully solve deep numerical reasoning problems, and this remains an active area of research.

4. Performance and comparison

4.1. Evaluation on FiQA, financial phrasebank and other datasets

Two datasets dominate the evaluation landscape: Financial PhraseBank [15] for sentence-level sentiment and FiQA [16] for aspect-based opinion mining. Li et al. [17] provide a thorough comparison of ChatGPT and GPT-4 against fine-tuned models on both benchmarks. On PhraseBank, GPT-4 with zero-shot prompting achieves an F1 score of 83%, and with 5-shot learning it reaches 86%, which is on par with the fine-tuned FinBERT baseline at 84% [17]. On FiQA, which involves aspect-based sentiment analysis, GPT-4 zero-shot attains a weighted F1 of 87.15%, and 5-shot learning pushes it to 88.11%, outperforming the fine-tuned RoBERTa baseline at 87.09% [17]. These results demonstrate that general LLMs, without any fine-tuning, consistently outperform lexicon-based baselines and frequently match or exceed the performance of domain-specific models like BloombergGPT [18].

A key observation is that GPT-4 already closes most of the gap in the zero-shot setting: on PhraseBank it trails FinBERT by only one percentage point (83% vs. 84%). With five-shot learning, it actually pulls ahead (86% vs. 84%). On FiQA, the pattern is similar. This difference is likely because FiQA involves aspect-based sentiment, which requires the model to identify specific aspects and assess sentiment toward each one, a task that is inherently more complex than simple sentence-level classification. The fact that LLMs can approach or match fine-tuned models on these benchmarks is impressive, but it is also worth keeping in mind that these are controlled settings with clean data. Real-world financial texts are often noisier and more ambiguous.

DOI: 10.54254/2755-2721/2026.LD35648

4.2. Comparison with fine-tuned models

When placed next to fine-tuned FinBERT [1] or domain-specific models such as BloombergGPT [18], LLMs entail clear trade-offs. Where labeled data is scarce, zero-shot LLMs substantially outperform fine-tuned models trained on a handful of examples. Where labeled data is plentiful, a carefully tuned FinBERT can still match a general LLM on sentence-level tasks, though the gap narrows considerably with five-shot prompting [17]. This suggests that in-domain fine-tuning remains a valuable tool, particularly for complex aspect-based sentiment analysis. The real advantage of LLMs lies elsewhere: flexibility. They can switch from three-way classification to aspect-based sentiment without retraining, while a fine-tuned model would need a new training cycle. So, rather than being in competition, the two approaches complement each other in practice. LLMs shine when annotations are rare or task requirements change often; fine-tuned models remain attractive for stable, high-volume production pipelines.

Yet this flexibility comes at a cost: prompt-based systems are sensitive to wording changes, and their performance can drop sharply when the input distribution shifts, a robustness issue that receives less attention than raw accuracy. Some of the reported gains, moreover, may stem from benchmark contamination—the models may have seen PhraseBank sentences during pre-training, a possibility that Li et al. [17] acknowledge but do not fully rule out. This uncertainty makes the robustness of zero-shot and few-shot methods in fully unseen settings an open question.

5. Challenges and limitations

5.1. Numerical reasoning errors

The close performance between GPT-4 and fine-tuned models reported in Section 4.1 should be interpreted with caution: the controlled benchmark setting does not reflect the numerical reasoning and hallucination risks discussed in Sections 5.1 and 5.2. LLMs can misread financial figures in ways that a human analyst would not. Consider a model that reads "net profit fell 30% year-on-year" but classifies it as positive because the surrounding text emphasizes future growth—the numerical decline is effectively overridden by contextual narrative. The FinQA dataset [5], though originally designed for question answering rather than sentiment analysis, exposes a limitation that is directly relevant to financial sentiment tasks where models must interpret profit figures and growth rates. Even the best models at the time of their release struggled to achieve high accuracy on financial numerical reasoning, and general-purpose LLMs face similar or greater difficulties. At core, LLMs are pattern matchers rather than calculators, a limitation that becomes especially costly in finance. While techniques like chain-of-thought prompting can help somewhat, they do not fully solve the problem, especially when the reasoning involves multi-step calculations or comparisons across multiple time periods.

5.2. Hallucination and reliability

Hallucination, which means inventing facts, is dangerous anywhere, but especially in finance. An LLM might fabricate a specific stock price or a bogus analyst quote and deliver it with complete confidence. Such errors could lead to wrong trades and big losses. Ji et al. [6] offer a comprehensive taxonomy of hallucination phenomena in natural language generation. Work like DecodingTrust [19] further reveals that adversarial inputs can easily push GPT models into generating biased or false outputs, a vulnerability that is equally worrying in financial contexts, underscoring the brittleness of

DOI: 10.54254/2755-2721/2026.LD35648

current systems. This matters because financial decisions often involve large sums of money, where even small errors carry significant consequences.

5.3. Deployment constraints

Putting LLMs into real financial workflows is not just a matter of accuracy. The cost of calling commercial APIs for high-frequency news streams adds up quickly. Inference latency makes these models unsuitable for time-sensitive trading, where decisions need to be made in milliseconds. And sending confidential documents to an external cloud service is often incompatible with data protection rules, especially in regulated industries like banking. These hard constraints explain the growing interest in open-source, locally deployable models like LLaMA [20] and toolkits like FinGPT [12], which promise better privacy and lower running costs, even if they require more technical expertise to set up.

6. Future directions and conclusion

6.1. Improving domain reasoning

To tackle reasoning and hallucination problems, retrieval-augmented generation (RAG) is gaining traction [21, 22]. The idea is to dynamically pull in relevant financial reports or news articles during inference, grounding the model's output in verifiable sources. Marrying RAG with financial knowledge graphs could further strengthen entity-aware reasoning. And as FinGPT [12] has shown, continued instruction tuning on curated financial data can permanently sharpen a model's domain reasoning. Progress probably requires combining several of these approaches rather than relying on any single one.

6.2. Toward trustworthy financial LLMs

Trust in an LLM's output will require more than just high accuracy. Models need to express uncertainty and cite their sources. Before any deployment in finance, systematic adversarial robustness checks, along the lines of DecodingTrust [19], should be standard practice. A pipeline where every sentiment judgment is backed by retrieved evidence would help significantly with catching silent hallucinations. Researchers and practitioners will need to collaborate closely to establish clear evaluation standards and deployment guidelines in this space.

6.3. Efficient and reproducible deployment

Open-weight models such as LLaMA [20] allow on-premises hosting, solving both privacy concerns and the problem of recurring API fees. Techniques like quantization and distillation can shrink these models with negligible accuracy loss, making real-time analysis feasible. An open-source approach also ensures reproducibility, which matters not only for academic research but also for auditable financial systems. And once a model is in production, a feedback loop with domain experts can continuously iron out mistakes and improve performance over time.

7. Conclusion

This survey has reviewed the current state of LLM-based financial sentiment analysis under zero-shot and few-shot settings. The key finding is that this technology offers a genuinely flexible and

DOI: 10.54254/2755-2721/2026.LD35648

data-light alternative to traditional fine-tuning, with prompt engineering serving as the central technique for achieving good results. The experimental results on benchmarks like Financial PhraseBank and FiQA demonstrate that well-prompted LLMs can approach or even match the performance of fine-tuned domain models, which is remarkable given that they require no task-specific training data at all. Returning to the three challenges raised in Section 1.2, it is clear that progress is being made on all fronts, though at different speeds. For numerical reasoning errors, retrieval-augmented generation combined with specialized instruction tuning appears to be the most promising path forward, even if it cannot yet fully solve the problem. For hallucination and reliability issues, trustworthiness frameworks like DecodingTrust [19] provide valuable tools for systematic evaluation, and RAG-based factuality checks offer a practical way to catch invented information before it reaches end users. For deployment constraints, the rapid development of open-weight models, such as LLaMA [20] and efficiency techniques like quantization are making on-premises deployment increasingly feasible.

Despite these advances, several limitations remain in the current body of work. Most evaluations are conducted on clean, curated benchmarks rather than noisy real-world data, and the gap between benchmark performance and production performance is still not well understood. Additionally, the financial domain has unique regulatory and compliance requirements that most current research does not fully address. Future work should focus on building more realistic evaluation datasets, developing domain-specific trustworthiness standards, and bridging the gap between academic prototypes and production-ready systems. If these efforts succeed, financial sentiment analysis systems may become accurate, adaptable, and dependable enough for real-world deployment.

References

[1] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv*. https://arxiv.org/abs/1908.10063.

[2] Zhao, W. X. et al. (2023). A survey of large language models. *arXiv*. https://arxiv.org/abs/2303.18223.

[3] Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. *Proceedings in the 2023 ACM International Conference on AI in Finance*, New York, 374–382. https://doi.org/10.1145/3604237.3626869

[4] Qin, C. et al. (2023). Is ChatGPT a general-purpose natural language processing task solver? *Proceedings in the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 1339–1384.

[5] Chen, Z. et al. (2021). FinQA: A numerical reasoning dataset over financial data. *Proceedings in the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, 3697–3711.

[6] Ji, Z. et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

[7] Brown, T. B. et al. (2020). Language models are few-shot learners. *Proceedings in Advances in Neural Information Processing Systems 33*, Online, pp. 1877–1901.

[8] Wei, J. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings in Advances in Neural Information Processing Systems 35*, New Orleans, pp. 24824–24837.

[9] Kojima, T. et al. (2022). Large language models are zero-shot reasoners. *Proceedings in Advances in Neural Information Processing Systems 35*, New Orleans, pp. 22199–22213.

[10] Wei, J. et al. (2022). Emergent abilities of large language models. *OpenReview*, https://openreview.net/forum?id=yzkSU5zdwD.

[11] Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? *SSRN*, https://ssrn.com/abstract=4412228.

[12] Yang, H., Liu, X. Y. and Wang, C. D. (2023). FinGPT: Open-source financial large language models. *arXiv*, https://arxiv.org/abs/2306.06031.

[13] Wang, X. et al. (2023). Self-consistency improves chain of thought reasoning in language models. *Proceedings in the Eleventh International Conference on Learning Representations*, Kigali.

[14] Liu, P. et al. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1-35.

DOI: 10.54254/2755-2721/2026.LD35648

[15] Malo, P. et al. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology, 65*(4), 782-796.
[16] Maia, M. et al. (2018). WWW'18 open challenge: Financial opinion mining and question answering. *Proceedings in the Web Conference 2018 Companion*, Lyon, pp. 1941-1942.
[17] Li, X. et al. (2023). Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks. *Proceedings in the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Singapore, pp. 408-422.
[18] Wu, S. et al. (2023). BloombergGPT: A large language model for finance. *arXiv*, https://arxiv.org/abs/2303.17564.
[19] Wang, B. et al. (2023). DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. *Proceedings in Advances in Neural Information Processing Systems 36*, New Orleans.
[20] Touvron, H. et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv*, https://arxiv.org/abs/2302.13971.
[21] Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings in Advances in Neural Information Processing Systems 33*, Online, pp. 9459–9474.
[22] Gao, Y. et al. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv*, https://arxiv.org/abs/2312.10997.