mucAI at Ahasis Shared Task: Sentiment Analysis with Adaptive Few Shot Prompting

Ahmed Abdou
Technical University of Munich
ahmed.abdou@dtum.de

Abstract

Sentiment Analysis is a crucial task in Natural Language Processing (NLP) focused on identifying and categorizing emotional tones or opinions within text. For Arabic customer reviews, sentiment analysis is particularly challenging. The language's rich diversity, with numerous regional dialects differing significantly from Modern Standard Arabic (MSA) and each other in lexicon, syntax, and sentiment expression, complicates consistent performance across dialects. In this paper, we present our approach, submitted to the AHASIS Shared Task 2025, focusing on sentiment analysis for Arabic dialects in the hotel domain. Our method leverages the capabilities of GPT-4o through adaptive few-shot prompting technique, where similar contextual examples are dynamically selected for each review using a k-Nearest Neighbors (kNN) search over train embeddings from a fine-tuned encoder model. This approach tailors the prompt to each specific instance, enhancing classification performance over minority class. Our submission achieved an F1-score of 76.0% on the official test set, showing stronger performance for the Saudi dialect compared to Darija.

1 Introduction

Sentiment analysis for the Arabic language presents unique challenges due to its complex linguistic landscape. Unlike languages with more homogeneous structures, Arabic encompasses Modern Standard Arabic (MSA) and numerous regional dialects that differ in syntax, lexicon, morphology, and semantic expressions. These variations become particularly pronounced when analyzing sentiment in domain-specific contexts, such as hotel reviews, where emotional expressions and idiomatic phrases can vary significantly across dialectal boundaries. In this paper, we tackle the AHASIS shared task on sentiment analysis on arabic dialects (Saudi and Darija) in the hospitality domain.

Large Language Models (LLMs) exhibit impressive in-context learning (ICL) abilities: with a handful of demonstrations in the prompt they adapt to new tasks on the fly (Brown et al., 2020). Yet a growing body of work shows that ICL is highly sensitive to which and how many examples are shown (Yoshida, 2024). Small, static prompts amplify demonstration bias: models over-predict labels that dominate the prompt or appear later in the example list, harming minority classes. Traditional approaches have often relied on fine-tuning pre-trained transformer models specific to Arabic, such as AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021), which have shown considerable success. However, the advent of LLMs like has opened new frontiers, offering powerful generative and reasoning capabilities. For example, using LLMs we can get the generated tokens by the model justifying the given classification label (Huang et al., 2023).

Motivated by these findings, we explore the use of GPT-4o (Hurst et al., 2024) for the AHASIS shared task. We evaluate three prompting strategies: (i) zero-shot prompting, where no examples are provided; (ii) static few-shot prompting, using a fixed set of manually curated demonstrations; and (iii) adaptive few-shot prompting, where examples are dynamically retrieved from the training set based on semantic similarity using AraBERT embeddings. Importantly, each retrieved example is paired with a GPT-4o-generated chain of thought conditioned on its gold label. For comparison, we also fine-tune AraBERT and MARBERT as encoder-based baselines. Our key contribution is the adaptive few-shot strategy, which consistently outperformed both static and zero-shot prompting on the test set. By tailoring examples to each input, this method offers a simple yet effective way to improve LLM performance in multilingual and

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 54–61, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-009

dialectal sentiment tasks.

2 Related Work

LLMs have recently approached or matched the performance of supervised task models without gradient updates. A systematic evaluation (Zhang et al., 2024) across 26 datasets showed GPT-3.5/4 and Llama-2 (Touvron et al., 2023) within 1–3 macro-F1 of fine-tuned RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020) baselines on sentence- and aspect-level sentiment. Building on this, several studies have evaluated general-purpose LLMs on Arabic sentiment tasks. Taqyim (Alyafeai et al., 2023) benchmarked GPT-3.5 and GPT-4 across seven Arabic datasets and found that GPT-4 narrows but does not eliminate the gap to supervised MARBERT (Abdul-Mageed et al., 2021). GPTAraEval (Khondaker et al., 2023) evaluated ChatGPT [1] on 44 Arabic tasks and reported consistent under-performance relative to smaller, Arabic-tuned pre-trained language models (PLMs). Moreover, in (Al-Thubaity et al., 2023), the authors showed that GPT-4 in a 5-shot setting reaches AraBERT (Antoun et al., 2020) performance, whereas GPT-3.5 and Google Bard (PaLM 2) (Anil et al., 2023) lag behind. A comparative evaluation (Alharbi et al., 2025b) of DeepSeek-R1 (Guo et al., 2025), Qwen2.5 (Qwen et al., 2025), and LLaMA-3 (Grattafiori et al., 2024) further demonstrates the efficacy of dialect-specific prompting and parameter-efficient fine-tuning (LoRA) (Hu et al., 2022) in Arabic sentiment analysis, showing that prompt-input alignment significantly enhances performance, especially for underrepresented dialects.

kNN-Prompting (Xu et al., 2023) embeds all training instances once and predicts each test query by a simple majority vote over its k closest neighbours, thus bypassing context-length limits and heavy calibration steps. kNN-ICL (Zhao et al., 2023) refines this idea by selecting similarity-based demonstrations on the fly, yielding consistent F1 gains under strict token budgets. The closest antecedent to our pipeline is OpenMedLM (Maharjan et al., 2024), which retrieves the five most similar patient questions, asks an LLM to write gold-label-conditioned rationales once, caches them, and inserts the triple (question, chain-of-thoughts (CoT), answer) at inference—achieving state-of-the-art accuracy on medical QA without fine-tuning.

[1] https://openai.com/index/chatgpt/

3 Task and Data

The AHASIS shared task (Alharbi et al., 2025a) involves sentiment classification of hotel reviews, specifically targeting Arabic dialects—Saudi and Darija. Each review in the dataset is annotated with two labels: the dialect (Saudi or Darija) and the sentiment expressed (positive, negative, or neutral). The primary objective of the task is to predict the sentiment label given a sentence in one of these dialects.

The provided training dataset consists of 860 sentences, evenly split with 430 sentences for each dialect. Within each dialect subset, there are 154 positive, 168 negative, and 108 neutral sentiment sentences. To conduct our experiments, we stratified the training data based on both sentiment and dialect, splitting it into 80% training and 20% development subsets obtaining 688 sentences for training and 172 sentences for development. Additionally, a separate test set comprising 216 sentences was provided, with annotations specifying only the dialect and excluding sentiment labels.

4 Method

Building on insights from prior work on retrieval-augmented prompting and chain-of-thought (CoT) reasoning, we introduce a dynamic kNN + label-conditioned CoT framework. Unlike static prompting methods, which rely on fixed demonstrations and a global decision boundary, our approach constructs an adaptive, query-specific prompt at inference time.

For each input review, we (i) embed it using a pre-trained Arabic encoder, (ii) retrieve its k nearest dialect-balanced neighbours from a cached set of labelled hotel reviews, (iii) generate a rationale for each neighbour using GPT-4o, conditioned on its gold label, and (iv) assemble these (review, label, CoT) triplets into the prompt.

We evaluate our approach alongside two encoder-based baselines (AraBERT and MARBERT), and compare three prompting strategies using GPT-4o: zero-shot, static few-shot, and our proposed adaptive few-shot prompting. All models are tested on a shared train/dev/test split to ensure comparability across settings.

4.1 Encoder-Based Baselines

We fine-tune two established Arabic transformer models, **AraBERT** and **MARBERT**, as classification baselines. These serve as strong nongenerative benchmarks for comparison with GPT-4o-based approaches.

4.2 Zero-Shot and Static Few-Shot Prompting

In the zero-shot setup, GPT-4o receives only a system instruction in Arabic asking it to: (1) assign one of the three sentiment labels (positive, neutral, negative), and (2) generate a brief natural language justification for its prediction.

For the static few-shot setting, we prepend a fixed set of curated examples—each containing an input review, its gold label, and a justification—to the prompt. This improves performance but has limited flexibility: when test instances deviate aspect-wise from the static examples, performance degrades. Additionally, expanding the static prompt to cover more cases is costly and difficult to maintain.

4.3 Adaptive Few-Shot Prompting

To address the limitations of static prompting, we propose an **adaptive few-shot strategy** that builds a tailored prompt for each input review. The process consists of two stages:

(1) Retrieval: For a given input review x, we compute its embedding using a pre-trained AraBERT encoder and retrieve the $k = 20$ nearest neighbors from the training set. These are then stratified by sentiment label, and we select the top $n = 3$ examples per class (if available), yielding up to 9 demonstrations. This approach balances semantic similarity and aspect alignment (e.g., topic or focus of the review) with label diversity.

(2) Prompt Construction with CoT: Each retrieved example is paired with a chain-of-thought (CoT) justification generated by GPT-4o, conditioned on its gold label (see prompt in Table 8). These structured examples—(review, label, justification)—are inserted into the prompt, followed by the target review x to be classified. This dynamic prompt ensures that each input is evaluated in a context shaped by semantically and topic relevant reasoning chains.

5 Results

In this section, we present the experimental results obtained across different prompting strategies using GPT-4o, alongside baseline results from the encoder models AraBERT and MARBERT. Our primary metric for evaluation is the macro-F1 score. Table 1 summarizes the overall results on both the development and test sets. The code used for the experiments is available on GitHub [2].

Model / Method	Dev	Test
AraBERT	83.7	73
MARBERT	85.2	73
Zero-Shot	80.5	75
Few-Shot	82.5	74
Adaptive Few-Shot	**84.5**	**76**

Table 1: Macro-F1 scores for sentiment classification across different experimental setups.

Table 2: Class level F1-score for Saudi dialect sentences

Prompting Strategy	Neg	Neu	Pos	Macro
Zero-Shot	96	70	84	83
Few-Shot	97	78	88	88
Adaptive Few-Shot	**97**	**81**	**89**	**89**

Table 3: Class level F1-score for Darija dialect sentences

Prompting Strategy	Neg	Neu	Pos	Macro
Zero-Shot	88	63	82	78
Few-Shot	89	62	83	78
Adaptive Few-Shot	**90**	**67**	**84**	**80**

Our adaptive few-shot method achieved the highest test set performance among the GPT-4o prompting techniques, with an F1-score of 76%. This placed our method sixth in the AHASIS shared task leaderboard. Interestingly, although MARBERT achieved the highest development set score of 85.2%, its performance dropped noticeably to 73% on the test set, similar to AraBERT.

Tables 2 and 3 provides a detailed breakdown of F1-scores for each sentiment class (Negative, Neutral, Positive) and the overall Macro F1-score, comparing our three prompting strategies across the Saudi and Darija dialects on the development set. Consistent with the overall test set performance noted in Table 1, the Adaptive Few-Shot strategy yielded the highest Macro F1-scores for both dialects: 89 for Saudi and 80 for Darija.

A key objective of the adaptive strategy was to address challenges with the neutral class. For the Saudi dialect, the Neutral F1-score improved substantially from 70 (Zero-shot) and 78 (Few-shot) to 81 with Adaptive Few-Shot. Similarly, for the Darija dialect, Adaptive Few-Shot improved the Neutral F1-score to 67 compared to (Zero-shot) 63

[2] https://github.com/AhmedAbdel-Aal/Ahasis_shared_task

and (Few-Shot) 62. Performance on the Negative and Positive classes was generally strong across all methods, particularly for the Saudi dialect where Zero-shot already achieved F1-scores of 96 (Negative) and 84 (Positive). The Adaptive Few-Shot method largely maintained or slightly enhanced these high scores while making its most significant impact on the Neutral class. Interestingly, the impact of the standard Few-Shot prompting varied by dialect. While it offered clear improvements for the Saudi dialect, increasing the Macro F1 from 83 (Zero-shot) to 88, it provided no overall benefit for the Darija dialect, where the Macro F1 remained at 78, and the Neutral F1-score even saw a slight decrease.

Comparing the two dialects, the overall F1-scores for Darija were consistently slightly lower than those for Saudi across all prompting methods. For instance, with Adaptive Few-Shot, the Macro F1 was 89 for Saudi versus 80 for Darija.

6 Discussion

The superior performance of the adaptive few-shot strategy can be attributed to its dynamic, instance-specific contextualization. Traditional fine-tuning creates a static, global decision boundary that applies uniformly across all test inputs, which can be suboptimal for linguistically diverse or ambiguous cases. Similarly, static few-shot prompting relies on a fixed set of demonstrations that may not align well with the semantics of a given test instance, limiting their ability to guide the model effectively. In contrast, our adaptive few-shot approach constructs query-specific prompt for each input. This adaptation enables the model to better capture subtle distinctions—particularly near the boundary between neutral and positive sentiment—by grounding its reasoning in semantically relevant examples. In effect, improving precision and recall for neutral class, see Table 4.

Table 4: Precision (P) and Recall (R) per sentiment class (%) across prompting strategies.

Strategy	**Negative**		**Neutral**		**Positive**	
	P	R	P	R	P	R
Zero-Shot	86	99	78	58	83	84
Few-Shot	89	97	81	60	82	89
Adaptive Few-Shot	89	99	85	65	85	89

More precisely, under zero-shot prompting, the neutral class achieved a precision of 78%, indicating that when the model predicted neutral, it was often correct. However, the low recall of 58% resulted in a modest F1-score of 67%. This suggests that the model relied on a general understanding of what counts as "neutral" in language, rather than learning how neutrality is defined in this specific dataset. As a result, it often misclassified factual or mildly opinionated reviews as positive (see Table 6 for an example).

Introducing static few-shot prompting led to moderate improvements. For the neutral class, precision improved slightly to 81%, and recall increased marginally to 60%, resulting in an F1-score of 69%. The modest gain in recall suggests that manually selected examples help the model better recognize prototypical neutral instances, but may still fall short in capturing the full diversity of this class. The most significant gains came from adaptive few-shot prompting, where performance improved across all classes. The neutral class in particular benefited, with precision rising to 85% and recall improving to 65%, leading to its highest F1-score of 74%. This shows that providing contextually relevant examples helped the model better handle ambiguity and make more consistent decisions. Compared to static prompts, the adaptive strategy offered examples that were closer in meaning and tone to the input, which helped the model better understand what neutrality looks like in this dataset. Reviews that contained mild opinions or balanced descriptions—previously misclassified as positive—were more often labeled correctly. This suggests that dynamic prompting helped the model adjust its decision boundary more accurately around the neutral class (see Table 7 for an example).

Given the computational overhead introduced by adaptive few-shot prompting, we analyze its runtime and token-level cost in practice. The final classification prompt includes 3 to 9 demonstrations per test instance, averaging 778 input and 239 output tokens. This range results from selecting up to 3 examples per sentiment class (negative, neutral, positive) from the 20 nearest neighbors; if the retrieved set lacks class diversity, fewer than 9 examples are included. These demonstrations are generated using justification generation prompt shown in Table 5, with an average of 95 input and 219 output tokens each. We cache all intermediate generations, so each training example is used for CoT generation at most once. In total, each test instance, on average, incurs one main LLM call (778 in + 239 out) and 3 to 9 smaller calls (95 in + 219 out each).

7 Limitations and Future Work

The effectiveness of the adaptive few-shot strategy is closely tied to the capabilities of the underlying language model (GPT-4o) and the quality of the encoder used for kNN retrieval. Limitations in the LLM's understanding of specific dialects, or in the encoder's ability to generate semantically meaningful embeddings, may propagate through the retrieval process and affect final predictions. In this work, the encoder choice was not extensively optimized, leaving room for improvement in retrieval quality. Additionally, we used semantic similarity as a proxy for selecting aspect-aligned reviews, but this approximation may not always capture the most relevant examples for each input—especially when sentiment is conveyed through subtle tone, emphasis, or implied preferences rather than explicit aspect terms. Our experiments were conducted exclusively with GPT-4o, and it remains unclear how well the observed improvements from adaptive prompting would transfer to other large language models. In addition, we did not perform extensive hyperparameter tuning or architecture exploration for the encoder-based baselines (AraBERT and MARBERT). These models were fine-tuned with standard settings to provide a comparative reference, but stronger results might be achievable with more targeted optimization. Finally, key components of our system—such as the number of neighbors (k), the number of few-shot examples, and the structure of the prompts—were selected based on preliminary experiments rather than exhaustive tuning. A more systematic exploration of hyperparameters and prompt formats could further enhance performance and provide insight into robustness and generalization.

An important future direction is to evaluate the impact of chain-of-thought generation by comparing our approach to a version of adaptive few-shot prompting that uses retrieved examples without CoT. This would help isolate the contribution of the reasoning component and better understand its role in guiding sentiment classification. Moreover, our experiments were limited to GPT-4o; testing this approach with open-source Arabic LLMs would provide insight into its generalizability and practicality in low-resource or non-proprietary settings.

8 Conclusion

In this shared task, we tackled sentiment analysis for Arabic dialects within the hotel domain, focusing on the AHASIS 2025 dataset. Our approach centered on inspecting the capabilities of the LLMs, specifically GPT-4o, through various prompting strategies. We introduced an adaptive few-shot prompting technique, where during inference, we dynamically selected relevant contextual examples for each review by performing a kNN search over cached embeddings from the training set, which were generated by a fine-tuned Arabert model. This approach aimed to improve generalization and address challenges like the neutral class by providing similar aspects, relevant context for each specific instance. The results demonstrate a clear improvement with the incorporation of our adaptive few-shot prompting. Specifically, on the test set, the Macro-F1 score achieved was 76.0% with the adaptive approach, surpassing the 75.0% from zero-shot and 74.0% from static few-shot prompting, as well as our fine-tuned encoder baselines which scored 73.0%. Our submission secured 6th place in the shared task leaderboard.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proc. ACL*, pages 7088–7105.

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. 2023. Evaluating ChatGPT and bard AI on Arabic sentiment analysis. In *Proceedings of ArabicNLP 2023*, pages 335–349, Singapore (Hybrid). Association for Computational Linguistics.

Maram Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025a. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Maram Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025b. Evaluating large language models on arabic dialect sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Zaid Alyafeai, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel.

2023. Taqyim: Evaluating arabic nlp tasks using chatgpt models.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proc. OSACT*, pages 9–15.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023. $k$nn prompting: Beyond-context learning with calibration-free nearest-neighbour inference. In *Proc. ICLR*.

Lui Yoshida. 2024. The impact of example selection in few-shot prompting on automated essay scoring using gpt models. In *International Conference on Artificial Intelligence in Education*, pages 61–73. Springer.

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 Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenting Zhao, Ye Liu, Yao Wan, Yibo Wang, Qingyang Wu, Zhongfen Deng, Jiangshu Du, Shuaiqi Liu, Yunlong Xu, and Philip S Yu. 2023. Knn-icl: Compositional task-oriented parsing generalization with nearest neighbor in-context learning. *arXiv preprint arXiv:2312.10771*.

Appendix: Prompts and Model Outputs

A.1 Justification Generation Prompt

Table 5: Prompt used to generate chain-of-thought (CoT) justifications for each retrieved review in the adaptive few-shot setting.

Justification Generation Prompt

```
Question: Justify the sentiment of the following
Saudi dialect hotel review.
Review: {review}
Label: {label}
Return your response in the following XML format:
<analysis>
 <overall_justification> your step-by-step
analysis </overall_justification>
</analysis>
```

A.2 Zero-Shot and Adaptive Few-Shot Prompt Examples

We show the GPT-4o outputs for the same development review under two prompting strategies: zero-shot and adaptive few-shot. The overall prompt structure remains consistent across all settings; the only difference lies in the addition of review examples in the few-shot and adaptive versions. In this case, the zero-shot prompt leads to an incorrect prediction, while the adaptive few-shot prompt produces the correct label. This illustrates how retrieved examples—selected based on both semantic similarity and aspect alignment—help the model better calibrate its decision boundary, particularly in distinguishing between neutral and positive sentiment.

Table 6: Example of zero-shot prompting for sentiment classification in the Saudi dialect.

Prompt

Question: what is the sentiment of the following hotel review in Saudi dialect?

Review: رخيص وفي احسن موقع حصلت خصم لاني اشتغل في جهة تابعة القطاع الخاص و لكن حتى الاسعار العادية بدون الخصم رخيصة بالنسبة للموقع
Return your response in the following XML format:
<analysis>
<overall_sentiment> positive/neutral/negative </overall_sentiment>
<overall_justification> your step-by-step analysis </overall_justification>
</analysis>

GPT-4o Output

<analysis>
<overall_sentiment> positive </overall_sentiment>
<overall_justification>
The review highlights several positive aspects of the hotel. Firstly, the reviewer mentions that the hotel is رخيص (cheap), which is generally a positive attribute for budget-conscious travelers. Secondly, the location is described as في احسن موقع (in the best location), indicating a favorable opinion. The reviewer also notes receiving a discount, and even mentions that regular prices are cheap for the location. Overall, the review expresses satisfaction with pricing and location.
</overall_justification>
</analysis>

True Label: Neutral
Predicted Label (Zero-Shot): Positive (Incorrect)

Table 7: Example of adaptive few-shot prompting for sentiment classification in the Saudi dialect. For brevity, only one retrieved example is shown.

Prompt (Excerpt)

Question: what is the sentiment of the following hotel review in Saudi dialect?
Examples:

Review: جودة بالنسبة للسعر بالنسبة لمدينة الاسكندرية من افضل الفنادق بهذا السعر و السعر مناسب
<analysis>
<overall_sentiment> Neutral </overall_sentiment>
<overall_justification>

1. The review states جودة بالنسبة للسعر which translates to "Quality for the price," indicating a positive aspect of the hotel. 2. The phrase افضل الفنادق بهذا السعر translates to "one of the best hotels at this price," which further emphasizes a positive sentiment regarding the hotel's value. 3. The review mentions السعر مناسب meaning "the price is appropriate," reinforcing the idea that the reviewer finds the pricing reasonable. 4. However, the overall tone of the review is factual and lacks strong emotional language or expressions of enthusiasm, which is often associated with positive sentiment. 5. The use of neutral language and the focus on price and quality without expressing strong feelings leads to the conclusion that the sentiment is neutral rather than overtly positive.
</overall_justification>
</analysis>
Now, analyze the following review:

Review: رخيص وفي احسن موقع حصلت خصم لاني اشتغل في جهة تابعة القطاع الخاص و لكن حتى الاسعار العادية بدون الخصم رخيصة بالنسبة للموقع
Return your response in the following XML format...

GPT-4o Output

<analysis>
<overall_sentiment> neutral </overall_sentiment>
<overall_justification>

1. The review begins with رخيص وفي احسن موقع', indicating positive language.
2. Mentions receiving a discount for private sector work, reinforcing affordability.
3. Even without the discount, prices are considered cheap for the location.
4. However, the tone is factual and lacks strong emotion.
5. Overall, the review is descriptive rather than opinionated, suggesting neutrality.
</overall_justification>
</analysis>

True Label: Neutral
Predicted Label (Adaptive Few-Shot): Neutral (Correct)