

Empirical Evaluation of Language Models for Arabic Financial Sentiment Analysis: A Comparative Study of Pretrained and General Large Language Models

Abdullah Almusned*, Mohammad Mehedi Hassan, Muhammad Al-Qurishi, Bader Alkhamees

Department of Information System-College of Computer and Information Sciences, King Saud University, Saudi Arabia

Abstract—This study presents a comparative study on the effectiveness of large and pre-trained language models in analyzing sentiment in Arabic financial news. While sentiment analysis using AI has advanced significantly in English-language finance, similar progress in Arabic remains limited. Most existing Arabic models, such as AraBERT and CAMElBERT, are trained on general or social media posts and struggle with domain-specific content like financial data. At the same time, modern LLMs like ChatGPT and Gemini have not been thoroughly tested on Arabic financial data. Our work addresses these gaps by evaluating five models—ChatGPT, Gemini, Nuha (an Arabic LLM), AraBERT, and CAMElBERT—on a dataset of 10,166 financial news headlines from Argaam, a leading Saudi financial platform. We conduct three key experiments. First, we conduct a standard sentiment classification comparison across all models. Second, prompt engineering experiment using zero-shot, one-shot, and three-shot strategies. Finally, persona-driven analysis where models respond from the viewpoint of different financial roles. The results show that Nuha consistently delivers the highest performance, except when prompts include user roles. They show promising results, but did not achieve the highest performance. ChatGPT maintains stable results across prompt variations and shows the most remarkable results when there are persona-driven prompts. Gemini performs well in zero-shot but degrades with added context. PLMs underperform, highlighting the need for better domain adaptation. Overall, this study provides practical insights into model behavior under different prompting conditions, emphasizes the importance of cultural alignment, and introduces persona-driven prompting as a valuable strategy for Arabic financial sentiment analysis.

Keywords—Arabic financial news; stock news; financial news; domain specific; ChatGPT; gemini; large language model; prompt strategies; saudi stocks; sentiment analysis

I. INTRODUCTION

The world of finance rapidly evolving thanks to technology like digital payments, online banking, and new ways to check identity and assess risk. At the same time, artificial intelligence (AI) makes understanding financial data easier. Machine learning (ML), and especially Natural Language Processing (NLP), now plays a significant role in analyzing stock markets [1], predicting trends [2], and spotting early signs of financial trouble [3]. One key task in this field is sentiment analysis, which looks at how people feel based on text found in places like financial news [4] [5], social media posts [6], or official company updates. These signals help investors make

informed investment decisions and reduce risk by helping them understand how news might affect stock prices [7],[8].

In the past, researchers used simple machine learning tools and small datasets to study sentiment. Nevertheless, now, more powerful tools such as Pre-trained Language Models (PLMs) and Large Language Models (LLMs) make it possible to handle complicated language and context much better [9],[10]. For example, FinBERT was explicitly designed to understand English financial text and performs well on tasks that require understanding of market tone [11]. Some general-purpose models such as ChatGPT and Gemini have also been tested in different financial settings and languages, especially English, demonstrating competitive performance with proper input [12], [13].

However, despite these advances in English financial NLP, comparable empirical evidence for Arabic financial sentiment analysis remains limited—particularly in stock market news. Arabic models such as AraBERT [14] and CAMElBERT [15] were mostly tested on general texts or social media posts, like telecom reviews [16] or tweets [17],[18]. These models face challenges due to Arabic's complex grammar, various dialects, and a lack of labeled data [19].

Our work helps fill that gap. It compares how well ChatGPT and Gemini (two powerful LLMs) perform against Nuha (an Arabic local LLM) and traditional PLMs like AraBERT and CAMElBERT. The evaluation uses over 10,000 headlines from Argaam, a trusted Saudi stock market news source spanning multiple sectors. This research seeks to identify the sentiment of each news headline, leading to improved decision-making in the financial market. What makes it unique is not just the comparison of models but also two new ideas it tests. First, it looks at how zero-shot, one-shot, and three-shot prompting styles affect LLM model accuracy—a topic not widely explored in Arabic NLP. Second, it tries persona-based prompting by asking the LLM models to act like real-life users, such as financial analysts or traders, to see if that improves their answers. So far, to the best of our knowledge, this is one of the first studies to compare the performance of Arabic PLMs with modern LLMs on actual Arabic financial texts.

In view of the limited empirical evidence on Arabic financial sentiment analysis using modern language models, and the increasing relevance of prompt formulation and contextual framing, this study aims to systematically investigate how model architecture and prompting strategies affect sentiment

*Corresponding author.

classification performance in Saudi financial news. Accordingly, the following research questions are formulated:

- *RQ1*: How do instruction-tuned large language models compare with Arabic pretrained language models in classifying sentiment in Saudi financial news headlines?
- *RQ2*: How do different prompting strategies (zero-shot, one-shot, and few-shot) influence sentiment classification performance across large language models?
- *RQ3*: Does persona-driven prompting enhance the accuracy and robustness of sentiment classification in Arabic financial news analysis?

The rest of the study is as follows. Section II, is the conceptual and theoretical IV reports the results. Section V discusses the findings. Section VI will be the conclusion.

II. CONCEPTUAL AND THEORETICAL FRAMEWORK

This study is grounded in several interconnected theoretical foundations from natural language processing and machine learning that collectively shape the understanding of Arabic financial sentiment analysis. These foundations provide the conceptual perspective through which model behavior is interpreted and experimental methodologies are theoretically supported.

In-context learning, introduced by [20], refers to the capability of large language models to acquire task-specific behaviors directly from examples embedded within prompts, without requiring parameter updates or gradient-based optimization. This paradigm departs from conventional supervised learning, which depends on extensive fine-tuning using labeled datasets. Authors in [21] demonstrated that the effectiveness of in-context learning depends not merely on the presence of examples, but on their quality, relevance, and structural alignment with inference instances, indicating that models infer latent task structures rather than memorize input-output mappings. Authors in [22] further expanded this understanding through chain-of-thought prompting, showing that explicitly structured reasoning steps facilitate complex multi-stage inference. However, [23] reported substantial variability in few-shot performance across different model architectures and training objectives.

Transfer learning principles, formalized by [24], posit that knowledge acquired during large-scale pre-training can be transferred to specialized downstream tasks, with effectiveness governed by domain proximity between source and target domains. Authors in [25] operationalized this concept through Universal Language Model Fine-tuning, demonstrating that sequential adaptation from general to domain-specific corpora outperforms task-specific training from scratch. Authors in [26] further showed that continued pre-training on domain-specific data yields significant performance gains when target domains exhibit specialized vocabulary and discourse patterns. In Arabic financial sentiment analysis, domain adaptation spans both linguistic dimensions—Modern Standard Arabic augmented with financial terminology—and topical dimensions associated with economic and market-related discourse. Comparisons between multilingual models (ChatGPT, Gemini), Arabic-local models (Nuha), and Arabic-specific pretrained models

(AraBERT, CAMeLBERT) enable examination of whether linguistic and topical alignment corresponds with classification performance.

A fundamental distinction in contemporary language models exists between architectures trained solely on language modeling objectives and those subjected to instruction-tuning, which involves explicit optimization for following natural language instructions. Authors in [27] demonstrated through InstructGPT that fine-tuning models on diverse instruction-following tasks using reinforcement learning from human feedback substantially improves zero-shot generalization and task alignment. This approach is grounded in meta-learning principles, whereby exposure to large collections of instruction-formatted tasks enables models to infer task intent and adapt behavior accordingly [28],[29]. Authors in [30] further established that instruction-tuning yields cumulative benefits, improving both zero-shot and few-shot performance. This distinction differentiates instruction-tuned large language models from pretrained language models that rely primarily on masked language modeling objectives.

Prompt engineering has emerged as a central factor influencing large language model effectiveness, with systematic variations in input formulation producing substantial performance differences. Authors in [31] showed that minor changes in wording, phrasing, or structural organization can lead to notable accuracy fluctuations. Authors in [32] identified prompt sensitivity as a fundamental challenge, demonstrating that models may exhibit unstable performance across semantically equivalent but syntactically different prompts. Authors in [33] further documented extreme sensitivity to demonstration ordering in few-shot settings, showing that permuting identical examples can significantly affect outcomes.

Recent studies indicate that large language models can simulate distinct perspectives when explicitly assigned roles, with persona-based prompting influencing reasoning processes and task performance. Authors in [34] demonstrated that multi-persona approaches enhance complex problem-solving by activating different knowledge distributions within learned representations. Authors in [35] provided theoretical grounding, arguing that role-based prompting leverages models' exposure during pre-training to diverse professional registers and domain-specific discourse patterns. Authors in [36] further showed that persona prompts systematically shift model behavior and can improve performance when role framing aligns with domain expertise. In financial discourse, role-based prompting offers a structured means of contextualization, reflecting the differing analytical priorities of analysts, traders, and financial advisors.

Financial sentiment analysis represents a specialized domain characterized by distinctive linguistic properties and a strong dependence on numerical and temporal context. Authors in [37] demonstrated that general-purpose sentiment lexicons systematically misclassify financial texts because commonly used words often carry domain-specific meanings. Authors in [11] operationalized this insight through FinBERT, achieving state-of-the-art results by modeling finance-specific semantic patterns. Authors in [38] further established that financial sentiment is frequently expressed implicitly through factual reporting rather than overt opinion markers, requiring models to integrate numerical reasoning with linguistic interpretation.

In Arabic financial discourse, these challenges are compounded by language-specific factors, including morphological richness, formal registers, and culturally grounded indirect communication styles, which introduce additional layers of semantic complexity.

Multilingual language models aim to learn shared representations across languages, enabling cross-lingual transfer from high-resource to low-resource contexts. Authors in [39] demonstrated this capability through XLM-RoBERTa, achieving competitive multilingual performance despite substantial imbalance in training data. Authors in [40] found that multilingual models acquire language-agnostic syntactic and semantic abstractions. However, [41] reported that cross-lingual transfer degrades for morphologically rich languages such as Arabic, particularly in tasks requiring cultural grounding or domain-specific reasoning. Authors in [42] further showed that multilingual models can support few-shot cross-lingual transfer, though effectiveness depends on linguistic typology and representational quality. These findings highlight the complexities of multilingual modeling in specialized financial domains.

Based on the theoretical foundations outlined above, this study proposes the following hypotheses to guide the empirical investigation:

- H1: Instruction-tuned large language models will outperform Arabic pretrained language models in zero-shot Arabic financial sentiment classification, attributable to their enhanced instruction-following capabilities and higher-level contextual abstraction.
- H2: The impact of prompting strategies will be model-dependent, such that few-shot prompting will benefit models exhibiting strong in-context learning alignment, whereas models optimized primarily for abstraction-based generalization may experience performance degradation when exposed to demonstration-based prompts.
- H3: Persona-driven prompting will improve sentiment classification performance by introducing role-specific contextual framing that aligns model reasoning with domain-relevant financial decision perspectives.

III. METHODOLOGY

This section describes the different experiments used to evaluate the performance of large language models (LLMs) and pretrained language models (PLMs) in Arabic financial sentiment classification as it shown in Fig. 1. We start by defining the datasets and their sources. Then, we define each LLM and PLM model used in this study. Afterward, we introduce the different experiments. Finally, we set and describe the evaluation metrics.

A. Dataset

The dataset utilized in our work consisted of 10,166 Saudi stock market news headlines collected from Argaam [43], a leading financial economic website in the region. Argaam is owned by the Saudi Research and Media Group, a publicly listed company on the Saudi Stock Market. The headlines covered the period from January 1, 2021, to January 1, 2024, and covered 233 listed companies from 11 different sectors

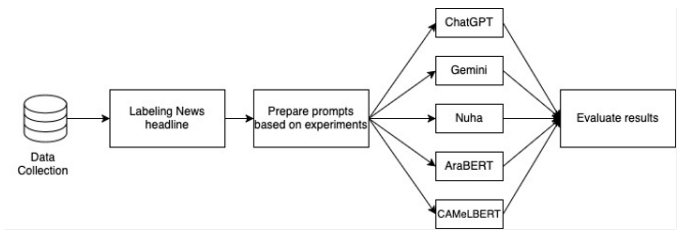


Fig. 1. Used methodology.

in the Saudi market. This comprehensive dataset offered a diverse representation of sentiments and writing styles, with the concise and focused nature of the headlines presenting both a valuable and challenging source for sentiment analysis due to their often implicit emotional small difference.

Each headline was assigned one of three sentiment labels: positive, negative, or neutral. The annotation process followed a hybrid approach, where ChatGPT, Gemini, and Nuha generated the initial labels, followed by manual validation by experts to determine the final reference labels.

B. Models

ChatGPT is a state-of-the-art large language model developed by OpenAI, built on the GPT (Generative Pretrained Transformer) architecture. It has been fine-tuned using reinforcement learning from human feedback (RLHF), which enhances its ability to generate contextually appropriate and coherent responses. Due to its extensive training on diverse and large-scale datasets, ChatGPT exhibits strong generalization capabilities, making it well-suited for tasks such as sentiment analysis, especially in domains like financial news where contextual nuance and linguistic variability are critical. In this study, we used GPT via APIs with a temperature of 0.1.

Gemini is Google DeepMind's advanced large language model, designed to compete at the frontier of generative AI. Leveraging cutting-edge transformer architectures and trained on multilingual, multi-modal datasets, Gemini aims to deliver high performance across a range of tasks, including natural language understanding and generation. Its multilingual capabilities and powerful reasoning abilities make it a promising candidate for complex applications such as sentiment classification in finance, particularly when dealing with rich, structured data inputs. For the purpose of this study, all safety settings were disabled to ensure that the model's raw interpretive capacity could be evaluated without filtering or moderation constraints.

Nuha, previously referred to as Juhaina, is a bilingual Arabic-English Large Language Model developed by Elm Company with the explicit goal of cultural and linguistic alignment with Arabic-speaking users. Nuha supports a wide range of tasks including instruction-following, open-domain question answering, and sentiment classification. It is post-trained using supervised fine-tuning (SFT) and human preference alignment methods to improve its generation quality and adherence to Arabic cultural norms. In our work, Nuha is evaluated alongside other LLMs to assess its performance on Arabic financial sentiment analysis, contributing to the exploration of locally aligned generative AI solutions.

CAMELBERT is a family of Arabic pretrained language models introduced by the CAMEL Lab at New York University Abu Dhabi. The model variants are trained on extensive corpora representing Modern Standard Arabic (MSA), Dialectal Arabic (DA), Classical Arabic (CA), and a mixture of these variants. CAMELBERT adopts the BERT architecture with standardized pretraining procedures, leveraging over 167 GB of high-quality Arabic text. Through systematic experimentation, the developers demonstrated that language variant proximity—rather than data volume—is the most significant factor influencing downstream task performance. In this study, CAMELBERT is used as an Arabic PLM, offering valuable insight into how diverse pretraining data affects sentiment analysis outcomes. Therefore, the ‘bert-base-arabic-camelbert-mix-sentiment’ model was employed without additional fine-tuning.

AraBERT is a monolingual Arabic language model based on the BERT-base architecture, developed by researchers at the American University of Beirut. It consists of 12 transformer layers and approximately 110 million parameters. AraBERT was trained using the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives on a large Arabic corpus comprising 70 million sentences, collected from Arabic Wikipedia, the 1.5B words El-Khair corpus, and the OSIAN news corpus. AraBERT has shown state-of-the-art results across various Arabic NLP tasks such as sentiment analysis, named entity recognition, and question answering. In our study, AraBERT serves as an Arabic Pretrained Language Model (PLM) for sentiment classification. Accordingly, the ‘AraBert-Arabic-Sentiment-Analysis’ model was utilized in its pretrained configuration, without fine-tuning.

C. Experiments

We conducted a series of experiments using multiple prompting strategies across general-purpose and Arabic-specific language models on Arabic financial sentiment data. This section defines each experiment, and the prompt used for it.

1) Experiment 1: Comparative analysis of sentiment analysis models: The first experiment compared five sentiment analysis models—AraBERT, CAMELBERT, ChatGPT, Gemini, and Nuha—on their ability to classify Arabic financial news into three sentiment categories: positive, negative, and neutral. Each model was applied to the same dataset of headlines, ensuring a consistent baseline for comparison as shown in Table I (Exp1). AraBERT and CAMELBERT were leveraging their pre-training on Arabic text. ChatGPT, Gemini and Nuha, as generative language models, were evaluated in zero-shot settings. This experiment aimed to provide insights into the models’ performance across a standard sentiment classification task.

2) Experiment 2: Zero-shot, one-shot, and few-shot prompt strategies: The second experiment focused on comparing the performance of ChatGPT, Gemini, and Nuha under different prompting approaches: zero-shot, one-shot, and few-shot strategies. In this setup, the models were tasked with sentiment classification of Arabic news headlines. Zero-shot prompting involved no prior examples, one-shot used a single example within the prompt, and few-shot included three examples of

TABLE I. PERSONA-DRIVEN PROMPTS

ID	Prompt
Exp1	Respond with only one word (<i>neutral</i> , <i>positive</i> , or <i>negative</i>) based on the sentiment of the input text.
P1	Act as a financial analyst expert in Saudi stocks. How would you assess the tone of the news in the following headline in relation to the stock? Respond only in one word: <i>positive</i> , <i>negative</i> , or <i>neutral</i> . Headline: {headline}.
P2	Act as a Saudi stock expert trader. Based on the headline {headline}, how would you interpret the news regarding its potential impact on the stock? Respond only in one word: <i>positive</i> , <i>negative</i> , or <i>neutral</i> .
P3	Act as a sentiment analysis model trained on Saudi financial news headlines. Based on the headline {headline}, Respond only in one word: <i>positive</i> , <i>negative</i> , or <i>neutral</i> .
P4	Act as an expert in Saudi stock trading. Based only on the headline, will you buy, sell, or hold your stocks the next day? Respond only in one word: <i>positive</i> , <i>negative</i> , or <i>neutral</i> (<i>positive</i> = buy, <i>negative</i> = sell, <i>neutral</i> = hold). Headline: {headline}.

Prompts used to evaluate the effect of different expert personas on sentiment classification of Saudi stock market news headlines.

each class to guide the models as shown in Table I (Exp2). This experiment was designed to evaluate how the inclusion of examples influenced the models’ ability to accurately interpret and classify sentiments, providing deeper insights into their adaptability and learning capabilities. CAMELBERT and AraBERT were not evaluated due to their limited capabilities in zero/few-shot contexts and absence of instruction-tuned architectures, but given their baseline performance in Experiment 1, their inclusion would likely have yielded similarly weak results under complex prompting scenarios.

3) Experiment 3: Persona-driven prompts based on user roles: The third experiment explored the use of role-specific personalized prompts to adapt sentiment analysis outputs based on the perspective of specific professional roles, such as financial analysts, Saudi traders, and other relevant job profiles. For this purpose, ChatGPT, Gemini, and Nuha were tested using tailored prompts designed to align the sentiment classification process with the priorities and contextual needs of these roles. Table I (P1-P4) details the prompts used for sentiment classification. For example, prompts for financial analysts emphasized sentiment relevant to market trends and economic impact, while prompts for Saudi traders focused on actionable insights and immediate relevance to trading activities. This experiment aimed to evaluate the potential of personalized prompts in enhancing the contextual appropriateness and practical utility of sentiment analysis. For the same reason in the second experiment, CAMELBERT and AraBERT were not evaluated.

D. Evaluation Metrics

To evaluate the performance of the five models in sentiment classification task, we employed a standard suite of metrics. These encompassed accuracy, precision, recall, F1-score, Cohen’s Kappa, and confusion matrices.

Accuracy measures the proportion of correctly classified instances. However, it can be misleading with imbalanced datasets.

Precision, Recall, and F1-score were used to address accuracy’s limitations, we used precision (true positives among

predicted positives), recall (true positives among actual positives), and their harmonic mean, the F1-score. These were calculated per class and using macro- and weighted-averaging to account for class imbalance.

Cohen's Kappa measures agreement between predicted and true labels, correcting for chance agreement. A Kappa of 1 indicates perfect agreement, 0 indicates chance agreement, and -1 indicates perfect disagreement. This is particularly useful for comparing models with different class biases.

The Confusion Matrix provides a tabular visualization of classification performance, showing true/false positives/negatives for each class and aiding in identifying misclassifications patterns.

IV. RESULTS

This section presents the findings intended as descriptive comparisons from three experiments designed to evaluate the performance of five large language models, ChatGPT, Gemini, Nuha, CAMELBERT, and AraBERT, in the sentiment classification tasks. The models were assessed using standard performance metrics, accuracy, precision, recall, F1-score, and Cohen's Kappa. The following subsections detail the outcomes of each experiment, highlighting key patterns and comparative insights.

A. Experiment 1: Comparative Analysis of Sentiment Analysis Models

The evaluation results presented in the comparative Table II offer significant insights into the performance of five language models: ChatGPT, Gemini, Nuha, CAMELBERT, and AraBERT. Among the evaluated models, Nuha achieved the highest observed performance, with an accuracy of 89.09%, a precision of 0.8958, a recall of 0.8909, and an F1-score of 0.8914. This suggests Nuha is highly effective in distinguishing between different sentiment categories while maintaining a strong balance between precision and recall. Gemini follows closely, with an accuracy of 88.69% and the highest precision 0.9027, indicating a conservative classification strategy that prioritizes correctness in positive sentiment predictions. However, its recall is marginally lower than Nuha. ChatGPT, while slightly trailing behind Gemini and Nuha, still performs competitively with an accuracy of 86.70%, a precision of 0.8861, a recall of 0.8670, and an F1-score of 0.8633, demonstrating a strong generalization capability.

Conversely, CAMELBERT and AraBERT exhibit significantly weaker performance, with CAMELBERT achieving 60.21% accuracy and AraBERT scoring the lowest at 55.14%. These results indicate their limited ability to accurately classify Arabic sentiment, likely due to suboptimal domain adaptation. The confusion matrices further illustrate classification errors and model biases, shedding light on misclassification tendencies.

Nuha minimizes classification errors, correctly predicting 4,779 positive, 937 neutral, and 3,341 negative instances, with only misclassified 767 positives, 96 neutral and 246 negatives, reinforcing its robustness. Gemini, while strong in both precision and recall, struggles with false negatives, misclassifying 1,000 positive instances, which may indicate a conservative

TABLE II. EXPERIMENT 1 PERFORMANCE METRICS

Model	Accuracy (%)	Precision	Recall	F1-score
ChatGPT	86.70	0.8861	0.8670	0.8633
Gemini	88.69	0.9027	0.8869	0.8876
Nuha	89.09	0.8958	0.8909	0.8914
CAMELBERT	60.21	0.7198	0.6012	0.5056
AraBERT	55.14	0.7325	0.5514	0.3989

Performance comparison of large language models and transformer-based baselines on Experiment 1. The best result in each metric is shown in bold.

bias in sentiment classification. ChatGPT performs well in classifying 5,469 positive, 816 neutral and 2,529 negative sentiments correctly, but exhibits a relatively high number of false negatives of 1,058 misclassified negative instances, suggesting a tendency to overestimate negativity in certain contexts. Fig. 2 shows the confusion matrix for each model.

Pre-trained models, CAMELBERT and AraBERT exhibited significant classification confusion, particularly in differentiating neutral and positive sentiment. CAMELBERT misclassified 956 neutral instances as positive, while AraBERT exhibits a strong bias toward predicting positive sentiment, incorrectly classifying 1,031 neutral instances as positive and 3,525 negative instances as positive. This suggests that AraBERT struggles to differentiate between neutral and positive sentiment polarities, a significant limitation for applications requiring precise sentiment analysis. Additionally, CAMELBERT's high misclassification rate for negative sentiment where 576 instances are incorrectly labeled as positive, suggests it lacks strong contextual understanding of negative sentiment indicators.

From the confusion matrices, two key patterns emerge. First, bias toward positive sentiment is apparent in AraBERT and CAMELBERT, with a disproportionate number of neutral and negative instances classified as positive. This suggests these models fail to capture sentiment variations, particularly in distinguishing neutral language from positive sentiment. Second, ChatGPT and Gemini exhibit higher false negative rates, where positive instances are sometimes misclassified as negative. While this could indicate a conservative classification strategy aimed at reducing false positives, it may lead to missed positive sentiment detections, which is crucial in sentiment-sensitive applications like our situation financial news.

To further assess inter-rater reliability and classification agreement, Cohen's Kappa coefficient was calculated for each model as it shows in Fig. 3. Cohen's Kappa is a more robust measure than accuracy as it accounts for chance agreement. The results reveal a clear distinction in model reliability. Nuha achieved 0.8107 and Gemini 0.8064 exhibited the highest agreement with ground truth, reinforcing their strong consistency in classification. ChatGPT with score 0.7537 follows closely, indicating good agreement but slightly more variance in predictions compared to Nuha and Gemini. CAMELBERT achieved 0.1526 demonstrates poor agreement, aligning with its high misclassification rates seen in the confusion matrices. AraBERT with score 0.0159 exhibits an almost random agreement level, confirming its severe misclassification issues and bias toward predicting positive sentiment.

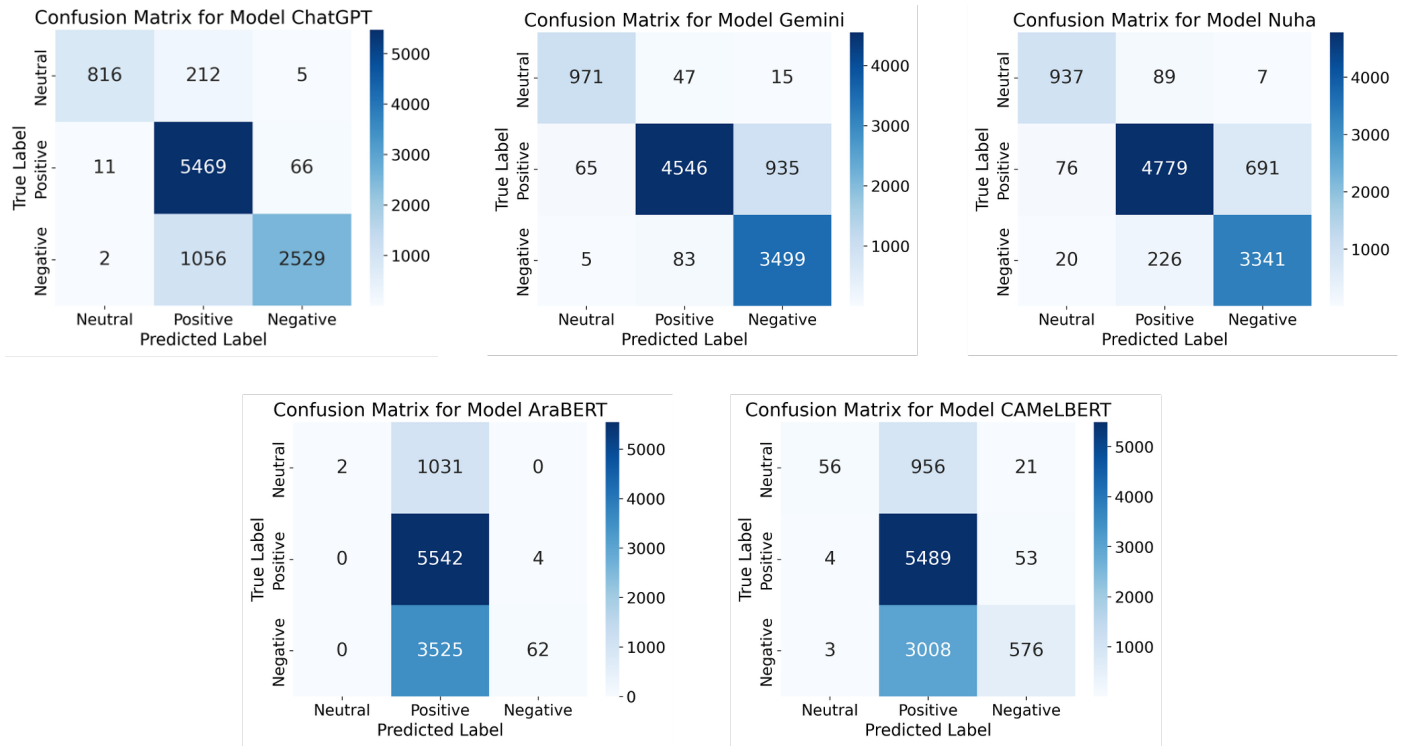


Fig. 2. Experiment 1 confusion matrix.

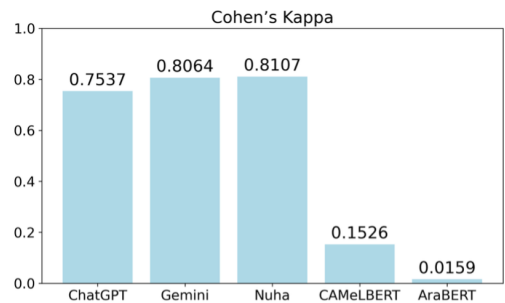


Fig. 3. Experiment 1 Cohen's kappa.

Overall, Nuha proves to be the most reliable model, excelling across all key performance metrics while maintaining low error rates across sentiment categories. Gemini and ChatGPT remain strong alternatives, with Gemini excelling in precision and ChatGPT offering a balanced classification approach. CAMELBERT and AraBERT, however, require significant improvements, particularly in distinguishing between neutral and positive sentiment, as well as in reducing their overall misclassification rates. These findings reinforce the importance of careful model selection in sentiment analysis tasks, particularly in fields where classification precision directly impacts decision-making processes.

B. Experiment 2: Zero-Shot, One-Shot, and Three-Shot Prompt Strategies

The second experiment evaluates the impact of prompting strategy—specifically Zero-Shot, One-Shot, and Three-

TABLE III. EXPERIMENT 2 PERFORMANCE METRICS

Model	Strategy	Accuracy (%)	Precision	Recall	F1-score
ChatGPT	Zero	86.70	0.8861	0.8670	0.8633
	One	87.40	0.8868	0.8740	0.8707
	Three	83.43	0.8599	0.8343	0.8254
Gemini	Zero	88.69	0.9027	0.8869	0.8876
	One	58.34	0.5671	0.5834	0.5632
	Three	56.36	0.5439	0.5636	0.5256
Nuha	Zero	89.09	0.8958	0.8909	0.8914
	One	91.97	0.9200	0.9197	0.9197
	Three	92.45	0.9247	0.9245	0.9245

Performance comparison of different prompting strategies (zero-, one-, and three-shot) across the evaluated models in Experiment 2. The best result for each model is shown in bold.

Shots—on the performance of three large language models: ChatGPT, Gemini, and Nuha. Table III summarized the performance metrics for this experiment.

Nuha emerged as the most consistently performant model across all prompt configurations. It achieved an accuracy of 89.09% in the zero-shot, improving to 91.97% in one-shot and further to 92.45% in the three-shot. Precision, recall, and F1-score followed the same ascending pattern. Cohen's Kappa values for Nuha also reflect strong inter-rater agreement, increasing steadily from 0.8107 to 0.8674. The confusion matrices [Fig. 4(c)] confirm this consistency, with Nuha classifying the majority of positive and negative instances correctly while maintaining balanced performance across all sentiment classes.

These consistent improvements suggest that Nuha benefits significantly from in-context learning, effectively incorporating example-driven prompts to enhance classification quality without suffering from overfitting or degradation.

ChatGPT demonstrated relatively stable performance across all prompting strategies, with minor fluctuations. Its zero-shot accuracy was 86.70%, marginally improving to 87.40% in one-shot, but decreasing to 83.43% in three-shots. Precision and recall peaked in the one-shot setting at 0.8868 and 0.8740, respectively, before declining in three-shots 0.8599 and 0.8343. The F1-score mirrored this trend, with values of 0.8633, 0.8707, and 0.8254. This slight decline in the three-shot setting suggests ChatGPT may struggle to integrate larger contextual prompts effectively. The Cohen's Kappa values of 0.7537, 0.7700, and 0.6931, respectively, support this interpretation, indicating strong agreement in zero- and one-shot, but reduced reliability when exposed to extended prompt length. The confusion matrices [Fig. 4(a)] further illustrate this trend that ChatGPT performs best with minimal context (zero- or one-shot), but misclassifies more negative instances as positive in three-shot, suggesting some semantic drift or context overload. Nevertheless, its relatively high and balanced scores across all metrics indicate that ChatGPT maintains a robust baseline performance and is generally resilient to prompt variation.

In contrast, Gemini exhibits a sharp and unexpected decline in performance when shifting from zero-shot to few-shot strategies. It achieves 88.69% accuracy, 0.8869 recall, and an F1-score of 0.8876 in zero-shot, performing comparably to Nuha and exceeding ChatGPT. However, in one-shot and three-shots, Gemini's accuracy drops to 58.34% and 56.36%, recall to 0.5834 and 0.5636, and F1-score to 0.5632 and 0.5256, respectively. The Cohen Kappa values also plummet 0.2185 and 0.1616, showing low agreement and near-random performance in those settings (Fig. 5). This pattern indicates that Gemini is highly sensitive to few-shot prompting, possibly due to inefficient handling of in-context learning examples or overfitting to provide demonstrations. The conclusion that Gemini excels in generalization without explicit examples, but fails to adapt when such examples are introduced.

The overall findings underscore three key insights. First, in-context prompting improves model performance only when the model architecture and training support such adaptation, as evidenced by Nuha's consistent gains. Second, not all models benefit from increasing the number of examples, with Gemini's performance declining sharply under few-shot conditions, suggesting possible overfitting or distraction effects. Third, ChatGPT offers balanced generalization, performing reliably across varying prompt styles but showing diminishing returns in more complex three-shot contexts.

In conclusion, the experiment highlights the non-linear and model-specific effects of prompting strategies. While Nuha leverages contextual prompts to enhance performance, and ChatGPT maintains stability, Gemini's decline under few-shot settings suggests structural or representational limitations that hinder its ability to utilize prompt examples. These findings emphasize the importance of understanding model behavior in relation to task design and prompt structure for optimal deployment of language models in sentiment analysis.

TABLE IV. EXPERIMENT 3 PERFORMANCE METRICS

Model	Persona	Accuracy (%)	Precision	Recall	F1-score
ChatGPT	P1	91.64	0.9208	0.9164	0.9166
	P2	89.41	0.9042	0.8941	0.8943
	P3	92.84	0.9287	0.9284	0.9282
	P4	90.66	0.9091	0.9066	0.9068
Gemini	P1	87.13	0.8940	0.8713	0.8719
	P2	85.92	0.8871	0.8592	0.8600
	P3	87.13	0.8981	0.8713	0.8718
	P4	87.18	0.8730	0.8718	0.8721
Nuha	P1	87.66	0.8925	0.8766	0.8771
	P2	85.72	0.8822	0.8572	0.8579
	P3	85.58	0.8776	0.8558	0.8564
	P4	86.68	0.8681	0.8668	0.8671

Performance comparison of the four persona-driven prompting strategies across the evaluated models in Experiment 3. The best result for each model is shown in bold.

C. Experiment 3: Persona-Driven Prompts Based on User Roles

The third experiment explores the influence of persona-driven prompt variation (P1–P4) on the performance of three large language models (ChatGPT, Gemini, and Nuha) in sentiment analysis. Each persona represents a distinct style of prompting designed to assess how prompt framing affects classification quality, financial analyst, Saudi stock expert trader, sentiment analysis, and expert in Saudi stock recommendations, respectively. The models were evaluated using accuracy, precision, recall, F1-score, Cohen's Kappa, and supported by detailed confusion matrix analysis to assess class-specific behavior as is shown in Table IV, Fig. 6, and Fig. 7.

ChatGPT demonstrated the highest and most consistent performance across the four evaluated prompts. Its best performance was under P3 (sentiment analysis), where it achieved 92.84% accuracy, 0.9287 precision, 0.9284 recall, and 0.9282 F1-score, with a Cohen's Kappa of 0.8732. The confusion matrix for P3 reveals that ChatGPT correctly classified 5,271 positive, 971 neutral and 3,196 negative instances. Errors were mainly concentrated in misclassifying negative as positive in 385 cases, which is expected in closely aligned sentiment boundaries. Even under its lowest-performing persona, P2 (Saudi stock expert trader), ChatGPT retained 89.41% accuracy, and a solid balance of 0.9042 precision, 0.8941 recall, and 0.8943 F1-score. The P2 confusion matrix shows stronger neutral classification with 987 correct cases but slightly more misclassification of positive and negative examples compared to P3 [Fig. 6(a)]. Under P1 (Financial analyst) and P4 (expert in Saudi stock recommendations), ChatGPT continued to perform robustly, with Cohen's Kappa values of 0.8554 and 0.8380, respectively, and confusion matrices indicating high true positives and effective neutral distinction. These results collectively indicate that ChatGPT is highly responsive to persona framing, with P3 offering the best contextual synergy.

Gemini, while not reaching ChatGPT's peaks, showed strong and consistent performance across personas, with accuracy values ranging from 85.92% to 87.18%. Its most effective performance appeared in P1 (financial analyst) and P3 (sentiment analysis) both with approximately 87.13% accuracy, with

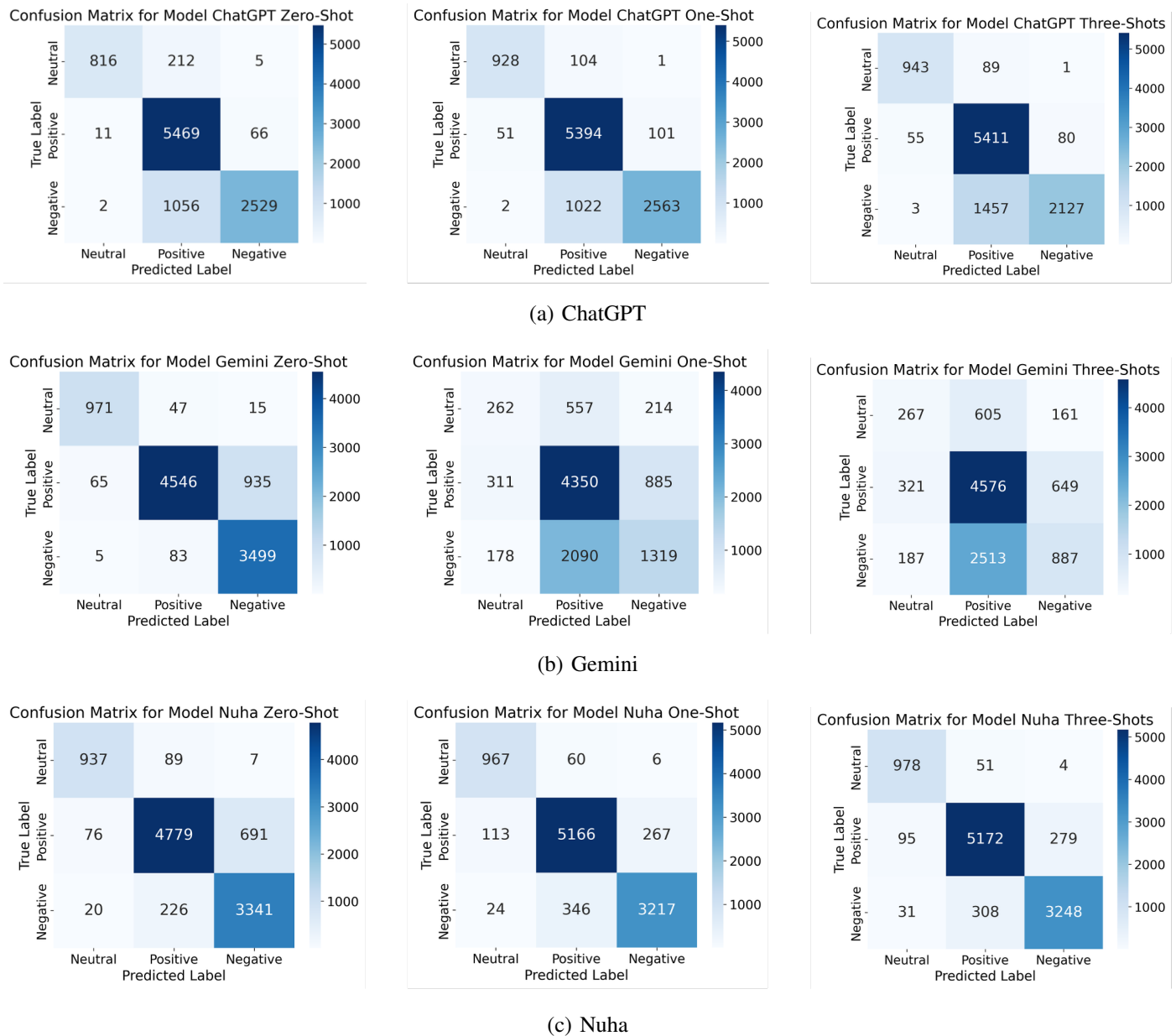


Fig. 4. Experiment 2 confusion matrix.

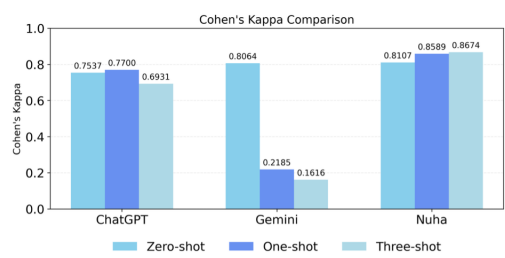


Fig. 5. Experiment 2 Cohen's kappa.

F1-scores of 0.8719 and 0.8718, and Cohen's Kappa values of 0.7824 and 0.7830, respectively. The confusion matrix for P1 indicates solid classification of negative sentiment with

3,525 correct cases and a modest level of misclassification between positive and negative 1,015 false-negatives. Under P3, Gemini improved neutral accuracy with 1,003 correct cases and slightly reduced false positives for positive sentiment [Fig. 6(b)]. P2 (Saudi stock expert trader) showed a drop in performance, with 1,208 false negatives and lower agreement 0.7625 Kappa, indicating sensitivity to specific persona structure. Despite this, Gemini's stability across personas reflects a moderate adaptability, suitable for prompt settings that are consistent and well-defined.

Nuha exhibited the greatest variance across personas, with accuracy ranging from 85.58% with P3 (sentiment analysis) to 87.66% with P1 (financial analyst). Precision, recall, and F1-score followed similar fluctuations. The best performance under P1 produced 0.8925 precision, 0.8766 recall, and 0.8771

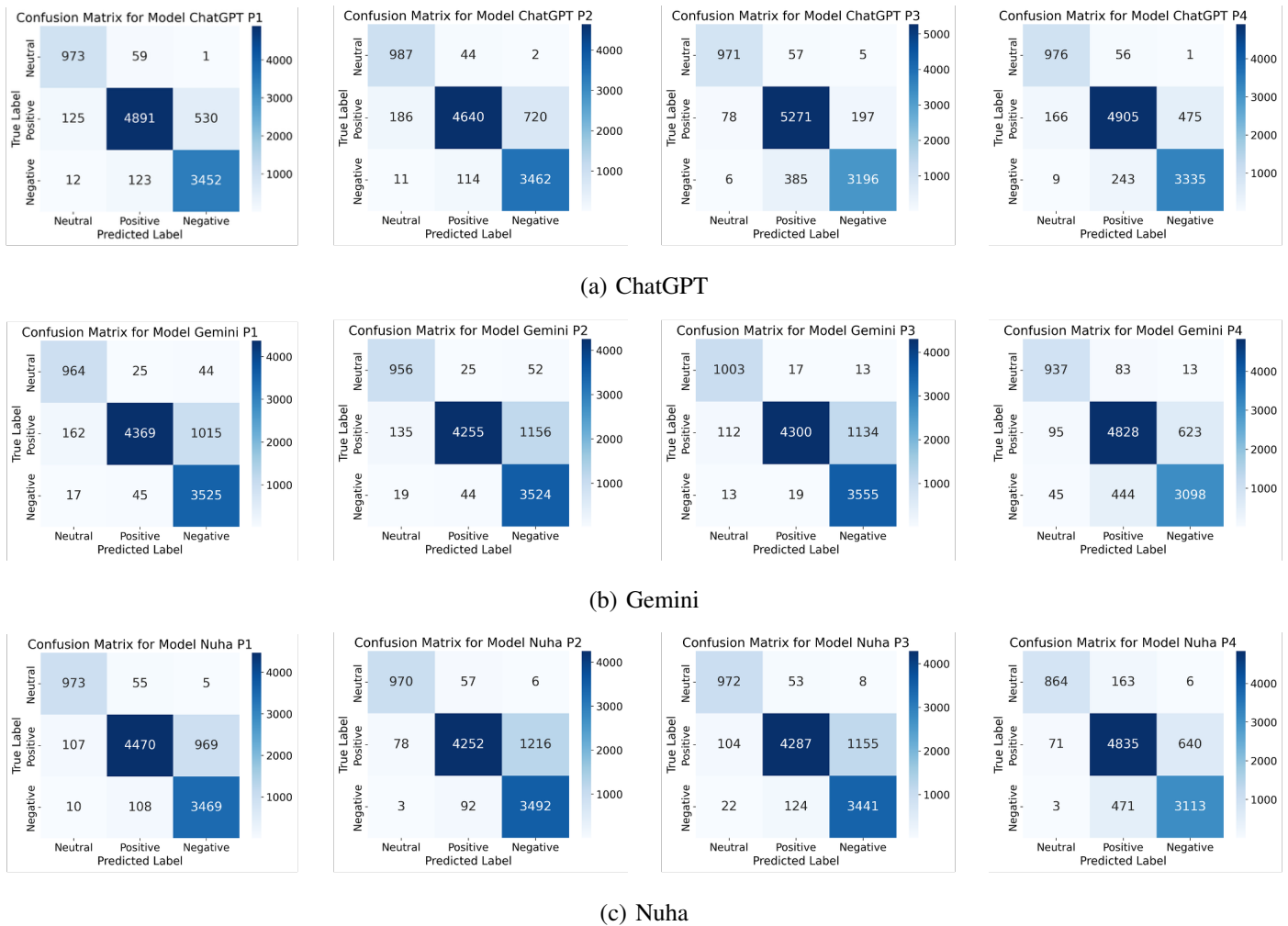


Fig. 6. Experiment 3 confusion matrix.

F1-score, with a Cohen's Kappa of 0.7896. Nuha's P1 confusion matrix shows strong classification of positive 4,470 correct cases, neutral 973 correct cases and negative 3,469 correct cases examples but a notable number of identifying 969 positives as negative cases. In P3, performance slightly declined: fewer positive and negative instances were correctly identified, with an increase in 1,155 false-positives cases and misclassified neutrals. P2 (Saudi stock expert trader) and P4 (expert in Saudi stock recommendations) followed a similar trend, with the P4 confusion matrix revealing the highest misclassification of neutral instances with 163 neutral examples classified as positive, which likely impacted overall precision [Fig. 6(c)]. Cohen's Kappa (Fig. 7) values for Nuha range from 0.7554 to 0.7896, which confirm its inconsistent agreement, particularly in neutral sentiment recognition, which may be influenced by prompt nuances embedded in each persona.

From an error analysis perspective, ChatGPT consistently showed the lowest number of high-impact errors—especially in distinguishing positive from negative sentiments—across all personas. Its three-way balance among neutral, positive, and negative classes was evident across confusion matrices, with P3 performing best. Gemini, while reliable in handling positive and negative sentiment, had greater difficulty distinguishing

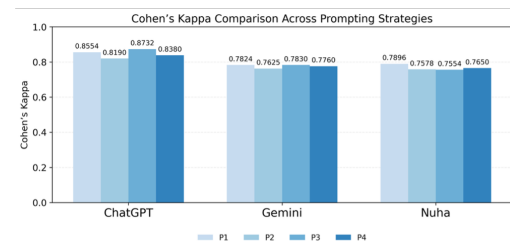


Fig. 7. Experiment 3 Cohen's kappa.

neutral sentiment, particularly under P2 and P4. Nuha struggled most in neutral sentiment detection, frequently confusing neutral with positive, especially in P4, where 163 neutral instances were misclassified as positive, lowering precision and Kappa agreement.

In terms of inter-rater reliability, ChatGPT outperformed both Gemini and Nuha, with Kappa values ranging from 0.8190 to 0.8732, showing strong and consistent agreement with ground truth across all personas. Gemini maintained moderate reliability from 0.7625 to 0.7830, while Nuha's reliability was more affected by personas variation between

0.7554 and 0.7896.

In summary, ChatGPT showed greater adaptability to persona-driven prompting, with the P3 configuration yielding the highest observed performance. Gemini remained stable and moderately effective, while Nuha exhibited a noticeable sensitivity to prompt framing, particularly in handling neutral sentiment. These findings suggest that prompt design—especially persona stylization—can significantly influence model behavior, and that ChatGPT is best suited for stock news. Consequently, persona tuning should be considered a critical component in applied sentiment analysis pipelines, especially when deploying large language models in real-world, user-facing environments, and domain-specific.

V. DISCUSSION

This study investigated the performance of five large language models (ChatGPT, Gemini, Nuha, CAMELBERT, and AraBERT) across three distinct experimental designs in the context of sentiment classification for the Saudi stock market financial news. Each experiment was developed to examine the models' capabilities from a different perspective: a direct zero-shot prompt (Experiment 1), prompt-based contextual learning using different shots techniques (Experiment 2), and persona-driven prompts based on user roles (Experiment 3). The evaluation utilized standard classification metrics—accuracy, precision, recall, and F1-score—supplemented by confusion matrix and Cohen's Kappa to measure inter-rater agreement beyond chance. Together, these experiments reveal critical insights into each model's behavior under varying prompting conditions.

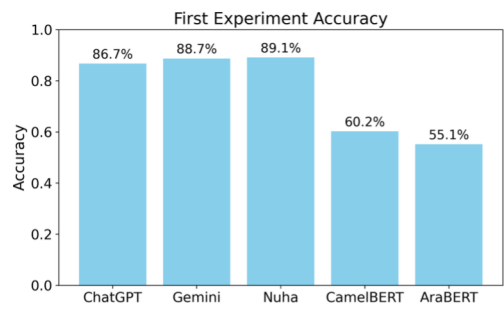


Fig. 8. Experiment 1 accuracy.

In the first experiment, which tested all five models on a standard classification task without prompt variation, Nuha, Gemini, and ChatGPT demonstrated high levels of performance as shown in Fig. 8, achieving accuracies of 89.09%, 88.69%, and 86.70%, respectively. These models also scored highly across precision, recall, and F1-score, suggesting strong overall balance. CAMELBERT and AraBERT, in contrast, exhibited markedly lower performance. CAMELBERT recorded an accuracy of 60.21%, and AraBERT dropped further to 55.14%, with corresponding F1-scores of 0.5056 and 0.3989. The significantly lower Kappa values for CAMELBERT with 0.1526 and AraBERT with 0.0159 indicate poor agreement with ground truth labels, suggesting that these models lack the capacity for nuanced sentiment differentiation, particularly when operating without prompt customization or instruction.

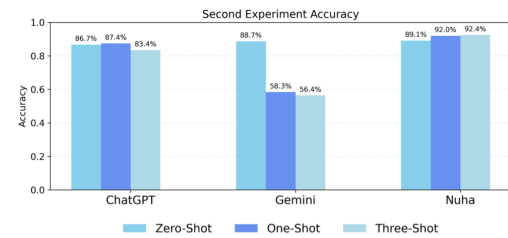


Fig. 9. Experiment 2 accuracy.

Regarding RQ1, the results demonstrate a pronounced performance gap between instruction-tuned large language models and Arabic pretrained language models in the task of Saudi financial sentiment classification. Across all evaluation metrics, ChatGPT, Gemini, and Nuha consistently outperformed AraBERT and CAMELBERT under zero-shot conditions. This outcome supports the premise that instruction-tuned models benefit from exposure to diverse task formulations during training, enabling them to interpret sentiment analysis as an instruction-following problem rather than a fixed classification task. In contrast, AraBERT and CAMELBERT—trained primarily through masked language modeling and dependent on downstream fine-tuning—showed limited adaptability when applied directly to financial sentiment classification without domain-specific retraining. Furthermore, the results indicate that the formal and implicitly evaluative nature of Saudi financial headlines, where sentiment is often embedded within regulatory and factual language, imposes high contextual reasoning demands. Such characteristics favor models capable of instruction-level abstraction over those relying predominantly on token-level pattern learning.

While in the second experiment, it introduced prompt-based context through zero-shot, one-shot, and three-shot prompting strategies. CAMELBERT and AraBERT were not evaluated due to their absence of instruction-tuned architectures. Here, Nuha exhibited the most robust and consistent improvement (Fig. 9), culminating in a three-shot accuracy of 92.45%, and achieving the highest F1-score with 0.9245 and 0.8674 Kappa among all models. ChatGPT showed moderate gains in the one-shot condition with 87.40% accuracy but experienced a decline in the three-shot configuration at 83.43%, possibly due to increased prompt complexity or context saturation. Gemini, in contrast, experienced substantial degradation in performance as prompting complexity increased. While its zero-shot performance remained strong with 88.69% accuracy, its one-shot and three-shot accuracies fell to 58.34% and 56.36%, respectively, accompanied by significant drops in precision, recall, and Kappa scores.

With respect to RQ2, the experiments reveal that the effect of prompting strategies is highly model-dependent. Nuha consistently benefited from richer contextual input, with performance improving from zero-shot to one-shot and three-shot prompting, reflecting effective in-context learning and stable integration of demonstrations. ChatGPT exhibited moderate sensitivity, achieving optimal performance under one-shot prompting while showing a slight decline as prompt complexity increased, suggesting partial context saturation. In contrast, Gemini displayed a marked and unexpected perfor-

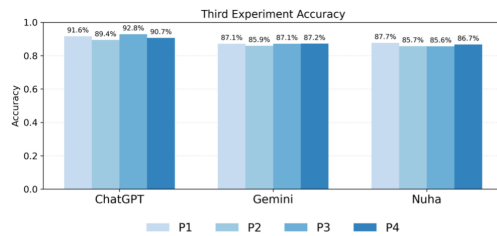


Fig. 10. Experiment 3 Accuracy.

mance degradation when transitioning from zero-shot to few-shot prompting. This behavior can be interpreted in terms of prompt interference, where models optimized for abstraction-based zero-shot generalization may experience misalignment when explicit demonstrations are introduced. These findings indicate that few-shot prompting does not universally enhance performance and may, in some cases, introduce contextual noise rather than beneficial guidance.

To complement the quantitative evidence provided by the confusion matrices, a brief qualitative examination of representative misclassifications was conducted. For instance, the headline “The company records operating losses despite increased revenues” was annotated as neutral in the ground truth but classified as negative by Gemini under few-shot prompting. This misclassification likely reflects the model’s tendency to prioritize loss-related lexical cues over balanced financial indicators when demonstrations emphasize adverse outcomes. In contrast, the headline “Suspension of stock trading until further notice” was consistently predicted as negative by all models, despite being labeled as neutral in the dataset due to its regulatory and procedural nature within the Saudi market context. This observation illustrates how models may overinterpret regulatory actions as sentiment-bearing events, underscoring the inherent difficulty of distinguishing procedural disclosures from evaluative sentiment in financial news reporting.

Experiment 3 focused on the impact of stylistic variation in the prompt, using four different personas (P1–P4) to simulate diverse user roles (Fig. 10). The personas were tested against the LLMs and not PLMs due to their absence of instruction-tuned architectures. In this setting, ChatGPT demonstrated the highest adaptability, achieving an average accuracy of 91.14% across all personas, with peak performance under Persona 3 with 92.84% accuracy, 0.9282 F1-score, and 0.8732 Kappa. These results underscore ChatGPT’s ability to generalize well across expressive prompt styles and reflect its strength in user-facing applications where language input is highly variable. Gemini maintained relatively stable performance across personas, with accuracies ranging from 85.92% to 87.18%, and minimal variation in agreement scores. While it did not achieve the highest metrics, its consistency suggests a degree of resilience to stylistic change, although it remained weaker than ChatGPT in terms of contextual nuance. Nuha, however, exhibited more volatility. Although it performed strongly under P1, its performance decreased in other conditions, particularly when neutral sentiment was present, suggesting a sensitivity to prompt tone and reduced stylistic generalization.

In response to RQ3, persona-driven prompting was found to exert a meaningful influence on sentiment classification

performance, particularly in the case of ChatGPT. Assigning explicit financial roles—such as analyst or trader—provided contextual framing that aligned model reasoning with domain-specific decision perspectives, resulting in higher accuracy and stronger agreement with ground-truth labels. ChatGPT demonstrated the highest robustness across persona variations, suggesting that its instruction-tuning and reinforcement learning mechanisms support effective adaptation to role-based contextual cues. Gemini showed moderate improvement but remained sensitive to persona formulation, while Nuha exhibited greater variability, particularly in neutral sentiment recognition. Collectively, these results indicate that persona-based prompting can enhance the practical relevance of financial sentiment systems; however, its effectiveness depends on the model’s ability to internalize and operationalize role-specific context.

The following analysis compares the effects of few-shot and persona-driven prompting on sentiment classification performance, as reported in Table III and IV. As shown in Table III, increasing the number of in-context examples does not lead to uniform performance gains across models. Nuha benefits from few-shot prompting, ChatGPT exhibits only marginal improvement, while Gemini experiences a notable decline. By contrast, Table IV indicates that persona-driven prompting yields more consistent and stable performance gains for instruction-tuned models, particularly ChatGPT, by aligning model reasoning with domain-relevant financial roles. Whereas few-shot prompting primarily affects how models infer task structure from demonstrations, persona-based framing appears to shape interpretation through contextual role expectations. Collectively, these findings suggest that contextual alignment achieved through role-based prompting may be more effective than increasing demonstration quantity alone for Arabic financial sentiment analysis, especially for models optimized for instruction-following behavior.

VI. CONCLUSION

This study presented a systematic empirical evaluation of Arabic pretrained language models and modern instruction-tuned large language models for financial sentiment analysis in Arabic news headlines within a Saudi market context. By examining model behavior across three experimental conditions—zero-shot classification, demonstration-based prompting, and persona-driven prompting—the study investigated how architectural design, training paradigms, and prompt formulation shape sentiment classification performance in a specialized financial domain.

The findings indicate that instruction-tuned large language models demonstrate greater flexibility and contextual sensitivity than Arabic pretrained language models when applied without task-specific fine-tuning. At the same time, the results reveal that the effectiveness of prompting strategies is strongly model-dependent. While Arabic-aligned models benefited from the inclusion of contextual demonstrations, certain general-purpose models achieved stronger performance under simpler zero-shot configurations and exhibited sensitivity to increased prompt complexity. These outcomes suggest that model scale alone does not ensure robustness, and that performance is critically influenced by the interaction between prompting strategy and training objectives.

Persona-driven prompting emerged as a particularly stable and effective mechanism for contextual alignment in financial sentiment analysis. Framing the task through role-specific perspectives—such as financial analysts or traders—enabled models to better interpret sentiment embedded within factual and regulatory financial language. This approach offers practical relevance for real-world financial applications, where sentiment interpretation is inherently shaped by decision context and stakeholder perspective rather than by generic polarity labels.

In conclusion, this study underscores the importance of aligning model selection and prompt design with domain-specific characteristics when deploying language models for financial text analysis. Rather than promoting a single universally superior model, the findings emphasize that effective Arabic financial sentiment analysis depends on the careful integration of linguistic context, model training paradigms, and prompt engineering. These insights contribute to ongoing scholarly discussions concerning the responsible and effective application of large language models in multilingual and domain-sensitive financial environments.

VII. LIMITATION AND FUTURE WORKS

This study has several limitations that should be acknowledged, which also highlight directions for future research. First, the dataset is constructed exclusively from Saudi financial news headlines obtained from Argaam, and therefore reflects the linguistic, regulatory, and institutional context of the Saudi market. While this specificity enhances contextual relevance, it restricts the generalizability of the findings to other Arabic financial environments, such as those in the UAE, Egypt, or the wider MENA region, where reporting practices and regulatory discourse may differ. Second, the analysis is limited to short news headlines, which often condense complex financial events into brief statements. Longer textual sources—such as full news articles, corporate reports, and social media content—may exhibit different sentiment dynamics and discourse structures, potentially influencing model performance and behavior. Third, although this study evaluates predictive performance, robustness, and prompting sensitivity, it does not incorporate interpretability techniques or user-centered validation. Such components are essential for understanding how financial practitioners interpret, trust, and operationalize AI-generated sentiment outputs.

Accordingly, future research should extend this work by constructing cross-market Arabic financial datasets, examining richer and more heterogeneous textual sources, and integrating explainability methods alongside human-in-the-loop evaluations. These directions would enable a more comprehensive assessment of both technical effectiveness and practical usability within real-world financial decision-support systems.

REFERENCES

- [1] C. Y. Lin and J. A. Lobo Marques, "Stock market prediction using artificial intelligence: A systematic review of systematic reviews," *Social Sciences & Humanities Open*, vol. 9, 2024.
- [2] T. Phuoc, P. T. K. Anh, P. H. Tam, and C. V. Nguyen, "Applying machine learning algorithms to predict the stock price trend in the stock market – the case of Vietnam," *Humanities and Social Sciences Communications*, vol. 11, no. 1, 2024.
- [3] G. Lombardo, M. Pellegrino, G. Adosoglou, S. Cagnoni, P. M. Pardalos, and A. Poggi, "Machine learning for bankruptcy prediction in the American stock market: Dataset and benchmarks," *Future Internet*, vol. 14, no. 8, p. 244, 2022.
- [4] A. Bahatkar, A. Thakre, A. Kharde, A. Singh, and A. Nagdive, "Sentiment analysis of financial news article using TextCNN algorithm and web scrapping," in *International Conference on Emerging Trends in Engineering and Technology, ICETET*. IEEE Computer Society, 2025.
- [5] P. Wattanakul and T. Phientrakul, "Thai stock price prediction from daily news contents," in *Lecture Notes in Networks and Systems*, ser. Lecture Notes in Networks and Systems. Springer, 2022, vol. 191, pp. 851–861, iCTCS 2020.
- [6] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, "Does Twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the H1N1 and the COVID-19 periods," *Cognitive Computation*, vol. 14, no. 1, pp. 372–387, 2022.
- [7] M. F. Hasan, V. S. Bhusari, L. R. Goranta, A. Dharmaraj, S. Kumar, and J. A. Dhanraj, "Artificial intelligence in financial management: Automating risk assessment and investment strategies," in *3rd International Conference on Advances in Computing, Communication and Materials, ICACCM 2024*. Institute of Electrical and Electronics Engineers Inc., 2024.
- [8] N. Seong and K. Nam, "Predicting stock movements based on financial news with segmentation," *Expert Systems with Applications*, vol. 164, p. 113988, 2021.
- [9] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, 2024.
- [10] B. Zhang, H. Yang, and X.-Y. Liu, "Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models," 2023. [Online]. Available: <http://arxiv.org/abs/2306.12659>
- [11] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019. [Online]. Available: <http://arxiv.org/abs/1908.10063>
- [12] A. Buscemi and D. Proverbio, "ChatGPT vs Gemini vs LLaMA on multilingual sentiment analysis," 2024. [Online]. Available: <http://arxiv.org/abs/2402.01715>
- [13] G. Fatouros, J. Soldatos, K. Kouroumalis, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with ChatGPT," *Machine Learning with Applications*, vol. 14, p. 100508, 2023.
- [14] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020. [Online]. Available: <http://arxiv.org/abs/2003.00104>
- [15] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," 2021. [Online]. Available: <http://arxiv.org/abs/2103.06678>
- [16] S. Aftan and H. Shah, "Using the AraBERT model for customer satisfaction classification of telecom sectors in Saudi Arabia," *Brain Sciences*, vol. 13, no. 1, 2023.
- [17] S. Al-Khalifa, F. Alhumaidhi, H. Alotaibi, and H. S. Al-Khalifa, "ChatGPT across Arabic Twitter: A study of topics, sentiments, and sarcasm," *Data*, vol. 8, no. 11, 2023.
- [18] A. Al-Thubaity, S. Alkhereyf, H. Murayshid, N. Alshalawi, M. Bin Omirah, R. Alateeq *et al.*, "Evaluating ChatGPT and Bard AI on Arabic sentiment analysis," 2023.
- [19] A. Louati, H. Louati, E. Kariri, F. Alaskar, and A. Alotaibi, "Sentiment analysis of Arabic course reviews of a Saudi university using support vector machine," *Applied Sciences (Switzerland)*, vol. 13, no. 23, 2023.
- [20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal *et al.*, "Language models are few-shot learners," 2020.
- [21] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" 2022.
- [22] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

- [23] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li *et al.*, “A survey on in-context learning,” 2024.
- [24] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [25] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” 2018.
- [26] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, “Don’t stop pretraining: Adapt language models to domains and tasks,” 2020.
- [27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin *et al.*, “Training language models to follow instructions with human feedback,” 2022.
- [28] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai *et al.*, “Multitask prompted training enables zero-shot task generalization,” 2022.
- [29] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester *et al.*, “Finetuned language models are zero-shot learners,” 2022.
- [30] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus *et al.*, “Scaling instruction-finetuned language models,” 2022.
- [31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [32] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, “Calibrate before use: Improving few-shot performance of language models,” 2021.
- [33] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity,” 2022.
- [34] Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji, “Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration,” 2024.
- [35] M. Shanahan, K. McDonell, and L. Reynolds, “Role play with large language models,” *Nature*, vol. 623, no. 7987, pp. 493–498, 2023.
- [36] L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, and Z. Akata, “In-context impersonation reveals large language models’ strengths and biases,” 2023.
- [37] T. Loughran and B. McDonald, “When is a liability not a liability? textual analysis, dictionaries, and 10-Ks,” *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.
- [38] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, “Good debt or bad debt: Detecting semantic orientations in economic texts,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.
- [39] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán *et al.*, “Unsupervised cross-lingual representation learning at scale,” 2020.
- [40] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4996–5001.
- [41] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, “XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization,” 2020.
- [42] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig *et al.*, “Few-shot learning with multilingual language models,” 2022.
- [43] Argaam, “Argaam,” <https://www.argaam.com/>, 2025. [Online]. Available: <https://www.argaam.com/>