Exploring the Power of Large Language Models for Vietnamese Implitcit Sentiment Analysis

Huy Luu Gia[1,2] **and Thin Van Dang** [1,2]
[1]University of Information Technology-VNU-HCM, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
23520618@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

We present the first benchmark for implicit sentiment analysis (ISA) in Vietnamese, aimed at evaluating large language models (LLMs) on their ability to interpret implicit sentiment accompanied by ViISA, a dataset specifically constructed for this task. We assess a variety of open-source and close-source LLMs using state-of-the-art (SOTA) prompting techniques. While LLMs achieve strong recall, they often misclassify implicit cues such as sarcasm and exaggeration, resulting in low precision. Through detailed error analysis, we highlight key challenges and suggest improvements to Chain-of-Thought prompting via more contextually aligned demonstrations. The dataset and code are available at the GitHub repository[1].

1 Introduction

Implicit Sentiment Analysis focuses on detecting sentiments conveyed indirectly through context, speaker intent, or pragmatic cues, rather than explicit polar words (Russo et al., 2015). While more challenging than Explicit Sentiment Analysis (ESA), ISA has benefited from recent advances in LLMs, which offer stronger reasoning capabilities (Paranjape et al., 2021; Liu et al., 2022). However, most ISA research has centered on English or other high-resource languages (Duan and Wang, 2024), while Vietnamese—despite being widely spoken—remains under-resourced. Prior Vietnamese sentiment analysis work has largely focused on ESA (Thin et al., 2023b), with no dedicated ISA datasets. Given the syntactic and pragmatic distinctions of Vietnamese, a focused ISA study is both necessary and timely.

To address this gap, our work focuses on evaluating the effectiveness of large language models for the task of implicit sentiment analysis in Vietnamese using state-of-the-art prompting strategies. Our key contributions are threefold: (1) we conduct a comprehensive evaluation of both open-source and proprietary LLMs on sentence-level ISA in Vietnamese; (2) we propose several directions to improve LLM performance on this challenging task, particularly in handling rhetorical and pragmatic features; and (3) we also release **ViISA**, a benchmark test set specifically designed for evaluating LLMs on ISA in Vietnamese.

2 Related works

Sentiment Analysis LLMs have achieved strong results in sentiment analysis, especially in zero/few-shot settings. Some works note their limitations across domains (Zhang et al., 2024), while others use continual learning to improve ABSA (Ding et al., 2024). For Vietnamese, research has focused on fine-tuning pretrained models (Thin et al., 2023b) and prompt engineering (Thin et al., 2024). A recent review (Thin et al., 2023a) highlights challenges in Vietnamese ABSA, but most work targets explicit sentiment.

Implicit Sentiment Analysis ISA is more complex, requiring inference beyond surface cues. Chain-of-thought prompting improves LLM reasoning for subtle sentiment (Fei et al., 2023), while other methods use coherence cues (Duan and Wang, 2024) or combine discourse features with structured prompts (Cui et al., 2023). These approaches stress the importance of reasoning for effective ISA.

3 Dataset Construction

ViISA is a sentence-level dataset for implicit sentiment analysis in Vietnamese, containing only Positive and Negative labels. Based on prior studies on indirect language in Vietnamese (Nguyen, 2020; Le, 2015), we developed annotation guidelines to capture key patterns of implicit sentiment.

[1]https://github.com/HuyGiaLuu/ViISA

Proceedings of the 18th International Natural Language Generation Conference, pages 202–214
October 29–November 2, 2025.

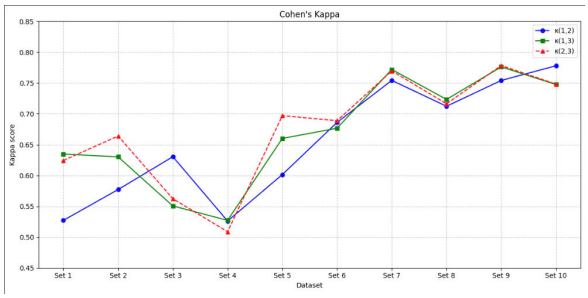

Figure 1: Inter-annotator agreement across 10 rounds of annotation.

Table 1: ViISA and GPT-4o result.

Statistic	Value
Total sentences	302
Negative samples	260
Positive samples	42
Avg. sentence length	20.76 words
Zero-shot GPT-4o Performance	
Accuracy	53.30
F1-weighted	66.21

1. **Definition:** Implicit sentiment arises when affect is inferred from context, tone, or speaker attitude, rather than direct lexical cues.
2. **Indicators:** Typical patterns include: (1) contradiction, (2) exaggeration or understatement, (3) wordplay or ambiguity, (4) rhetorical or sarcastic expressions, (5) tonal irony, and (6) contextual paradoxes.
3. **Annotation:** Annotators used example-driven guidelines focusing on implicit sentiment traits.
4. **Disambiguation:** Difficult cases were resolved collaboratively and used to refine the guideline.

We constructed the dataset using a two-stage process: first, a prompting-based method with LLMs was applied to extract Vietnamese implicit sentiment sentences from an existing dataset; second, human annotators filtered these sentences using predefined guidelines. This approach reduces annotation effort while maintaining linguistic quality. The filtering prompt was designed based on *Few-shot Chain-of-Thought prompting* (Wei et al., 2022), and includes three key components:

- **Role Assignment**: The LLM is prompted to act as a Vietnamese language expert specialized in detecting implicit sentiment.
- **Task Instructions**: The LLM performs a step-by-step process: identify implicit sentiment, return the original sentence, list relevant linguistic features, and explain the decision to support human verification.
- **Feature Descriptions and Examples**: The prompt provides definitions of key implicit sentiment features along with three illustrative examples for each.

For this filtering task, we used **GPT-4o**. As the source data, we adopted the **VLSP 2016** sentiment analysis dataset (Nguyen et al., 2018), which was released at the VLSP 2016 workshop. This dataset consists of Vietnamese technological product review sentences collected from popular online platforms such as *TinhTe.vn*, *VnExpress.net*, and *Facebook*. The full prompt template is provided in the Appendix A.

We conducted annotation over 10 rounds, with inter-annotator agreement tracked for consistency (Figure1). The label imbalance (Table 1) reflects the tendency of implicit sentiment in Vietnamese to appear more frequently as negative or sarcastic expressions.

4 Evaluation Methodology

4.1 Large Language Models

We employ both open-source and closed-source LLMs, applying the aforementioned prompting strategies to evaluate performance on the **ViISA**. The closed-source group includes models from the GPT and Gemini families[2], while the open-source group includes models from the LLaMA, Qwen3, DeepSeek, and Gemma families, as well as Vietnamese-focused models[3].

4.2 Prompting Strategy

We apply different prompting strategies for LLMs on the ViISA dataset, including:

- **Zero-shot reasoning with role-play** (Kong et al., 2024).
- **Plan-and-Solve Prompting (PS prompting)** (Wang et al., 2023).
- **Few-shot Chain-of-Thought (Few-shot CoT)** (Wei et al., 2022).

[2]GPT family: GPT-4o, GPT-4o-mini, GPT-4.1, GPT-4.1-mini. Gemini family: Gemini-2.5-pro, Gemini-2.5-flash, Gemini-2.0-flash, Gemini-2.0-flash-lite.

[3]LLaMA: LLaMA-4-Scout-17B, LLaMA-3.3-70B, LLaMA-3.1-8B. Qwen3: Qwen3-32B, Qwen3-235B. DeepSeek: Deepseek-v3. Gemma: Gemma-3-27B, Gemma-3-12B. Vietnamese-focused: FPT.AI-KIE-v1.7, SaoLa3.1-medium.

- **Active Prompting with Chain-of-Thought** (Wan et al., 2023).
- **Zero-Shot Reasoning with Self-Adaptive Prompting (COSP)** (Diao et al., 2024).

(The full detailed prompts and implementation details for each prompting strategy can be found in Appendix B and Appendix H.)

5 Result Analysis

5.1 Comparison of Prompting Strategies.

Discussion. Among all prompting strategies, **Few-shot CoT** (Wei et al., 2022) consistently achieves the highest performance, as evidenced by the results in Table 2 and Table 3. This highlights the benefit of providing in-context examples, particularly for tasks like implicit sentiment analysis in a low-resource language such as Vietnamese. By contrast, **Zero-shot reasoning with role-play** (Kong et al., 2024), **Plan-and-Solve** (Wang et al., 2023), and **COSP** (Wan et al., 2023) do not offer any concrete examples, which limits their effectiveness on tasks requiring a nuanced understanding of context and lexical cues. Although **Plan-and-Solve** (Wang et al., 2023) attempts to guide the reasoning process with structured questions, it still lacks real demonstrations. **COSP** (Wan et al., 2023), while computationally intensive and reasoning-oriented, also underperforms due to its reliance on LLM-generated paths without human-curated examples. **Active Prompting with CoT** (Wan et al., 2023) attempts to refine **Few-shot CoT** (Wei et al., 2022) by selecting the best examples via an LLM. However, applying a single set of examples across different models reduces their generalization ability. Our findings suggest that optimal prompting for LLMs should include model-specific few-shot examples for best performance. (Details of each prompting performance G and limitation analysis I are provided in the appendix .)

5.2 Error Analysis

We conduct an error analysis for the best-performing method, **Few-shot CoT** with Gemini-2.5-flash, based on its confusion matrix (Figure 2) and per-class precision and recall scores (Table 4):

Overall, the *Positive* class shows both lower precision and recall, with precision being particularly low. This indicates that among the samples predicted as *Positive*, a large portion were actually *Negative*, suggesting a tendency of the LLM to incorrectly predict implicit negative sentiment as positive.

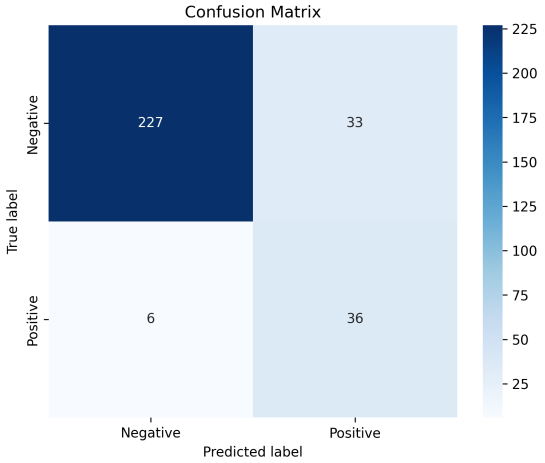

Figure 2: The confusion matrix of Few-shot CoT prompting on Gemini-2.5-flash.

Table 4: Precision and Recall for each sentiment class.

Class	Precision	Recall
Negative	97.32	87.31
Positive	52.17	85.71

False Positives (36 samples). We find that the model struggles with Vietnamese rhetorical and pragmatic cues, often misattributing sentiment by confusing main and background entities. It fails to detect exaggeration and implausibility—key signals of sarcasm—and tends to interpret literal meanings without recognizing implicit emotion or contradiction. This suggests a limited grasp of Vietnamese discourse and pragmatic context. For example, given a review *"Với giá đó thì bạn nên mua cho bản thân, gia đình và nếu có nhiều thì mua luôn cho cả xóm!"*, the model fails to recognize the use of hyperbole ("mua luôn cho cả xóm" / *buy for the whole neighborhood*), which makes the sentence unrealistic and sarcastic. Using an impossible scenario to praise a product ironically reverses the sentiment of the surface-level text.

False Negatives (6 samples). We observe that the model often fails to distinguish between literal and sarcastic expressions, misinterpreting humorous or ironic cues. It also struggles to correctly identify the main sentiment target in sentences with multiple entities, frequently assigning sentiment to secondary or background entities. Additionally, the model lacks contextual and commonsense

Table 2: Performance of the best-performing closed-source LLMs across prompting strategies.

Prompt Strategy	Best LLM	Accuracy	Micro F_1	Macro F_1	Weighted F_1
Zero-shot reasoning role-play	Gemini-2.5-pro	86.09	86.09	74.42	**86.89**
Plan-and-Solve Prompting	GPT-4.1	80.79	80.79	64.08	81.77
Few-shot CoT	Gemini-2.5-flash	**87.09**	**87.09**	**78.48**	85.21
Active Prompting with CoT	GPT-4o	83.77	83.77	68.86	84.42
Zero-shot Reasoning (COSP)	GPT-4.1	69.87	69.87	57.90	74.10

Table 3: Performance of the best-performing open-source LLMs across prompting strategies.

Prompt Strategy	Best LLM	Accuracy	Micro F_1	Macro F_1	Weighted F_1
Zero-shot reasoning role-play	Gemma-3-27b	79.14	79.14	55.09	78.81
Plan-and-Solve Prompting	Deepseek-v3	78.48	78.48	49.36	77.08
Few-shot CoT	Gemma-3-27b	**79.47**	**79.47**	**65.05**	**81.25**
Active Prompting with CoT	Deepseek-v3	**79.47**	**79.47**	63.99	81.03
Zero-shot Reasoning (COSP)	Deepseek-v3	67.22	67.22	55.90	72.03

reasoning, leading to incorrect interpretations of implausible comparisons or indirect expressions. For example, give a review *"Em Note4 của em nó còn chưa chịu hỏng thì bao giờ em mới được dùng Note7(6) đây?!"*. The model fails to detect the ironic rhetorical question ("còn chưa chịu hỏng thì bao giờ mới được dùng" / *still not broken so when can I get a new one?*), which is used humorously to praise the durability of the "Note4", not to express impatience or frustration.

5.3 Improvement Directions

We identify several directions to enhance the performance of the **Few-shot Chain-of-Thought** (Wei et al., 2022) approach for implicit sentiment analysis in Vietnamese. First, constructing or collecting demonstration examples that better reflect Vietnamese pragmatic usage and up-to-date contexts is essential. These examples should include more challenging cases involving multiple sentiment-bearing entities, conflicting emotional cues, exaggeration, and contradictions with commonsense knowledge. Second, we propose integrating ideas from **Active Prompting with Chain-of-Thought** (Wan et al., 2023), where each model dynamically selects its own in-context examples from a candidate pool. This pool can be larger and more diverse, allowing the model to choose examples that best align with its own understanding and reasoning patterns. Finally, the demonstrations can be further improved by providing structured and detailed reasoning steps tailored to implicit sentiment in Vietnamese. Inspired by the **PS Prompting** (Wang et al., 2023) framework, each reasoning step can be accompanied by guiding questions to scaffold the model's inference process more effectively.

6 Conclusion

This paper introduces a benchmark for evaluating large language models on implicit sentiment analysis in Vietnamese, with a focus on rhetorical and pragmatic aspects such as sarcasm, exaggeration, and contextual irony. We explore several state-of-the-art prompting strategies and observe that while LLMs achieve relatively high recall, their precision remains low—mainly due to difficulties in accurately detecting sentiment-bearing targets and interpreting indirect or ambiguous expressions. These challenges suggest that existing prompting methods may not fully account for the linguistic and cultural subtleties of Vietnamese, highlighting the need for more targeted strategies tailored to the nature of implicit sentiment. In future work, we aim to improve model performance by constructing higher-quality in-context examples and designing more structured, reasoning-guided prompts to enhance both accuracy and generalization.

Acknowledgements

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under the grant number NCM2025-26-02.

References

Jin Cui, Fumiyo Fukumoto, Xinfeng Wang, Yoshimi Suzuki, Jiyi Li, and Wanzeng Kong. 2023. Aspect-category enhanced learning with a neural coherence model for implicit sentiment analysis. In *Findings of the ACL: EMNLP 2023*, pages 11345–11358, Singapore.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the ACL*, pages 1330–1350, Bangkok, Thailand.

Xuanwen Ding, Jie Zhou, Liang Dou, Qin Chen, Yuanbin Wu, Arlene Chen, and Liang He. 2024. Boosting large language models with continual learning for aspect-based sentiment analysis. In *Findings of the ACL: EMNLP 2024*, pages 4367–4377.

Zhihua Duan and Jialin Wang. 2024. Implicit sentiment analysis based on chain-of-thought prompting. In *Proceedings of the 2024 7th International Conference on Advanced Algorithms and Control Engineering*, pages 368–371. IEEE.

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 the 61st Annual Meeting of the ACL*, pages 1171–1182, Toronto, Canada.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4099–4113, Mexico City, Mexico.

Thi Thuy Ha Le. 2015. Biểu đạt lịch sự trong hành động ngôn từ phê phán tiếng việt và tiếng anh. *Ngôn ngữ & Đời sống*, (2 (232)):40–44.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169.

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

Thu Hanh Nguyen. 2020. Hành vi ngôn ngữ trách trong tiếng việt. *HNUE Journal of Science, Social Sciences*, 65(8):119–128.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 443–450, Denver, Colorado.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023a. A systematic literature review on vietnamese aspect-based sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–28.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2024. Prompt engineering with large language models for vietnamese sentiment classification. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 181–192, Tokyo, Japan.

Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023b. Vietnamese sentiment analysis: an overview and comparative study of fine-tuning pretrained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–27.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the ACL: ACL 2023*, pages 3493–3514.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the ACL*, pages 2609–2634.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 24824–24837. Article No. 1800.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the ACL: NAACL 2024*, pages 3881–3906, Mexico City, Mexico.

A Prompt for Filtering Implicit Sentiment Examples

Filtering Prompt for Implicit Sentiment Detection

You are a Vietnamese language assistant with expertise in identifying sentences that convey **implicit sentiment**, such as metaphor, sarcasm, or irony (i.e., sentiments not expressed directly but through language, structure, or contextual signals).
Please read the input sentence. If there is strong evidence of implicit sentiment, perform the following three steps:

1. **Original**: Copy the original sentence.
2. **Emotion Feature**: List all matching features from the list below that help identify implicit sentiment.
3. **Explanation**: Provide a detailed explanation of why the sentence likely expresses the implicit sentiment, focusing on subtle tones, ambiguity, abnormal structure, or pragmatic contradiction.

Only mark a sentence **if there is clear evidence** of sarcasm or implicit emotion. Otherwise, return **SKIP**.

Six Features of Implicit Sentiment (Emotion Feature):

- **Contradiction between literal meaning and implied intent**
  - Examples:
    - "Bạn chăm chỉ ghê, cả năm đi học đúng ba ngày."
    - "Trình bày bài làm gọn gàng, toàn bộ trống trơn luôn!"
    - "Bạn thông minh thật, trả lời sai hết tất cả câu hỏi luôn!"
- **Exaggeration, understatement, or sarcastic softening**
  - Examples:
    - "Anh ấy chạy nhanh như rùa bò."
    - "Cậu ấy hát hay đến mức chó cũng tru theo."
    - "Bài này dễ như lên trời vậy đó!"
- **Wordplay or lexical ambiguity**
  - Examples:
    - "Cậu ấy rất có tiềm năng… tiềm mãi mà chưa thấy năng đâu."
    - "Bạn ấy thi tốt thật… tốt nghiệp luôn khỏi phải học lại nữa."
    - "Cậu học lực khá… khá là đáng lo."
- **Rhetorical questions or sarcastic exclamations**
  - Examples:
    - "Bạn làm được thế này mà không thi Nobel à?"
    - "Có ai giỏi như bạn không? Nộp bài trắng tinh như tờ giấy!"
    - "Trời đất, sao bạn siêng thế? Đến lớp một tháng một lần!"
- **Sarcastic tone (spoken language)**
  - Examples:
    - "Woa, tuyệt vờiiii, bạn làm bài mà chỉ sai… hết toàn bộ thôi đó!"
    - "Ôi, giỏi quá haaa, vắng mặt liên tục mà vẫn đậu!"
    - "Cảm ơnnn nhiều nha, nhờ bạn mà nhóm mình bị trừ hết điểm."
- **Contextual irony or paradox**
  - Examples:
    - "May ghê, vừa kịp giờ!" (when arriving 30 minutes late)
    - "Nhờ bạn mà bài nhóm bị điểm 0, tuyệt thật!"
    - "Anh ấy rất chuyên cần – thường đến lớp ngay khi tiết học kết thúc."

Special Notes: Long, playful, or overly complimentary sentences may be sarcastic. Watch for exaggerations, illogical comparisons, or reversed meanings.

Input: "<text>"
If the sentence contains implicit sentiment, respond using the following format:

Original: <original sentence>
Emotion Feature: <list of matching features>
Explanation: <detailed explanation>

If NO implicit sentiment is found, respond with: SKIP

B Prompting Strategy: Zero-shot Role-Play

Step 1: Role-setting prompt The model is instructed to take on the role of a Vietnamese sentiment analysis expert:

Role-setting Prompt

You are a highly sophisticated Vietnamese language expert specializing in sentiment analysis. Your primary task is to accurately classify the sentiment of Vietnamese sentences, with a particular focus on **implicit sentiment**. I am your client who needs sentiment analysis.

Step 2: Role-feedback prompt We record the model's response to the role-setting prompt to confirm role immersion. An example of feedback from the model:

Model's Response (Feedback)

As a Vietnamese sentiment expert, I'm ready to help you detect implicit sentiment. Please send the sentence to analyze.

Step 3: Classification prompt After confirming the role, the model is given a classification task for a Vietnamese sentence:

Zero-shot Role-play Prompt (Full Input)

User (role-setting prompt):
You are a highly sophisticated Vietnamese language expert specializing in sentiment analysis. Your primary task is to accurately classify the sentiment of Vietnamese sentences, with a particular focus on **implicit sentiment**. I am your client who needs sentiment analysis.

Assistant (model response):
Understood. I will help you identify the implicit sentiment in Vietnamese sentences with high accuracy.

User (sentiment question):
Based on your expertise, classify the sentiment of the following Vietnamese text.
Respond with only one of the following words: **Positive**, **Negative**.

Text: "Với giá đó thì bạn nên mua luôn cho cả xóm!"

Sentiment:

C Plan-and-Solve Prompt

Plan-and-Solve Prompt (Implicit Sentiment)

System prompt:
You are an expert in sentiment analysis, focusing on implicit sentiment through nuanced reasoning.

User prompt:

Q: Analyze the implicit sentiment of the following sentence. Classify it as Positive or Negative.
Sentence: "Với giá đó thì bạn nên mua luôn cho cả xóm!"
A: Let's first understand the problem and devise a complete plan with focus on implicit sentiment analysis.
Then, let's carry out the plan and reason step by step. Every step should answer the subquestions:
"What is the relationship between context and emotional words in a sentence?
From that connection, what emotion does the speaker express in the sentence?"
At the end, output only one word and only one word, which must be exactly one of these:
Positive or Negative. No other explanation or text should be returned.

D Few-shot CoT Prompt

Few-shot Chain-of-Thought Prompt (Implicit Sentiment)

System prompt:
You are a sentiment analysis expert for Vietnamese text. Your task is to detect the implicit sentiment in each sentence. Analyse the sentence step by step, then answer with one single word only: Positive or Negative.

Few-shot examples:

Example 1:
Q: Không biết ai thiết kế giao diện này, mà mình mở lên là cái miệng chữ o mắt chữ a luôn á chời
A: Let's think step by step. Diễn tả bất ngờ và ngạc nhiên tích cực khi thấy giao diện đẹp → Sentiment: Positive

Example 2:
Q: Samsung đúng là tệ dễ sợ , luôn đi sau người ta nhưng mà là đi sau của 5-10 năm sau , hahaa
A: Let's think step by step. Dù chê nhưng lại hàm ý khen đi trước thời đại → Sentiment: Positive

Example 3:
Q: Apple đúng là nhàm chán , dậm chân tại chổ mãi , tại chổ top 1.
A: Let's think step by step. Nói nhàm chán nhưng vẫn đứng đầu → Sentiment: Positive

Example 4:
Q: Hi. Mình cũng đang hóng Note 6 ra để mua ... Note 4, hy vọng lúc đó Note 4 sẽ còn khỏang 8t. Note 6 thì quá đỉnh, mình không chống nổi.
A: Let's think step by step. Giả vờ khen Note 6 nhưng thật ra chỉ chờ mua Note 4 giá rẻ → Sentiment: Negative

Example 5:
Q: Các hãng đt tq bây h cho ra đời những chiếc đt rất ấn tượng nhưng Oppo k nằm trọng số đó
A: Let's think step by step. Nêu tổng thể tốt nhưng riêng Oppo bị loại trừ → Sentiment: Negative

Example 6:
Q: Cái này có trên Nokia 1100 từ chục năm trước rồi.
A: Let's think step by step. Chê tính năng lỗi thời, không mới mẻ → Sentiment: Negative

Prompt for inference:

Now analyze the following sentence. Only respond with one word: Positive or Negative.

Q: {text}
A: Let's think step by step.

E Implicit Sentiment Examples for Few-shot CoT

Sentence	Sentiment
Không biết ai thiết kế giao diện này, mà mình mở lên là cái miệng chữ o mắt chữ a luôn á chời	Positive
Lúc vừa định hỏi tới đâu rồi thì quay qua thấy hàng đã hiện ở cửa như một tia chớp.	Positive
Samsung đúng là tệ dễ sợ, luôn đi sau người ta nhưng mà là đi sau của 5-10 năm sau, hahaa	Positive
Haizz, chỉ cần 10 năm nữa là apple sẽ bị các hãng khác đuổi kịp mất, lo ghê há.	Positive
Ừ con điện thoại ấy cũng bình thường, tầm tầm thôi à, mà thường top 1 và tầm 10 điểm chứ gì.	Positive
Iphone hay bị chê ghê há, nhưng mà tới lúc bán thì thi nhau xếp hàng chen chút mua, rồi chê dữ rồi há.	Positive
Cái app này khiến mình đắn đo khi phải kiếm app khác xóa đi để trống bộ nhớ mà tải nó.	Positive
Đang loay hoay cài app thì nhân viên tới giúp ngay mà không cần phải gọi gì.	Positive
Cuối cùng thì nhà vua cũng đã trở lại, ngai vàng vẫn sẽ thuộc về apple.	Positive
Đúng là xiaomi chả có gì ngoài cấu hình, pin, camera, màn hình đứng top 1.	Positive
Trong khi mọi người phải vác theo mấy cục sạc dự phòng nặng nề thì tiếc quá hai ngày rồi mình vẫn chưa cần dùng tới sạc.	Positive
Apple đúng là nhàm chán, dậm chân tại chỗ mãi, tại chỗ top 1.	Positive
Cô giáo không biết có bán thuốc không sao cả lớp như uống thuốc ngủ á.	Negative
Các hãng điện thoại Trung Quốc bây giờ cho ra đời những chiếc điện thoại rất ấn tượng nhưng Oppo không nằm trong số đó.	Negative
Đọc thấy rất hứng thú nhưng khi đến đoạn 'bộ nhớ trong 32 GB' tôi không đọc nữa.	Negative
iPad Pro thật tuyệt vời, để mình thêm 9 triệu nữa đi mua Surface Pro 4 mới được xách tay về vậy!!!	Negative
10 năm một thiết kế, thật là đột phá công nghệ.	Negative
Hi. Mình cũng đang hóng Note 6 ra để mua ... Note 4, hy vọng lúc đó Note 4 sẽ còn khoảng 8 triệu. Note 6 thì quá đỉnh, mình không chống nổi.	Negative
Z5 dùng chip 810 đã thấy nóng khủng khiếp, giờ lên 820 chắc thành cái chảo nướng.	Negative
Thấy chip của MediaTek là sợ lắm... rồi.	Negative
Cái này có trên Nokia 1100 từ chục năm trước rồi.	Negative
Dự là màu xanh sẽ cháy hàng bởi vì nếu không xài màu xanh thì chẳng ai biết bạn đang dùng iPhone7 cả.	Negative
Ừ máy đẹp ghê đó nhưng nhường phần mấy bạn.	Negative
SE gì? Ai Phôn Sẽ "Ế" đúng hơn. haha	Negative

F Zero-shot Self-Adaptive Prompting (CoSP) prompt

Stage 1: Prompt for Generating Multiple Reasoning Paths

Chain-of-Thought Prompt (Stage 1)

Question: <Vietnamese sentence>
Think step-by-step and then give the final sentiment prediction (Positive or Negative):

Stage 2: Prompt for Final Prediction using In-Context Demonstrations

Final CoT Prompt with Selected Demos (Stage 2)

Question: <Demo 1>
Reasoning: <Demo 1 reasoning>
Answer: Positive/Negative

Question: <Demo 2>
Reasoning: <Demo 2 reasoning>
Answer: Positive/Negative

...

Question: <New sentence>
Let's think step-by-step and then give the final sentiment prediction (Positive or Negative):

G Model Performance on Prompting Strategies

Table 5: Performance of various models under Role-Play Prompting.

Model	Accuracy	Micro-F1	Macro-F1	F1-weighted
FPT.AI-KIE-v1.7	0.6788	0.6788	0.5049	0.7167
SaoLa-3.1-medium	0.6788	0.6788	0.5324	0.7213
LLaMA-3.3-Swallow-70B-Instruct-v0.4	0.7152	0.7152	0.5220	0.7414
LLaMA-3.3-70B-Instruct	0.7583	0.7583	0.5914	0.7799
LLaMA-4-Scout-14B-16E	0.7649	0.7649	0.5322	0.7704
LLaMA-3.1-8B-Instruct	0.7715	0.7715	0.5276	0.7726
Deepseek-v3-0324	0.7848	0.7848	0.4936	0.7708
Gemma-3-27B-it	0.7914	0.7914	0.5509	0.7881
Gemma-3-12B-it	0.7980	0.7980	0.5558	0.7926
GPT-4o-mini	0.7616	0.7616	0.5022	0.7616
GPT-4o	0.7616	0.7616	0.5818	0.7797
GPT-4.1-mini	0.7649	0.7649	0.5488	0.7742
GPT-4.1	0.8079	0.8079	0.6214	0.8132
Gemini-2.0-flash	0.8179	0.8179	0.6079	0.8150
Gemini-2.5-flash	0.8377	0.8377	0.6991	0.8465
Gemini-2.0-flash-lite	0.8444	0.8444	0.5715	0.8183
Gemini-2.5-pro	**0.8609**	**0.8609**	**0.7442**	**0.8689**

Table 6: Performance of various models using Plan-and-Solve Prompting.

Model	Accuracy	Micro-F1	Macro-F1	F1-weighted
SaoLa-3.1-medium	0.6523	0.6523	0.3610	0.7044
LLaMA-3.3-Swallow-70B-Instruct-v0.4	0.6854	0.6854	0.5091	0.7215
LLaMA-4-Scout-17B-16E	0.7020	0.7020	0.5539	0.7394
LLaMA-3.1-8B-Instruct	0.6821	0.6821	0.4810	0.7142
LLaMA-3.3-70B-Instruct	0.7086	0.7086	0.5587	0.7444
Qwen3-32B	0.7086	0.7086	0.5481	0.7425
Qwen3-235B-A22B	0.7550	0.7550	0.4031	0.7810
Gemma-3-12B-it	0.6722	0.6722	0.3067	0.7102
Gemma-3-27B-it	0.7616	0.7616	0.6054	0.7846
Gemini-2.0-flash	0.7517	0.7517	0.5802	0.7739
Gemini-2.0-flash-lite	0.7318	0.7318	0.3775	0.7598
Gemini-2.5-pro-preview-03-25	0.7748	0.7748	0.6550	0.8018
Gemini-2.5-flash	0.7815	0.7815	**0.6729**	0.8089
Deepseek-v3-0324	0.7848	0.7848	0.4936	0.7708
GPT-4.1-mini	0.7119	0.7119	0.5806	0.7500
GPT-4o-mini	0.7417	0.7417	0.3453	0.7551
GPT-4o	0.7947	0.7947	0.6224	0.8065
GPT-4.1	**0.8079**	**0.8079**	0.6408	**0.8177**

Table 7: Performance of various models under Few-shot-CoT Prompting.

Model	Accuracy	Micro-F1	Macro-F1	F1-weighted
Qwen3-235B-A22B	0.3344	0.3344	0.3306	0.3674
Qwen3-32B	0.4106	0.4106	0.3896	0.4713
FPT.AI-KIE-v1.7	0.6424	0.6424	0.5028	0.6930
LLaMA-3.1-8B-Instruct	0.6523	0.6523	0.5280	0.7029
LLaMA-4-Scout-17B-16E	0.7119	0.7119	0.6052	0.7534
LLaMA-3.3-Swallow-70B-Instruct-v0.4	0.7252	0.7252	0.5601	0.7546
LLaMA-3.3-70B-Instruct	0.7351	0.7351	0.6420	0.7738
SaoLa-3.1-medium	0.7682	0.7682	0.6054	0.7884
Deepseek-v3-0324	0.7881	0.7881	0.6492	0.8085
Gemma-3-12B-it	0.7119	0.7119	0.3700	0.7515
Gemma-3-27B-it	0.7947	0.7947	0.6505	0.8125
GPT-4.1-mini	0.7649	0.7649	0.6460	0.7941
GPT-4o-mini	0.7781	0.7781	0.5886	0.7902
GPT-4o	0.8146	0.8146	0.6796	0.8297
GPT-4.1	0.8576	0.8576	0.7593	**0.8703**
Gemini-2.0-flash-lite	0.8079	0.8079	0.4201	0.8154
Gemini-2.5-pro	0.8311	0.8311	0.7458	0.8521
Gemini-2.0-flash	0.8510	0.8510	0.7140	0.8569
Gemini-2.5-flash	**0.8709**	**0.8709**	**0.7848**	0.8521

Table 8: Performance of models using Active Prompting with Chain-of-Thought.

Model	Accuracy	Micro-F1	Macro-F1	F1-weighted
Qwen3-235B-A22B	0.3874	0.3874	0.3735	0.4409
Qwen3-32B	0.4570	0.4570	0.4242	0.5233
FPT.AI-KIE-v1.7	0.6656	0.6656	0.5131	0.7098
LLaMA-3.3-70B-Instruct	0.7252	0.7252	0.6234	0.7647
LLaMA-4-Scout-17B-16E	0.7285	0.7285	0.6109	0.7653
LLaMA-3.1-8B-Instruct	0.7384	0.7384	0.3681	0.7619
LLaMA-3.3-Swallow-70B-Instruct-v0.4	0.7550	0.7550	0.5702	0.7736
SaoLa-3.1-medium	0.7550	0.7550	0.6289	0.7850
Gemma-3-12B-it	0.7450	0.7450	0.5627	0.7665
Gemma-3-27B-it	0.7781	0.7781	0.6079	0.7944
Deepseek-v3-0324	0.7947	0.7947	0.6399	0.8103
Gemini-2.5-pro-preview-03-25	0.7848	0.7848	0.4844	0.8266
Gemini-2.0-flash-lite	0.8146	0.8146	0.6276	0.8181
Gemini-2.5-flash-preview-04-17	0.8278	0.8278	**0.7329**	**0.8478**
Gemini-2.0-flash	0.8311	0.8311	0.6968	0.8425
GPT-4o-mini	0.7715	0.7715	0.3650	0.7786
GPT-4.1-mini	0.7748	0.7748	0.6372	0.7985
GPT-4.1	0.8245	0.8245	0.6990	0.8393
GPT-4o	**0.8377**	**0.8377**	0.6886	0.8442

Table 9: Performance of various models using Zero-shot Self-Adaptive Prompting.

Model	Accuracy	Micro-F1	Macro-F1	F1-weighted
Deepseek-v3-0324	0.6722	0.6722	0.5590	0.7203
GPT-4o	0.6358	0.6358	0.5396	0.6915
GPT-4.1-mini	0.6523	0.6523	0.5482	0.7048
GPT-4o-mini	0.6623	0.6623	0.5436	0.7116
GPT-4.1	**0.6987**	**0.6987**	**0.5790**	**0.7410**

H Implementation of Prompting Strategies

Zero-shot Reasoning with Role-play

This method is implemented via a two-turn prompting scheme. First, a role-establishing prompt instructs the LLM to act as a Vietnamese language expert in implicit sentiment analysis; the model's role-confirmation response is recorded. Then, a task prompt asks the LLM to classify a target sentence as either *Positive* or *Negative*, within its assumed role. The full input includes the role prompt, the role-confirmation, and the classification prompt. This strategy could not be applied to Qwen3 models due to insufficient Vietnamese training data.

Plan-and-Solve (PS) Prompting

PS prompting uses a two-part input: (1) a task prompt presenting the classification goal and the target sentence, and (2) a PS prompt instructing the LLM to first "devise a complete plan"and then "carry out the plan"through structured multi-step reasoning with intermediate questions. This method is incompatible with FPT.AI-KIE-v1.7, which lacks multi-step reasoning ability without in-context examples.

Few-shot Chain-of-Thought (Few-shot CoT)

Few-shot CoT prompting constructs prompts that guide the LLM to reason step-by-step using in-context examples. The LLM is assigned the role of a Vietnamese sentiment analysis expert. Three examples per class (Positive, Negative) are randomly sampled from a labeled pool, each accompanied by a brief explanation of the reasoning process behind the sentiment label.

Active Prompting with Chain-of-Thought

This strategy builds on Few-shot CoT but improves example selection. Instead of sampling randomly, candidate examples are scored by entropy after 10 rounds of zero-shot inference. The most uncertain samples per label (highest entropy) are selected. To reduce cost, GPT-4o is used as the selector, and the selected examples are reused across smaller models.

Zero-shot Reasoning with Self-Adaptive Prompting (COSP)

COSP is implemented as a two-stage prompting process. In Stage 1, the model performs zero-shot CoT prompting, generating seven reasoning paths. The top five are selected based on a scoring function $F_p = \text{Normalized Entropy} + \lambda \times \text{Repetitiveness}$. In Stage 2, these paths serve as few-shot demonstrations to regenerate seven new paths. The final label is chosen via majority vote. Due to high computational cost, this method is applied only to GPT-family and Deepseek-v3 models.

I Limitations of Prompting Strategies

Zero-shot Reasoning with Role-Play

This strategy guides the model to adopt a predefined role (e.g., teacher or language assistant) in a zero-shot setting. Its effectiveness relies heavily on how clearly the role is defined and whether the model is trained on the target language (e.g., Vietnamese). While it outperforms naive zero-shot prompting, the lack of multi-step reasoning limits its ability to handle complex or subtle expressions, especially those involving sarcasm or implicit cues.

Plan-and-Solve (PS) Prompting

PS prompting structures reasoning by prompting the model to first generate a plan, then follow a sequence of task-specific questions. This approach encourages logical progression and increases interpretability. While it improves over simple zero-shot reasoning, its performance depends on how well the guided steps match the linguistic and contextual complexity of the task, which may vary across domains and languages.

Few-shot Chain-of-Thought (Few-shot CoT)

Few-shot CoT enhances reasoning by including in-context examples with step-by-step explanations. It is especially useful in low-resource settings like Vietnamese. However, the model's success hinges on the quality and contextual fit of the examples. Inappropriate examples may mislead the model or fail to generalize, particularly in the presence of sarcasm, irony, or indirect sentiment cues.

Active Prompting with Chain-of-Thought

This method extends Few-shot CoT by selecting examples based on uncertainty (entropy) measured through multiple zero-shot runs. The most uncertain cases are chosen to improve robustness. While promising, this approach is often model-specific: examples selected for one LLM (e.g., GPT-4o) may not transfer effectively to others due to architectural and training differences.

Self-Adaptive Prompting (COSP)

COSP performs two stages of reasoning: it first generates multiple paths per input and then uses selected ones as few-shot demonstrations in a second round. Despite its innovative framework, COSP is computationally intensive and yielded the lowest performance in our experiments. Its main limitation lies in relying solely on self-generated demonstrations, which often lack the nuance and contextual grounding provided by carefully curated human examples.