Prompt Engineering with Large Language Models for Vietnamese Sentiment Classification

Dang Van Thin[1,2] and **Duong Ngoc Hao**[1,2] and **Ngan Luu-Thuy Nguyen**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{thindv,ngannlt,haodn}@uit.edu.vn

Abstract

Sentiment Analysis (SA) remains an active research area in Natural Language Processing due to its significance in academia and industry. Recent advancements in large language models (LLMs), including closed-source and open-source models, have demonstrated their potential for enhancing SA tasks. While existing research focuses on high-resource languages like English, this paper aims to conduct a comprehensive investigation into the effectiveness of prompt engineering with various LLMs for Vietnamese SA tasks. Specifically, we experiment with three prompt templates designed in Vietnamese and English, combined with two prompt engineering strategies (zero-shot and few-shot prompting), across the GPT family (GPT 3.5, GPT 4, and GPT 4o) and open-source models (Llama-3, SeaLLM) on six benchmark datasets. Our experimental results demonstrate that employing LLMs with appropriate prompt templates and strategies yields satisfactory performance, surpassing several strong baselines in sentiment classification tasks.

1 Introduction

Sentiment Analysis is one of the active research branches in the field of Natural Language Processing (NLP), with the goal of analyzing and automatically extracting opinions and emotional information aimed at the entities mentioned in the text (Liu, 2022). This task has attracted much attention from researchers because of its potential in real-world applications. Besides, organizations can utilize sentiment analysis applications to monitor multiple social media platforms in real-time and take immediate supportive actions (Feldman, 2013). However, manually conducting the analysis of such a large amount of data will be time-consuming and costly. Therefore, these practical needs have provided strong motivations for much research on the topic of opinion mining.

In recent years, large language models have revolutionized the field of Natural Language Processing, allowing machines to understand human language with increased efficiency (Zhao et al., 2023; Chang et al., 2023). These LLMs are developed based on the Transformer architecture (Vaswani et al., 2017) and trained on the large-scale raw corpora. This helps these models address various challenging NLP tasks in a zero-shot manner. In particular, recent extensive work has been utilising the LLMs to solve the sentiment analysis and has also received the attention of research communities. However, most of the previous studies focused on investigating the performance of LLMs for high-resource languages like English (Zhang et al., 2023b,a; Fatouros et al., 2023; Amin et al., 2023b; Xu et al., 2023; Deng et al., 2023; Amin et al., 2023a). Therefore, exploring the effectiveness of current LLMs in low-resource languages is a crucial research topic, especially for downstream tasks.

For the Vietnamese language, Sentiment Analysis has garnered attention from the research community for more than a decade. Inspired by the initial study (Kieu and Pham, 2010), there has been a significant amount of research in the field of SA at various data domain levels such as education (Nguyen et al., 2018b), hotels (Duyen et al., 2014), and e-commerce (Vo et al., 2017; Nguyen et al., 2018a), etc. Besides, the development of traditional tasks in document-level and sentence-level SA tasks (Thin et al., 2023c), research topics in the field of SA in Vietnamese have focused mainly on aspect-based sentiment analysis tasks (Thin et al., 2023b). Most of the previous works developed methods based on the power of machine learning models (Do et al., 2023), deep learning (Loc et al., 2023) or pre-trained language models (Thin et al., 2023a; Thin and Nguyen, 2023). Exploring the effectiveness of LLMs for a regional language on downstream tasks is one of the crucial research topics. To the best of our knowledge, there is no research exploring the effectiveness of large language models for addressing various Vietnamese SA tasks. In order to bridge this research gap, this paper aims to investigate the effectiveness of various open-source LLMs and GPT series models in handling Vietnamese SA tasks across different scenarios.

2 Related Work

2.1 Vietnamese Sentiment Classification

For the Vietnamese language, the topic of Sentiment Analysis has also received significant attention from the scientific research community, particularly in the past five years. In detail, Thin et al. (2023c) was the first attempt to investigate the effectiveness of fine-tuning pre-trained language models on various Vietnamese benchmark datasets for sentiment classification. Thin et al. (2023b) provided a systematic survey of current research on the ABSA task for the Vietnamese language. The study analyzed different aspects of the topic, including the current approaches, evaluation metrics, and available benchmark datasets. Particularly, Do et al. (2023) presented a Contextualized Window Attention (CWA) method to acquire the context of these groups rather than focusing on an individual word. Another work by Thin et al. (2023a) investigated two ensemble methods: soft-voting and feature fusion, utilizing various pre-trained language models for sentiment classification and aspect-category SA tasks. Loc et al. (2023) proposed a deep learning architecture combined with contextual embeddings from a pre-trained language model.

2.2 Large Language Models for SA

Recently, the development of large language models has received substantial interest across both academic and industrial communities (Zhao et al., 2023; Chang et al., 2023). Most existing LLMs are developed based on the Transformer architecture, as described by Vaswani et al. (2017), and are trained on massive unlabeled corpora. With the growth of LLMs, there have been a number of research efforts aiming at evaluating the performance of LLMs or ChatGPT across Sentiment Analysis tasks (Zhang et al., 2023b,a; Fatouros et al., 2023; Amin et al., 2023b; Xu et al., 2023; Deng et al., 2023; Amin et al., 2023a). Specifically, Zhang et al. (2023b) carried out a systematic evaluation to examine the performance of LLMs in zero-shot and few-shot settings, comparing them with fine-tuned T5 models across various SA tasks and benchmarks. The authors explored three open-source LLMs of the Flan model family and two versions of the OpenAI model. Similarly, the work of Zhang et al. (2023a) investigated three open-source LLMs in both zero-shot and few-shot scenarios on five datasets specific to the software engineering domain. Instead of using the same LLMs as in the previous work (Zhang et al., 2023b), the authors opted for three publicly available LLMs, each with 13 billion parameters. Fatouros et al. (2023) explored the potential of ChatGPT with zero-shot prompting in the finance domain. Amin et al. (2023b) also investigated the capabilities of ChatGPT models, including GPT-4 and GPT-3.5, on various affective computing tasks. The study of Xu et al. (2023) designed a specialized prompt template and examined the limitation of ChatGPT for a complex task, namely the quadruplet ABSA task. The authors (Deng et al., 2023) presented a novel architecture for analyzing market sentiment on social media based on the LLM.

From the analysis above, it is clear that most prior research has focused on evaluating the performance of Large Language Models in the English language. To the best of our knowledge, there has been no exploration into the performance of various LLMs for SA tasks in regional and low-resource languages. As a result, the use of LLMs for these languages is a critical issue. One of the crucial research topics is investigating how existing LLMs can more effectively support the processing of these languages, particularly in downstream applications. Therefore, this paper aims to evaluate the effectiveness of prompt engineering on different current LLMs in the zero-shot and few-shot settings on Vietnamese SA tasks.

3 Methodology

3.1 Prompt Template Design

Large language models can produce different responses depending on the information provided in the prompt template. Therefore, designing effective prompts is challenging due to the variability in the underlying knowledge and background information of different LLMs (Hasan et al., 2024). A well-crafted prompt is crucial for LLMs to understand the task and generate the desired response accurately. As a result, in this work, we explore three prompt templates for both Vietnamese and

English languages. We present three designs for prompt engineering below:

- **Direct Question Prompting**: This prompt format is highly effective for tasks requiring specific answers. It minimizes ambiguity by directly instructing the model to classify sentiment, making it ideal for straightforward tasks or situations where clarity is crucial.

- **Labeling Instructions**: Providing clear instructions ensures the model understands what is expected. This method is particularly effective where consistency and accuracy in response generation are crucial.

- **Role-Playing Prompt**: This approach capitalizes on the ability of LLMs by assigning them a specific role, like a sentiment analysis expert. This can create more engagement in classifying the sentiment polarity class for the input review.

Each template has its strengths and holds potential for exploring the sentiment classification task in various levels of input reviews and domains, especially for low-resource languages such as Vietnamese. Figure 1 illustrates the three prompt template designs in English for the sentiment classification task.

3.2 Prompt Engineering Strategy

Beyond the use of prompt templates, prompt engineering offers a powerful approach to effectively harnessing LLMs for diverse NLP tasks. Given the wide range of prompt engineering techniques and their task-specific nature, this study focuses on applying zero-shot prompting (Wei et al., 2021; Reynolds and McDonell, 2021) and few-shot prompting (Brown et al., 2020a) to the sentiment classification problem. A brief overview of these strategies follows.

- **Zero-shot Prompting**: This strategy involves providing a model with a task instruction without any accompanying examples. The model must generate output based solely on its general knowledge and understanding of the given task.

- **Few-shot Prompting**: This technique incorporates k-shot examples into the prompt to improve in-context learning abilities using demonstrations. Contrary to the approach in the previous work (Min et al., 2022), we randomly select k input-label samples for each sentiment class from the training set. We evaluated using three k-shot settings: 1-shot, 3-shot, and 5-shot. For the ACSC task, we random sample K (k=1,3) examples for each aspect category.

3.3 Large Language Models

In this study, we utilize three major closed-source (GPT 3.5, GPT 4 and GPT 4o) and two open-source LLMs (Llama-3 8B and SeaLLM v3 7B) that have significantly advanced NLP in Vietnamese language. Furthermore, these models are at the forefront of language modelling capabilities and provide robust support for the Vietnamese language.

- **GPT 3.5 Turbo**: GPT-3.5 Turbo is an advanced model in the GPT architecture series developed by OpenAI (Brown et al., 2020b). It enhances the capability to understand natural contexts.

- **GPT 4** (Achiam et al., 2023): This model enhanced capabilities in understanding and generating human-like text. GPT-4 demonstrates exceptional ability in various NLP downstream tasks, especially reasoning tasks.

- **GPT 4o**: GPT-4o is a multilingual and multimodal model that represents an update and optimization of the GPT-4 model. This model has the ability to respond faster and better recognize context to provide answers.

The list of open-source large language models is investigated in this work is present as below:

- **Llama-3 8B Instruct**: is a family of models developed by Meta based on the Llama-2 architecture (Touvron et al., 2023). The models utilize a new tokenizer that expands the vocabulary size up to 128K, enabling efficient multilingual text encoding.

- **SeaLLM v3 7B** (Wenxuan et al., 2024): is the latest models to the SeaLLMs family (Phi et al., 2024), specifically designed for Southeast Asian languages.

4 Experimental Setup

4.1 Experimental Settings

To investigate the performance of GPT-3.5-Turbo, GPT4o and GPT-4, we used the key from Azure

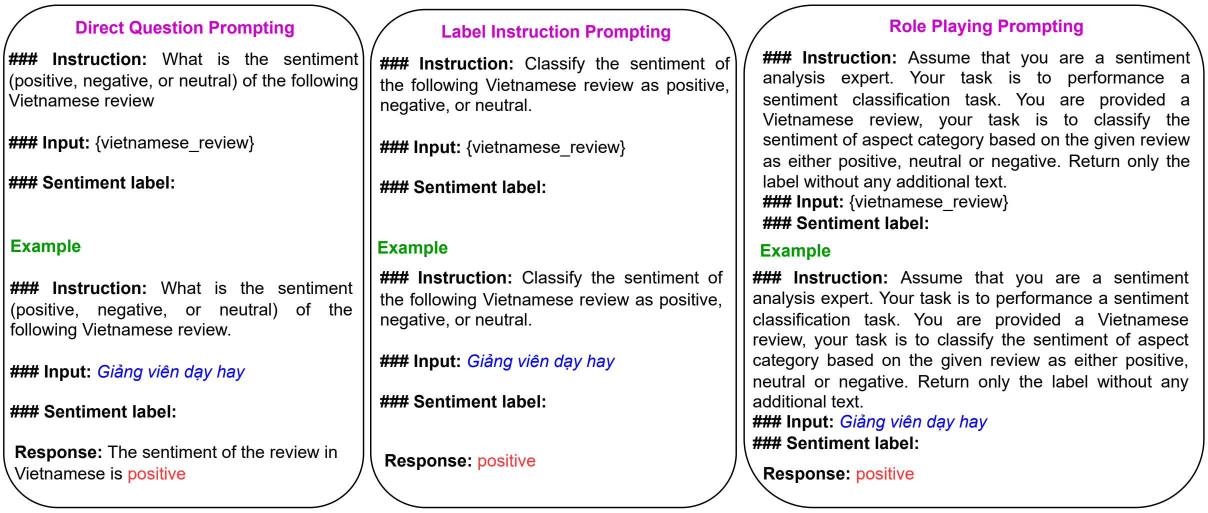

Figure 1: Three Prompt Template designs for Sentiment Classification task.

OpenAPI because of its stability and minimal impact on response time. Two open-source LLMs can be accessed through the Huggingface platform. All experiments were conducted on a single NVIDIA A100 with 80GB GPU and a token length limit of 4096 for the zero-shot and few-shot prompting. The temperature parameter was set to zero to ensure consistency for LLMs, thereby yielding deterministic predictions in the inference phrase.

4.2 Datasets and Evaluation Metrics

For the sentiment classification task, we utilize sentence-level and document-level data from diverse domains. We employ publicly available datasets such as UIT-VSFC (Nguyen et al., 2018b) for the education domain, VLSP (Nguyen et al., 2018a) for social media, and HSA (Duyen et al., 2014) for the hotel domain. We use the same number of samples in our training and testing sets as the corresponding original datasets. For the aspect-category sentiment classification task, we use three datasets for different domains from two previous works, including the restaurant and hotel (Thin et al., 2021), smartphone (Luc Phan et al., 2021). Due to the imbalanced distribution of aspect and sentiment labels in these datasets, we restructured the test set by selecting 50 samples for each aspect category and sentiment extracted from the test and development sets. The training set size is maintained as in prior studies.

4.3 Baseline Comparison Models

To comprehensively evaluate the performance of our results, we compare them against the following approaches:

Fine-tuning pre-trained BERT-based language models (Thin et al., 2023c) have achieved state-of-the-art performance across numerous NLP downstream tasks. For this approach, we re-report the results from previous studies for the sentiment classification task and implement the new models for the ACSA task. We use different robust pre-trained BERT-based language models for the Vietnamese language.

Fine-tuning pre-trained Encoder-Decoder language models can address the understanding tasks by converting them into the text generation problem. In this work, we fine-tuned several of these models, including viT5 (Phan et al., 2022), mT5 (Xue et al., 2021). We use the hyper-parameters as a recommendation in previous works (Thin and Nguyen, 2023; Thin et al., 2023c) for the classification tasks.

5 Results and Discussion

5.1 Zero-shot Strategy

Table 1 and Table 2 present the performance of the zero-shot strategy with different prompt templates on three close-source LLMs for different datasets. As can be observed in Table 1, the “Role-Playing” template tends to have higher Macro F1 and Micro F1 scores across different models, languages, and datasets compared to the other two templates except for the hotel domain. The role-playing approach might encourage the LLM to understand the task better. Therefore, LLMs might focus on relevant aspects of the text and make more accurate sentiment predictions. Moreover, using the “Role-Playing”

Table 1: The results of different prompt templates based on zero-shot strategy on close-source LLMs for the Sentiment Classification. (Best results are highlighted in each column).

Model	Language	Prompt Template	UIT-VSFC		HSA		VLSP		Average
			Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	
GPT 3.5	Vietnamese	Direct Question	64.56	76.03	67.85	77.76	64.66	67.24	69.68
		Labeling Instruction	57.10	66.55	67.11	73.52	67.97	68.48	66.78
		Role-Playing	68.77	82.00	63.47	78.21	68.68	68.79	71.82
	English	Direct Question	65.23	78.71	73.27	82.30	68.63	69.90	72.84
		Labeling Instruction	64.51	77.38	72.13	81.69	65.41	67.24	71.39
		Role-Playing	68.69	81.15	63.60	80.79	69.14	69.24	72.10
GPT 4o	Vietnamese	Direct Question	67.58	80.39	70.58	82.15	59.59	65.52	70.97
		Labeling Instruction	67.28	80.20	70.28	81.54	65.41	68.86	72.26
		Role-Playing	68.76	81.30	74.06	81.24	71.24	72.67	74.88
	English	Direct Question	55.74	79.19	67.72	79.12	49.09	60.95	65.30
		Labeling Instruction	67.97	80.54	70.28	81.54	50.18	61.52	68.67
		Role-Playing	68.96	81.21	74.74	80.33	72.01	72.67	74.99
GPT 4	Vietnamese	Direct Question	69.78	82.38	72.86	82.00	73.69	74.86	75.93
		Labeling Instruction	67.95	80.01	73.18	81.54	72.57	73.52	74.80
		Role-Playing	69.12	81.43	76.38	83.02	**74.71**	**75.43**	76.52
	English	Direct Question	64.98	77.01	73.87	82.75	75.22	75.71	74.92
		Labeling Instruction	64.22	76.06	75.00	82.90	73.60	74.10	74.31
		Role-Playing	**69.31**	**82.93**	**76.74**	**83.06**	74.15	74.76	**76.83**

template makes the interaction with the LLM more engaging and natural, potentially leading to better performance (Sondos Mahmoud Bsharat, 2023).

We also observed that English prompt templates generally outperformed their Vietnamese counterparts across most datasets and prompt templates. However, the performance difference between the two languages was not statistically significant. Even using the Vietnamese prompt with the GPT 4 model gives better results on two metrics for the VLSP dataset. This is primarily due to the fact that most LLMs are initially pre-trained on massive English text corpora, providing them with a stronger foundation in understanding and generating English text compared to other languages. This finding matches those observed in earlier studies (Tran et al., 2024).

As shown in Table 1 and Table 2, the results show that GPT-4 performs better than GPT-3.5 and GPT-4o for most datasets. On average, GPT-4 consistently outperformed the other two models across both SC and ACSC tasks, regardless of the prompt template used. Interestingly, for the more complex ACSC task, the performance difference between GPT-4 and GPT-4o was insignificant when using the 'Role-Playing' template in both languages. Besides, experimental results suggest that the impact of prompt template design diminishes when using large language models like GPT-4 and GPT-4o, likely due to their enhanced ability to understand a broader range of languages and dialects. For example, GPT-4 using a Vietnamese prompt template achieved the best performance on VLSP datasets, with Macro F1 and Micro F1 scores of 74.71% and 75.43%, respectively. Compared to the two smaller open-source LLMs (Llama-3 8B Instruct and Seallm v3 7B), the GPT series models significantly outperform in zero-shot prompting scenarios (see Table 3 and Table 4). In addition, the Llama-3 model gives the best results compared to Sea-LLM v3 in most of the datasets except for the UIT-VSFC.

5.2 Few-shot Strategy

Tables 3 and 4 present the performance of various LLMs under few-shot scenarios for the SC and ACSC datasets, respectively. Generally, k-shot prompting significantly enhances performance compared to zero-shot prompting across most models. However, we observe performance degradation in some high-parameter models like GPT-4 and GPT-4o on the HSA dataset as the number of shots increases. This might be attributed to overfitting, where the model relies on provided examples rather than understanding the underlying task.

Figure 2 demonstrates that using a few-shot prompt with GPT-4 enhanced the overall performance than zero-shot prompting for the UIT-VSFC and HSA datasets. In the case of VLSP, the few-shot approach also improved results, but the difference is not significant in three LLMs. The reason is that the VLSP dataset is a challenging dataset annotated at the document level and contains many vocabulary, syntax and grammar errors. Besides, we noticed that two open-source LLMs (Llama-3 and Sea-LLM) with 5-shot prompting achieved a comparable performance with the GPT-3.5 and GPT-4o in three SA datasets. For the ACSC dataset, the Llama-3 8B Instruct also give better results than GPT-3.5 in the Hotel and Phone datasets. Moreover, the experimental results show that increas-

Table 2: The results of different prompt templates based on zero-shot strategy on close-source LLMs for the Aspect-Category Sentiment Classification.

Model	Language	Prompt Template	Restaurant		Hotel		Smartphone		Average
			Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	
GPT 3.5	Vietnamese	Direct Question	60.25	63.33	66.26	78.14	57.54	75.36	66.81
		Labeling Instruction	51.72	60.00	62.66	77.65	43.20	64.97	60.03
		Role-Playing	51.38	56.67	69.15	83.04	55.33	74.54	65.02
	English	Direct Question	66.91	69.67	69.08	81.75	68.54	79.02	72.66
		Labeling Instruction	64.30	67.67	66.23	80.63	67.55	78.82	70.87
		Role-Playing	56.68	64.50	69.50	82.13	60.16	76.99	68.33
GPT 4o	Vietnamese	Direct Question	55.51	65.00	71.47	85.85	66.22	82.28	71.06
		Labeling Instruction	61.62	67.67	71.26	84.24	65.62	81.26	71.95
		Role-Playing	67.36	71.83	72.27	86.82	71.89	83.32	75.58
	English	Direct Question	63.86	63.83	68.37	86.01	62.83	82.48	71.23
		Labeling Instruction	62.50	63.33	70.84	87.14	63.37	82.48	71.61
		Role-Playing	**71.90**	74.33	73.36	84.89	72.53	83.30	76.72
GPT 4	Vietnamese	Direct Question	70.46	73.67	71.93	86.41	68.40	81.47	75.39
		Labeling Instruction	70.44	73.00	72.89	86.25	**73.75**	81.67	76.33
		Role-Playing	68.18	72.00	73.22	86.17	70.75	81.87	75.37
	English	Direct Question	72.93	72.00	71.48	85.77	69.95	83.10	75.87
		Labeling Instruction	69.69	74.17	73.51	**87.94**	69.31	82.28	76.15
		Role-Playing	71.42	**74.83**	**73.71**	85.93	73.26	**83.87**	**77.00**

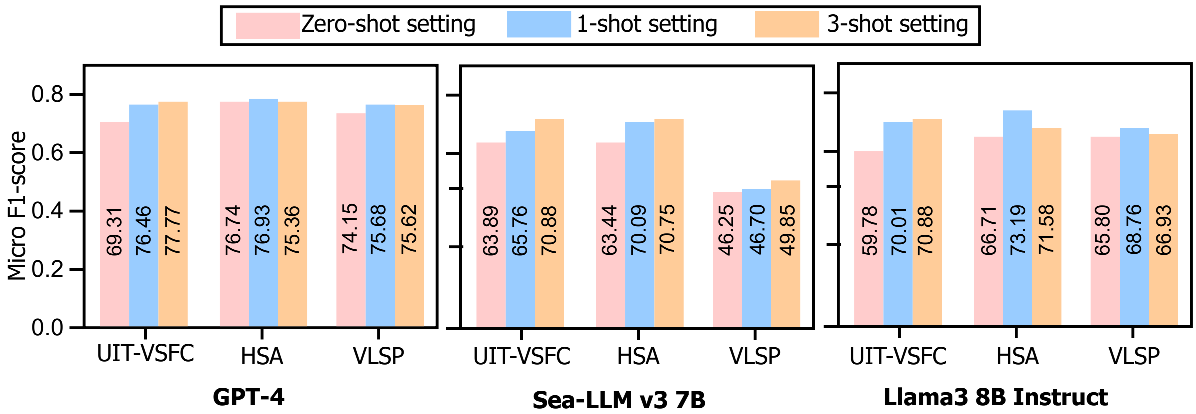

Figure 2: Performance Comparison of GPT-4, Sea-LLM v3 and Llama-3 in Zero-Shot vs Few-Shot Prompting (k=1 and k=3) on three SA benchmark datasets.

ing the k-shot example improves the performance on various datasets in different LLMs. Our results are consistent with previous studies (Zhang et al., 2023b) in the English language.

5.3 Comparison to baselines

In comparison to other baseline approaches, two prompting strategies demonstrate competitive performance across AC and ACSC datasets. Specifically, in the SA datasets, the few-shot prompting approach achieves a weighted F1-score of 91.27% on the UIT-VSFC dataset, surpassing most baseline models except for viT5, XLM-R, and PhoBERT. For the HSA and VLSP datasets, both prompt strategies outperform previous approaches, with improvements of +2.39% and +1.52%, respectively. The comparison of different approaches to the best results of the two prompt strategies is shown in Table 5.

As depicted in Table 6, it can be seen that fine-tuning pre-trained language models in a classification-based approach are strong baselines with the highest performance for the ACSC task, followed by the results of prompt strategies. Despite the complexity of the ACSC task, LLMs with prompt engineering have not yet been able to surpass the performance of fine-tuned small pre-trained language models. Nonetheless, our experiments demonstrate that LLMs can achieve reasonable performance on the ACSC task without requiring the development of new datasets or training custom models.

6 Error Analysis

To better understand LLM performance, we conduct an error analysis based on GPT-4's best results using a few-shot prompting strategy across different datasets. We manually select these incorrect predictions and categorize error types by model.

First, we analyze the confusion matrix to under-

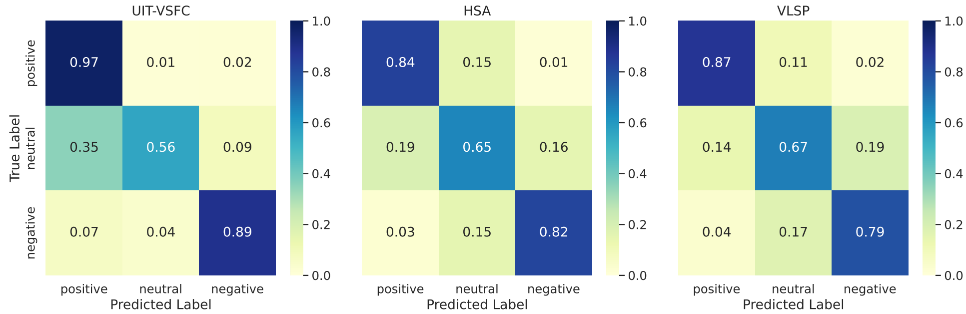

Figure 3: Confusion matrix for three SA datasets.

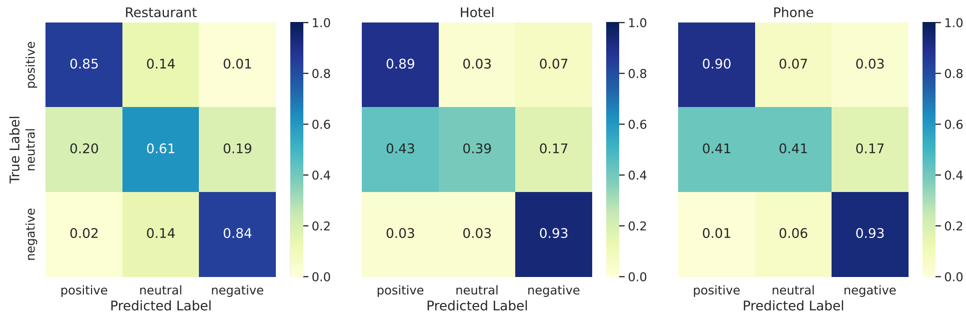

Figure 4: Confusion matrix for three ACSC datasets.

Table 3: Few-shot performance of different LLMs for three SA datasets.

Model	UIT-VSFC		HSA		VLSP	
	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
Llama-3 8B Instruct						
0-Shot	59.78	68.41	66.71	69.59	65.80	65.90
1-Shot	70.01	84.30	73.19	79.43	68.76	68.95
3-Shot	70.88	84.14	71.58	79.12	66.93	68.10
5-Shot	75.96	87.05	70.19	80.03	69.39	69.90
Sea-LLM v3 7B						
0-Shot	63.89	75.36	63.44	69.44	46.08	52.10
1-Shot	65.76	78.49	70.09	77.31	46.70	50.38
3-Shot	71.72	83.86	70.75	75.64	49.82	52.38
5-Shot	71.76	85.06	70.86	77.76	56.14	57.14
GPT 3.5						
0-Shot	71.69	84.65	63.60	80.79	56.14	63.24
1-Shot	74.30	87.21	70.11	81.45	69.52	71.05
3-Shot	72.97	85.79	71.73	80.94	67.33	69.71
5-Shot	73.69	86.83	71.01	81.54	69.10	71.14
GPT 4o						
0-Shot	68.96	81.21	74.74	80.33	72.01	72.67
1-Shot	74.72	86.77	76.46	81.85	77.22	77.14
3-Shot	76.09	88.66	75.29	80.03	76.20	76.38
5-Shot	77.41	89.86	75.38	79.73	**77.70**	77.62
GPT 4						
0-Shot	69.31	82.93	76.74	**83.06**	74.15	74.76
1-Shot	76.46	89.01	**76.93**	82.45	75.68	76.10
3-Shot	77.77	89.51	75.36	80.79	75.62	76.48
5-Shot	**80.41**	**91.25**	75.16	80.18	77.57	**77.71**

Table 4: Few-shot performance of different LLMs for aspect-level sentiment classification datasets.

Model	Restaurant		Hotel		Phone	
	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
Llama-3 8B Instruct						
0-Shot	56.08	60.50	70.21	82.23	67.71	77.80
1-Shot	54.33	60.33	71.39	84.00	65.53	77.39
3-Shot	55.30	61.17	71.59	84.16	65.58	77.39
Sea-LLM v3 7B						
0-Shot	36.94	46.00	56.13	76.05	51.64	68.64
1-Shot	45.93	54.50	65.94	82.88	59.46	74.95
3-Shot	45.29	54.50	64.49	82.80	59.56	74.34
GPT 3.5						
0-Shot	56.68	64.50	69.50	82.13	60.16	76.99
1-Shot	63.75	66.00	71.06	83.76	58.39	74.54
3-Shot	64.56	68.17	69.89	81.35	61.35	74.95
GPT 4o						
0-Shot	71.90	74.33	73.36	84.89	72.53	83.80
1-Shot	71.84	74.17	73.88	85.23	73.11	**84.26**
3-Shot	74.74	76.67	72.32	82.80	**75.07**	82.28
GPT 4						
0-Shot	71.42	74.83	72.71	85.93	70.26	81.87
1-Shot	72.34	75.00	**74.99**	**86.50**	70.58	82.08
3-Shot	**75.66**	**77.67**	73.71	85.13	74.86	83.71

stand better the prediction ability of each label in our best-performing models. The results are shown in Figure 3 and Figure 4 for SC and ACSC tasks, respectively. In analyzing the three SA datasets, we observe that the models effectively classify both negative and positive reviews. Additionally, the percentage of misclassifications between positive and negative labels is minimal in all three datasets. This demonstrates that LLMs are able to classify the positive and negative reviews effectively in most datasets. Two confusion matrices also reveal that most reviews related to the neutral label are incorrectly predicted. Moreover, in some datasets like UIT-VSFC, Hotel, and Phone, the proportion of incorrect data samples is notably higher for neutral and positive labels. The reason for this result is the definition of "neutral" class in the annotation guidelines for each dataset. For example, in Table 7, the review with Id 3, "nói chung là ổn," is an-

Table 5: Weighted F1-score of two prompt strategies against other approaches on three SA datasets. Some results is adapted from (Thin et al., 2023c).

Type	Model	HSA	UIT-VSFC	VLSP
Baselines	MLP (Nguyen et al., 2018a)	-	-	69.40
	MaxEnt (Nguyen et al., 2018b)	-	87.94	-
	LD-SVM (Nguyen et al., 2018c)	-	90.20	-
	VietSentiLex (Vo and Yamamoto, 2018)	77.00	-	-
	BiLSTM-CNN (Le et al., 2020)	-	93.51	-
	Two-channel CNN (Nguyen et al., 2020)	-	88.90	64.00
	Two-channel LSTM (Nguyen et al., 2020)	-	89.30	69.50
	mT5	73.07	89.27	63.27
	viT5	80.80	92.54	75.66
	viBERT_FPT	74.02	90.64	69.98
	viELECTRA_FPT	74.10	89.87	67.33
	mBERT	77.15	91.41	68.53
	XLM-R	74.57	92.55	73.06
	PhoBERT	80.94	**93.45**	76.05
This work (Best results)	Zero-shot Prompting	**83.33**	85.04	74.15
	Few-shot Prompting	81.35	91.27	**77.57**

Table 6: Macro F1-score of two prompt strategies against other baselines on three ACSC datasets.

Type	Model	Restaurant	Hotel	Phone
Baselines	VisoBERT	**82.90**	78.89	86.16
	XLM-R	81.79	77.20	83.81
	PhoBERT	82.82	**79.90**	**86.46**
	mT5	75.12	73.13	71.85
	viT5	77.17	75.14	76.32
This work (Best results)	Zero-shot Prompting	71.90	73.71	73.75
	Few-shot Prompting	75.66	74.99	75.07

notated as “positive” but is predicted as ’neutral’ due to the word ’ổn’ (“okay”). In Vietnamese, this word expresses a moderate emotion and is generally considered neutral sentiment, similar to the example with ID10 in the UIT-VSFC dataset. Besides, we found that the model tends to give the wrong prediction with reviews containing two opposing sentiments. These reviews often are annotated as “neutral” labels based on the guidelines (as examples in Id 2). The lack of this assumption in the models leads to incorrect predictions.

For the SC datasets, we also found that the model often gives the wrong prediction with implicit sentiment, insufficient context, comparison review, or conditional reviews. For instance, in the examples with Id 1, Id 12, and ID 13 in Table 7, it can be seen that these reviews contain implicit sentiments. Therefore, the model must be able to reason to detect the right sentiment label. To address this challenge, the chain-of-thought reasoning prompting technique (Fei et al., 2023) is one of the effective solutions for classifying implicit sentiment in reviews. The model mispredicted some reviews that lack context, such as examples in Id 7, 8, 9, and 14. These samples are ambiguous, and making a decision depends heavily on the definitions of the guidelines and the domain experts. Moreover, the model often fails to predict the comparison review as the example with Id 11 (“Mua ipad air2 cũ ngon hơn nhiều” (*Buying a used ipad air2 is much better*)). We can see that the user compares the current product to the ’old ipad air2’ and expresses that the current product is not good enough to buy. Therefore, the sentiment label is negative. One type of error we also noticed that the model predicted incorrectly was conditional review, as in the examples with Id 4 and 5. It is difficult for a model to identify the right sentiment label for these reviews as human opinions.

In the ACSC task, we noted that the model frequently struggled to accurately predict implicit sentiment, which necessitates analyzing the underlying implications of reviews. As illustrated by examples 1, 2, and 11 in Table 8, the model often misinterprets the context of reviews related to the Drinks#Quality, Drinks#Style_Option and Rooms#Quality aspect categories. These categories typically convey positive sentiments when compared to other aspects. Besides, the model sometimes gives the wrong prediction for some aspect categories that are mentioned in the review but does not express the polarity, such as, for example, in Id 3 and 5. As the same error type as the SC dataset, some review contains the “neutral” vocabulary (ổn (okay) or bình thường (ordinary)), but the model predicts a positive class.

Another type of error occurs when the model is not able to identify the information for the given aspect category, which leads to incorrect classify of the sentiment polarity label. For example, in review with Id 8 as “Quá thất vọng. Đang xài u10 chuyển qua con này do thiết kế màu đẹp hơn nhưng đơ, xài loạn cảm ứng. (*Very disappointed. Switched to this phone due to its nicer color design, but it’s laggy and has an unresponsive touchscreen.*)”, we can easily identify the phrase representing the information for the “Design” aspect as ‘thiết kế màu đẹp hơn” (its nicer colour design), and the corresponding sentiment label is positive. However, it is possible that due to information ambiguity, the model incorrectly predicts the corresponding sentiment label for the “Design” aspect category as negative. To address this situation, future work can require the models to extract the text related to the aspect category before classifying its sentiment polarity. This approach could potentially enhance the overall performance of the ACSC task.

Table 7: Error examples for three sentiment classification datasets.

Id	Dataset	Review	Gold Label	Prediction
1	HSA	Gần đến sáng mới thấy mát ... (*It only starts to feel cool near dawn ...*)	negative	neutral
2		Khách sạn có địa điểm tốt nhưng phòng hơi nhỏ và bí. (*The hotel is well-located but the rooms are somewhat small and stuffy.*)	neutral	negative
3		Nói chung là ổn (*Overall, it's okay*)	positive	neutral
4		nếu có thêm bồn tắm nữa thì không còn gì để phàn nàn. (*If there were a bathtub, there would be nothing to complain about.*)	neutral	positive
5		Nếu phòng lớn hơn một chút sẽ tốt hơn. (*If the room were a bit larger, it would be better.*)	negative	neutral
6	UIT-VSFC	nên cho sinh viên slide để học. (*Students should be given slides to study.*)	negative	positive
7		máy chiếu rõ hơn. (*The projector should be clearer.*)	negative	positive
8		không điểm danh. (*Do not take attendance.*)	neutral	positive
9		dạy full english. (*Teach fully in English.*)	negative	neutral
10		thầy dạy khá ổn. (*The teacher teaches quite okay.*)	neutral	positive
11	VLSP	Mua ipad air2 cũ ngon hơn nhiều (*Buying a used ipad air2 is much better*)	negative	positive
12		ước gì có em này (*Wish I had this one*)	positive	neutral
13		lại là oppo (*It's Oppo again*)	negative	neutral
14		Đùa chứ giờ còn chưa mua nổi note 4?? (*Joking, but I still can't afford a note 4??*)	neutral	negative

7 Conclusion

In this study, we focused on evaluating the performance of various LLMs across different prompt templates and engineering strategies for Vietnamese sentiment classification tasks. To our knowledge, this is the first comprehensive investigation of LLMs for diverse Vietnamese datasets. Our extensive experiments demonstrated that the GPT-4 model, combined with a role-playing template in English, consistently achieved the highest performance across most datasets. Moreover, the few-shot prompting strategy effectively enhanced overall performance for both SC and ACSC tasks, regardless of whether the LLMs were open-source or closed-source. Compared to previous baseline approaches, employing LLMs with prompt engineering, particularly for datasets with limited training data, significantly improved overall performance. The findings presented in our paper can contribute to research on developing AI applications across various data domains, as they address the significant cost associated with annotating datasets for training machine learning models.

Acknowledgements

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2024-26-02. Dang Van Thin was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS117.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023a. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

Mostafa M. Amin, Rui Mao, Erik Cambria, and Björn W. Schuller. 2023b. A wide evaluation of chatgpt on affective computing tasks.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Table 8: Error examples for three aspect-category sentiment classification datasets.

Id	Domain	Review	Aspect Category	Gold Label	Prediction
1		70K cốc trà sữa cũng đáng. (*A 70K cup of milk tea is also worth it.*)	Drinks#Quality	positive	neutral
2	Restaurant	Nước có size khổng lồ uống mệt nghỉ luôn. (*The drink has a giant size, drinking it is exhausting.*)	Drinks#Style Options	positive	negative
3		Về đồ ăn khá tệ so với mức giá voucher 185k. (*The food is quite bad compared to the 185k voucher price.*)	Food#Prices	neutral	negative
4		Menu khá sang choảnh nha ko bình dân tí nào. (*The menu is quite luxurious, not casual at all.*)	Drinks#Style Options	positive	negative
5		Sản phẩm tốt trong tầm giá. Cấu hình cao, thiết kế đẹp, bộ nhớ 128GB. Qúa tốt để chơi game (*Good product for the price range. High configuration, beautiful design, 128GB memory. Too good for gaming.*)	Storage	neutral	positive
6	Phone	máy có thiết kế đẹp tuy nhiên cấu hình thấp đáng tiếc cho thương hiệu nokia vì không thấu hiểu người dùng (*The device is beautifully designed, but its low configuration is disappointing for Nokia.*)	Design	positive	negative
7		Sản phẩm tốt, dung lượng pin lớn dùng đc nhiều ngày, loa nghe to rõ đáp ứng tốt. (*Good product, large battery capacity that lasts many days, loud and clear speaker meets expectations.*)	Storage	positive	neutral
8		Quá thất vọng. Đang xài u10 chuyển qua con này do thiết kế màu đẹp hơn nhưng đơ, xài loạn cảm ứng. (*Very disappointed. Switched to this phone due to its nicer color design, but it's laggy and has an unresponsive touchscreen.*)	Design	positive	negative
9		Nhân viên cũng ổn. (*The staff is okay.*)	Service#General	neutral	positive
10	Hotel	Tôi thấy họ cũng nhiệt tình hỗ trợ khách hàng, nhưng chuyển tới chuyển lui vậy cũng bất tiện. (*I find them enthusiastic in customer support, but moving around like that is inconvenient.*)	Service#General	positive	negative
11		Nhân viên phục vụ thái độ cũng được và giá cả thì không xứng đáng với chất lượng phòng. (*The service staff's attitude is okay, but the price does not match the room quality.*)	Rooms#Quality	positive	negative
12		Phòng ở bình thường, không có vấn đề gì phát sinh cả. (*The room is ordinary, with no issues arising.*)	Rooms#General	neutral	positive

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Proceedings of NIPS*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. Llms to the moon? reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1014–1019.

Hoang-Ha Do, Xuan-Hieu Pham, Duc-Hiep Nguyen, Quoc-An Nguyen, Duy-Cat Can, and Hoang-Quynh Le. 2023. Enhancing aspect-based sentiment analysis with contextualized window attention mechanism. In *Proceedings of KSE*, pages 1–6.

Nguyen Thi Duyen, Ngo Xuan Bach, and Tu Minh Phuong. 2014. An empirical study on sentiment analysis for vietnamese. In *Proceedings of ATC*, pages 309–314, Vietnam. IEEE.

Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. 2023. Transforming sentiment analysis in the financial domain with chatgpt. *Machine Learning with Applications*, page 100508.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of ACL*, pages 1171–1182, Toronto, Canada.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis. In *Proceedings of LREC-COLING*, pages 17808–17818, Torino, Italia. ELRA and ICCL.

Binh Thanh Kieu and Son Bao Pham. 2010. Sentiment analysis for vietnamese. In *Proceedings of KSE*, pages 152–157.

Lac Si Le, Dang Van Thin, Ngan Luu-Thuy Nguyen, and Son Quoc Trinh. 2020. A multi-filter bilstm-cnn architecture for vietnamese sentiment analysis. In *Proceedings of ICCCI*, pages 752–763. Springer.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Cu Vinh Loc, Truong Xuan Viet, Tran Hoang Viet, Le Hoang Thao, and Nguyen Hoang Viet. 2023. Pretrained language model-based deep learning for sentiment classification of vietnamese feedback. *International Journal of Computational Intelligence and Applications*, page 2350016.

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence. In *Knowledge Science, Engineering and Management*, pages 647–658, Cham. Springer International Publishing.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP*, pages 11048–11064, Abu Dhabi, United Arab Emirates.

Huyen TM Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2018a. Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.

Quan Hoang Nguyen, Ly Vu, and Quang Uy Nguyen. 2020. A two-channel model for representation learning in vietnamese sentiment classification problem. *Journal of Computer Science and Cybernetics*, 36(4):305–323.

Van Kiet Nguyen, Vu Duc Nguyen, Phu XV Nguyen, Tham TH Truong, and Ngan Luu-Thuy Nguyen. 2018b. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *Proceedings of KSE*, pages 19–24. IEEE.

Vu Duc Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2018c. Variants of long short-term memory for sentiment analysis on vietnamese students' feedback corpus. In *Proceedings of KSE*, pages 306–311.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of NAACL*, pages 136–142.

Xuan Nguyen Phi, Zhang Wenxuan, Li Xin, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. Seallms - large language models for southeast asia. In *ACL 2024 System Demonstrations*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA.

Zhiqiang Shen Sondos Mahmoud Bsharat, Aidar Myrzakhan. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*.

Dang Thin and Ngan Nguyen. 2023. Aspect-category based sentiment analysis with unified sequence-to-sequence transfer transformers. *VNU Journal of Science: Computer Science and Communication Engineering*.

Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023a. A study of vietnamese sentiment classification with ensemble pre-trained language models. *Vietnam Journal of Computer Science*.

Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023b. A systematic literature review on vietnamese aspect-based sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(8).

Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023c. Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Van Dang Thin, Ngan Luu-Thuy Nguyen, Tri Minh Truong, Lac Si Le, and Duy Tin Vo. 2021. Two new large corpora for vietnamese aspect-based sentiment analysis at sentence level. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4).

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dinh Dien. 2024. Vimedaqa: A vietnamese medical abstractive question-answering dataset and findings of large language model. In *Proceedings of ACL*, pages 356–364.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of NIPS*, 30.

Huynh Quoc Viet Vo and Kazuhide Yamamoto. 2018. VietSentiLex: a sentiment dictionary that considers the polarity of ambiguous sentiment words. In *Proceedings of PACLIC*, Hong Kong.

Quan Vo, Huy Nguyen, Bac Le, and Minh Nguyen. 2017. Multi-channel lstm-cnn model for vietnamese sentiment analysis. In *Proceedings of KSE*, pages 24–29, Vietnam. IEEE.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *Proceedings of ICLR*.

Zhang Wenxuan, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages.

Xiancai Xu, Jia-Dong Zhang, Rongchang Xiao, and Lei Xiong. 2023. The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: A comparative analysis. *arXiv preprint arXiv:2310.06502*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*, pages 483–498, Online.

Ting Zhang, Ivana Clairine Irsan, Ferdian Thung, and David Lo. 2023a. Revisiting sentiment analysis for software engineering in the era of large language models. *arXiv preprint arXiv:2310.11113*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.