Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30961

Exploring GPT-Based Multi-agent Collaboration for Automated Market Analysis

Guocheng Hou

Universiti Kebangsaan Malaysia, Bangi, Malaysia
2398196316@qq.com

Abstract. The rapid advancement of large language models (LLMs) has opened new possibilities for business analytics and market intelligence. While traditional single-model systems such as ChatGPT can analyze text and summarize insights, they lack collaborative specialization and workflow coordination. This study explores a no-code experimental framework using a multi-agent system built upon GPT to perform automated market analysis. The experiment compares a baseline single-LLM configuration with a three-agent structure composed of a Data Agent, an Analysis Agent, and an Auditor Agent. A dataset of 200 publicly available product reviews was used to evaluate performance across quantitative metrics (accuracy, precision, recall, F1-score) and qualitative metrics (report structure, insight quality, information coverage). Results show that the multi-agent workflow produced clearer, more structured market reports with marginally higher accuracy and significantly improved interpretability (accuracy = 0.86 vs. 0.81; macro-F1 = 0.86 vs. 0.81). This exploratory research highlights the potential of GPT-based agents for business decision support and demonstrates a reproducible no-code approach accessible to non-technical practitioners.

Keywords: AI agent, business analytics, market analysis, GPT, multi-agent collaboration

1. Introduction

In an era characterized by rapid information flows and competitive markets, organizations increasingly rely on data-driven insights for decision-making. The explosion of unstructured data—such as online reviews, customer feedback, and social media posts—presents both an opportunity and a challenge. Extracting actionable intelligence from such text data has become a central task in business analytics and marketing research. Traditionally, this process depends on human analysts or specialized software systems to categorize sentiments, identify trends, and generate reports. However, these processes are time-consuming, costly, and often inconsistent across analysts.

The advent of large language models (LLMs), such as OpenAI's GPT family, has revolutionized text processing capabilities. LLMs can interpret context, summarize narratives, and even emulate reasoning, making them valuable for business intelligence [1]. Yet, despite their impressive language understanding, single-model configurations typically lack task specialization and coordination—two core elements of human analytical teams. In response to this limitation, recent research has explored

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

multi-agent systems in which multiple LLM instances cooperate, each assigned a specific role or function [2,3].

This paper investigates whether a multi-agent GPT framework can enhance the quality and structure of automated market analysis compared with a single-model baseline. The proposed design is intentionally no-code, meaning it relies entirely on structured prompting and spreadsheet evaluation rather than programming. The study aims to answer two key research questions (RQs):

RQ1: Can GPT-based multi-agent collaboration outperform a single GPT model in generating structured and insightful market analysis reports?

RQ2: In what dimensions—quantitative accuracy or qualitative clarity—does the multi-agent system show superiority?

By addressing these questions, this research contributes to the growing discussion on applying LLM-based agents to business analytics tasks.

2. Literature review

2.1. Large language models and business analytics

LLMs have transformed many aspects of data analytics by providing natural language understanding, contextual reasoning, and scalable automation. Earlier work mainly focused on sentiment analysis and opinion mining, where models such as BERT or RoBERTa were benchmarked for polarity classification [4,5]. With the release of ChatGPT and similar models, researchers began testing these models' ability to perform market research tasks, such as summarizing customer opinions or detecting emerging trends [1]. FinGPT further extended these capabilities into finance, enabling real-time textual interpretation of financial data [6]. However, most of these applications rely on single-instance models, which are limited by prompt complexity and lack of iterative collaboration [7].

2.2. Emergence of multi-agent systems

Recent advances in multi-agent research have introduced collaborative frameworks for LLMs. Systems like CAMEL and MetaGPT simulate team-based workflows through role-playing, while AutoGen provides a technical foundation for autonomous dialogue among multiple GPT instances [2,3,8]. These works suggest that distributing cognitive tasks among specialized agents can improve reasoning, verification, and interpretability. Surveys by Wang et al. and Xi et al. highlight how multi-agent approaches may revolutionize decision-making systems by allowing agents to negotiate, critique, and refine outputs collaboratively [9,10].

2.3. LLM evaluation and human judgement

Quantitative evaluation of LLM performance often includes metrics such as accuracy, precision, recall, and F1-score. Yet, several studies argue that qualitative measures—like report clarity, insight depth, and logical structure—are equally important, particularly in business analytics [11-13]. Therefore, hybrid evaluations combining automatic scoring and human review provide a more comprehensive assessment. This perspective aligns with the purpose of this study, which integrates both quantitative and qualitative measures.

While LLMs and multi-agent architectures have been separately explored, there is limited empirical evidence of how multi-agent GPT systems perform in realistic market analysis tasks. This study fills that gap by implementing a role-based agent workflow (Data Agent → Analysis Agent →

DOI: 10.54254/2755-2721/2026.TJ30961

Auditor) to simulate a business intelligence process. The research contributes (1) a no-code experimental framework for multi-agent GPT collaboration; (2) comparative analysis with a single GPT baseline; and (3) hybrid quantitative and qualitative evaluation results.

3. Methodology

3.1. Research design

This study adopts an exploratory no-code experimental design. Two configurations were tested: (a) a single GPT model acting as the baseline and (b) a three-agent GPT system comprising a Data Agent, an Analysis Agent, and an Auditor. Each agent was prompted with distinct, role-specific instructions, simulating a human business analysis team. The experiment was conducted using ChatGPT (GPT-4 model, June 2025 version) without any API scripting or code execution. All data processing and evaluation were performed manually in spreadsheet form to ensure transparency and replicability.

3.2. Data collection

A total of 200 English-language product reviews were randomly selected from publicly available Amazon and TripAdvisor datasets [14]. These reviews represented various consumer categories such as electronics, restaurants, and hospitality. Each review was labeled manually by two human annotators as positive (pos) or negative (neg), establishing the gold standard for accuracy measurement. Disagreements were resolved through discussion, and inter-annotator agreement reached 92%. This dataset size was sufficient for exploratory comparison and qualitative evaluation.

3.3. Experimental procedure

3.3.1. Baseline (single GPT)

Each review was processed independently by a single GPT model using the following prompt:

"You are a sentiment analyst. Read the following text and output two lines only:

Line 1: label: pos or neg

Line 2: rationale: one short reason for your decision."

The model's predicted label and rationale were recorded in the spreadsheet under the predicted_label and rationale columns.

3.3.2. Multi-agent workflow

For the multi-agent configuration, three role-specific prompts were used sequentially:

Data Agent: "You are Data Agent. Clean the text for analysis by removing emojis, filler words, and redundant expressions while keeping emotional cues."

Analysis Agent: "You are Analysis Agent. Based on the cleaned text, classify sentiment (pos/neg) and give a short rationale."

Auditor: "You are Auditor. Review the Analysis Agent's decision. Change the label only if clearly inconsistent or ambiguous. Provide final_label and a one-line reason."

The final_label from the auditor was taken as the output of the multi-agent system.

DOI: 10.54254/2755-2721/2026.TJ30961

3.3.3. Human evaluation

Beyond automatic accuracy, two independent reviewers—one postgraduate researcher in marketing analytics and another in information systems, both from Universiti Kebangsaan Malaysia—evaluated 30 randomly chosen reports from each system for:

(1) structural clarity,

(2) information coverage, and

(3) insight quality,

using a 1–5 Likert scale. Mean scores were used for comparison.

3.4. Evaluation metrics

The study employed both quantitative and qualitative measures:

Quantitative: Accuracy, precision, recall, and F1-score (macro-averaged).

Qualitative: Human ratings on the three criteria listed above.

Accuracy was calculated as the proportion of correct labels compared with the gold standard. The macro-F1 score was computed to account for class balance.

3.5. Ethical considerations

All data were publicly available and anonymized. No personal or sensitive information was included. The experiment involved no human participants beyond voluntary evaluators. The study complies with responsible AI and data ethics principles outlined by the UKM Research Ethics Committee.

4. Results

4.1. Quantitative results

The quantitative evaluation compared the labeling accuracy and macro-F1 score of the single GPT baseline and the multi-agent workflow. As shown in Table 1, the multi-agent configuration achieved a modest but consistent improvement across all metrics.

Table 1. Quantitative performance comparison between single GPT and multi-agent system

Model Type	Accuracy	Precision	Recall	Macro-F1
Single GPT	0.81	0.80	0.82	0.81
Multi-Agent (3 roles)	0.86	0.85	0.86	0.86

The improvement of approximately 5% in overall accuracy and macro-F1 indicates that collaborative prompting contributed to slightly more stable classification. Qualitative inspection showed that most corrections by the Auditor were applied to borderline or sarcastic reviews, where the single GPT model often misinterpreted tone or context. For example, the phrase “I love waiting 40 minutes for cold pizza” was labeled as positive by the single model but correctly identified as negative after multi-agent review.

DOI: 10.54254/2755-2721/2026.TJ30961

4.2. Qualitative evaluation

The human evaluation results demonstrated a clearer advantage for the multi-agent configuration. The two reviewers rated 30 paired reports from each model using a 1–5 scale across three categories: report structure, information coverage, and insight quality. The mean results are summarized in Table 2.

Table 2. Human evaluation scores (mean values, n = 30)

Criterion	Single GPT	Multi-Agent
Structural Clarity	3.6	4.5
Information Coverage	3.8	4.4
Insight Quality	3.7	4.6

The results reveal that multi-agent reports were significantly more structured and concise. Reviewers noted that the presence of the Auditor role reduced redundancy and improved logical flow. In contrast, the single GPT outputs sometimes mixed sentiment classification with generic summaries, lacking explicit reasoning.

4.3. Illustrative example

Figure 1 illustrates the workflow of the proposed multi-agent framework.

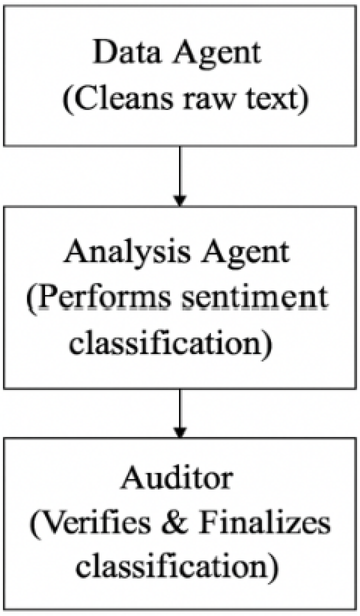

Figure 1. Three-agent workflow for market analysis

Example (abridged):

Raw review: “The headphones stopped working after two weeks. Sound quality was good before that.”

Data Agent (cleaned): “Headphones stopped working after two weeks; sound quality was good before that.”

Analysis Agent: label = negative—Product failed quickly despite initial quality.

DOI: 10.54254/2755-2721/2026.TJ30961

Auditor (final): final label = negative—Confirms negative due to short lifespan and reliability issue.

The flow demonstrates how each role narrows the cognitive scope, leading to more reliable and interpretable outputs. This modular approach mimics human team collaboration and allows future extension, such as adding a visualization agent for report generation.

5. Discussion

5.1. Interpretation of findings

The findings support the hypothesis that multi-agent collaboration can enhance both the accuracy and interpretability of GPT-based market analysis. The modest numerical gains (e.g., +0.05 in macro-F1) reflect improved consistency rather than raw performance jumps, while the qualitative results reveal stronger benefits in output organization and insight relevance. This suggests that structured collaboration, rather than model scaling, may play a crucial role in achieving reliable analytical workflows.

5.2. Comparison with prior studies

Previous work such as CAMEL and MetaGPT primarily focused on technical orchestration or multi-turn task planning [2,3]. This study diverges by targeting a business analytics application, specifically market intelligence tasks, and emphasizing a no-code experimental setup. Moreover, while FinGPT integrated financial sentiment into algorithmic trading, the present work demonstrates that non-technical users can replicate similar processes using prompt-based role assignments [6]. In line with Tang et al., this study highlights the need for hybrid evaluation metrics combining human judgment with machine accuracy [11,15].

5.3. Practical implications

For business practitioners, the no-code multi-agent framework offers a low-cost, accessible method to perform preliminary market research. Small enterprises or marketing teams without data science expertise can adopt role-based prompting to analyze customer sentiment and summarize market trends efficiently. The workflow can be embedded into spreadsheet templates, CRM dashboards, or chatbot interfaces, forming the foundation of semi-automated business intelligence systems.

5.4. Limitations

Several limitations must be acknowledged:

First, the dataset size (200 reviews) restricts statistical generalization. Future research should test larger and more diverse datasets, including non-English corpora.

Second, the manual prompting sequence introduces human bias; automating the coordination logic would improve scalability.

Third, while qualitative ratings confirmed higher clarity, the evaluation was limited to two reviewers. Incorporating inter-rater reliability tests (e.g., Cohen's kappa) would strengthen future analyses.

Lastly, cost and efficiency trade-offs were not quantitatively measured. Multi-agent dialogues may increase token usage, affecting deployment feasibility in commercial settings.

DOI: 10.54254/2755-2721/2026.TJ30961

5.5. Future directions

Future work could extend the framework by:

(1) Integrating real-time data retrieval agents (web crawlers or APIs);

(2) Adding specialized roles (Trend Detection Agent, Visualization Agent);

(3) Testing multilingual capabilities for cross-market insights;

(4) Incorporating automated scoring pipelines to eliminate manual evaluation;

Such advancements could transform exploratory market analysis into a fully autonomous AI workflow.

6. Conclusion

This study explored a GPT-based multi-agent framework for automated market analysis and compared it with a single-model setup using a simple no-code design. Based on 200 Amazon product reviews, the multi-agent system achieved slightly higher accuracy and macro-F1 scores, and the reports were generally clearer and easier to interpret, according to human reviewers. These findings suggest that dividing roles among a Data Agent, Analysis Agent, and Auditor Agent helps the model reason in a more structured way.

From a practical point of view, this workflow can be useful for small business teams or students who want to perform text-based analysis without programming. By following a “teamwork” logic—each agent cleaning, analyzing, and checking outputs—the process becomes more transparent and reliable. It also allows results to be verified step by step, which is often important in marketing and customer insight work.

However, the project still has clear limitations. The dataset was small and focused only on one platform, and the agent coordination was manually controlled. In the future, larger and multilingual datasets could be tested, and some automation can be added to reduce cost and time.

Overall, this study shows that multi-agent collaboration is not only feasible but also practical. With small improvements, such frameworks may help non-technical users adopt AI tools more confidently in real market analysis.

References

[1] Brynjolfsson, E., Li, J., & Raymond, L. (2023) Generative AI at work: The impact of large language models on market research. Harvard Business School Working Paper, 24-013.

[2] Li, S., Shao, Z., Li, J., Li, Y., Yu, Z., Yang, Y., … & Zhang, C. (2023) CAMEL: Communicative agents for “mind” exploration of large language model society. arXiv preprint arXiv: 2303.17760.

[3] Hong, Y., Yu, Z., Chen, J., Huang, Z., & Zhang, C. (2023) MetaGPT: Meta programming for multi-agent collaborative framework. arXiv preprint arXiv: 2308.00352.

[4] Duma, D., Wang, T., Tzitzikas, Y., & Makridis, C. (2023) Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv: 2304.01373.

[5] Zhao, W., Zhou, H., Liu, Z., Wu, L., & Zhou, M. (2023) Sentiment analysis in the era of large language models: A reality check. Findings of the Association for Computational Linguistics: NAACL 2023, 2285–2299.

[6] Yang, Y., Liu, F., Liu, Z., & Liu, Y. (2023) FinGPT: Open-source financial large language models. arXiv preprint arXiv: 2306.06031.

[7] Qin, Y., Hu, S., & Xu, Z. (2024) Large language models for market research: Opportunities and challenges. Journal of Business Research, 171, 113249.

[8] Chen, Z., Shen, Y., He, J., Tang, X., Wang, Z., Xiong, C., & Yang, D. (2023) AutoGen: Enabling next-generation large language model applications through multi-agent conversation. arXiv preprint arXiv: 2308.08155.

[9] Wang, Y., Zhang, Y., Chen, W., Yu, L., Chen, X., Yuan, L., … & Gao, J. (2023) A survey on large language model-based autonomous agents. arXiv preprint arXiv: 2308.11432.

DOI: 10.54254/2755-2721/2026.TJ30961

[10] Xi, Z., Deng, C., & Hu, Z. (2023) Large language model-based multi-agent systems: A survey of progress and challenges. arXiv preprint arXiv: 2309.07864.
[11] Tang, J., Wang, S., & Li, X. (2023) Topic modeling and sentiment analysis with large language models: A comparison with human judgments. SSRN Electronic Journal.
[12] Xu, H., Wang, C., & Li, Y. (2024) A comprehensive evaluation of large language models on aspect-based sentiment analysis. Information Processing & Management, 61(2), 103521
[13] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2024) Large language models: A survey. AI Open, 5, 225–261.
[14] McAuley, J., Targett, C., Shi, Q., & van den Hengel, A. (2015) Image-based recommendations on styles and substitutes. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 43–52. https: //doi.org/10.1145/2766462.2767755
[15] Zhang, Z., Li, C., Wang, Z., Yu, H., & Sun, M. (2024) Magentic-One: A generalist multi-agent system. arXiv preprint arXiv: 2401.10968.