# SMASH at StanceEval 2024: Prompt Engineering LLMs for Arabic Stance Detection

**Youssef Al Hariri**
University of Edinburgh
Edinburgh, UK
y.alhariri@ed.ac.uk

**Ibrahim Abu Farha**
University of Sheffield
Sheffield, UK
i.abufarha@sheffield.ac.uk

## Abstract

This paper presents our submission for the Stance Detection in Arabic Language (StanceEval) 2024 shared task conducted by Team SMASH of the University of Edinburgh. We evaluated the performance of various BERT-based and large language models (LLMs). MARBERT demonstrates superior performance among the BERT-based models, achieving F1 and macro-F1 scores of 0.570 and 0.770, respectively. In contrast, the Command-R model outperforms all models with the highest overall F1 score of 0.661 and macro F1 score of 0.820.

## 1 Introduction

Stance detection, also known as stance classification, prediction, or identification, is an NLP task that aims to determine the viewpoint of a text's author (Favor, Against, or Neutral) towards a specific target topic (Küçük and Can, 2020). Furthermore, it has gained significance both as a standalone problem and in conjunction with applications such as sentiment analysis, argument mining, sarcasm detection, rumor detection, fact-checking, and fake news detection (Alhindi et al., 2021; Küçük and Can, 2020). Arabic NLP presents unique processing challenges due to its diverse dialects and complex orthographic features (Darwish et al., 2021; Abu Farha and Magdy, 2020). Furthermore, the limited resources, such as annotated corpora for Arabic compared to English, intensify these challenges (Abu Farha and Magdy, 2021; Darwish et al., 2021; Abdul-Mageed et al., 2011). Hence, the task of stance detection is gaining increasing attention in underrepresented languages such as Arabic.

StanceEval 2024 shared task (Alturayeif et al., 2024), co-organized with the ArabicNLP 2024 conference, is motivated by the aforementioned factors, aiming to adapt and overcome these challenges by confronting the distinct aspects of Arabic language processing.

| **Target Topic** | **Size** | **Train** | **Val** | **Test** |
|---|---|---|---|---|
| COVID-19 vaccine | 1,373 | 933 | 234 | 206 |
| Digital Transformation | 1,348 | 916 | 229 | 203 |
| Women Empowerment | 1,400 | 952 | 238 | 210 |
| **Total** | 4,121 | 2,801 | 701 | 619 |

Table 1: Mawqif dataset topics' distribution.

This paper describes our participation in the StanceEval shared task, where we achieved 4th place. It also provides a detailed overview of the models we used to tackle the task. We compared the performance of BERT-based models and prompt engineering on large language models (LLMs). Our official submission was based on a prompt engineering approach utilizing Command-R model, which achieved an F1 score of 0.670 and macro F1 score of 0.821. For BERT-based models, MARBERT was the best, with an F1 of 0.568 and macro F1 score of 0.761

The rest of this paper is organized as follows: Section 2 provides a detailed description of the dataset used for this work. Section 3 outlines the experimental setup and briefly describes the models we utilized. Subsequently, the results, discussions, and findings are presented in Section 4, and the paper concludes with a summary and implications of the research in Section 5.

## 2 Data

The dataset used in the StanceEval 2024 shared task is Mawqif, a multi-label Arabic dataset for target-specific stance detection (Alturayeif et al., 2022, 2024). Mawqif dataset contains 4,121 sentences covering three topics, 'COVID-19 vaccine', 'digital transformation', and 'women empowerment' as shown in Table 1. Mawqif dataset is a multi-label dataset with labels for stance (Favor, Against, None), sentiment (Positive, Negative, Neutral), and sarcasm (Sarcastic and Non-sarcastic). For evaluation, the shared task organizers split Mawqif dataset

*Proceedings of The Second Arabic Natural Language Processing Conference*, pages 800–806
August 16, 2024 ©2024 Association for Computational Linguistics

into two subsets: a training and a blind test set. However, we split the training dataset into training and validation sets for our experiments. Table 1 shows the distribution of the dataset subsets we used.

# 3 Methodology

## 3.1 Models

This section provides an overview of the models used in the experiments. The experiments span two types: fine-tuned models and zero-shot models.

### 3.1.1 Fine-tuned Models

This section delves into the specifics of the models we used in the experiments, which were fine-tuned on the training data for the task. The models include the following:

- **AraBART** (Kamal Eddine et al., 2022): an Arabic model based on BART (Lewis et al., 2020) in which the encoder and the decoder are pretrained end-to-end. It is pre-trained on a corpus of 73GB.
- **AraBERT**: an Arabic-specific BERT model provided by (Antoun et al., 2020). We utilized the two AraBERT models, `v0.2-base` and `v2-base`. Both models are pre-trained on 24GB of data from Wikipedia, news articles, and the Open Source International dataset (OSIAN).
- **mBERT** (Devlin et al., 2018): a multilingual BERT model developed by Google AI and trained on 104 languages from Wikipedia's data.
- **CAMeLBERT** (Inoue et al., 2021): we utilized two models, the dialectal Arabic (`DA`) and the mixed model (`Mix`), which trained on mixed data of Modern Standard Arabic, dialectal Arabic, and classical Arabic.
- **MARBERT** (Abdul-Mageed et al., 2021): a model trained on a set of 128GB dataset, consisting of 1B tweets.
- **MARBERTv2** (Abdul-Mageed et al., 2021): a version of MARBERT model but trained further on 61GB of MSA data in addition to an 8.6GB of Arabic news dataset.
- **QARiB** (Chowdhury et al., 2020; Abdelali et al., 2021): is a dialectal Arabic BERT model. It was trained on data from tweets and a combination of Arabic GigaWord, Abulkhair Arabic Corpus, and OPUS.

### 3.1.2 Zero-shot Models

We tested the models below in a zero-shot setup. For the experiments, we utilize the quantized versions through Ollama[1]. The following are the models' details and the specific variants used for the experiments:

- **AceGPT** (Huang et al., 2023): An open-source, culturally aware LLM was developed to align with the values of Arabic-speaking communities. We utilized the quantized model of (`13b`[2]).
- **Gemma** (Gemma Team et al., 2024): an open model built by Google DeepMind (Google AI, 2024). We utilized the 7b instruct model (`7b-instruct-v1.1-fp16`).
- **LLAMA 3**: an LLM developed by Meta (Meta, 2024) . In the experiments we employed the versions: (`70b-instruct`), and (`8b-instruct-fp16`).
- **Command-R**[3]: an LLM from Cohere, optimized for conversational interaction and long context tasks. It was trained on a massive corpus in multiple languages with main focus on critical business use-cases. We utilized the quantized version (`35b-v0.1-q8_0`).
- **Command-R+**[4]: an updated LLM from Cohere in the Command-R family, which incorporates complex RAG functionality and multi-step tool use (agents). We used the model (`104b-q2_K`) in our experiments.
- **WizardLM-2**[5] (`7B`): an LLM developed by Microsoft AI. We used the model (`wizardlm2:7b-fp16`).
- **Mistral 7b** (Jiang et al., 2023): an LLM model with open weights in 8x7b and 8x22b parameter sizes. We utilized (`instruct-v0.2-fp16`).
- **Mixtral** (`8x7b`): a high-quality sparse mixture of experts model (SMoE) with open weights provided by (Jiang et al., 2023).

## 3.2 Experimental Setup

In our experiments, we tackled the task as a classification problem and used fine-tuned models (i.e., BERT and BART-based models) and zero-shot models (i.e., LLMs).

[1] https://ollama.com/

[2] Ollama's model tag: salmatrafi/acegpt:13b

[3] https://docs.cohere.com/docs/command-r

[4] https://docs.cohere.com/docs/command-r-plus

[5] https://wizardlm.github.io/WizardLM2/

| English | Arabic |
|---|---|
| Women empowerment | تمكين المرأة |
| Covid-19 Vaccine | لقاح كوفيد |
| Digital Transformation | التحول الرقمي |

Table 2: Arabic translation of target topics.

For the BERT and BART-based models, we fine-tuned each model for 7 epochs using the *AdamW* optimizer, with an initial learning rate of 2e-05. The batch size was set to 128, with the weight decay parameter of 0.01.

Furthermore, for the zero-shot setup, we experimented with various prompts to achieve the best performance. We systematically evaluated a diverse set of language models for their ability to detect stance in Arabic text. Each model was prompted to classify the stance of each sentence as 'favor', 'against', or 'neutral' towards the target topic. The output of the models was then processed to extract the assigned label. This includes matching both American and British versions of the word favour/favor, and the rare cases of mixed Arabic answers such as (فavor).

After multiple experiments and iterations, we noticed that prompting the model to produce answers in Arabic significantly impairs the performance, exhibiting 'hallucination' behavior, as the output would usually be random Arabic sentences without a clear answer. Furthermore, integrating labels in Arabic, as opposed to English, resulted in a deterioration of performance. Thus, we decided to go with an English prompt that relies on asking the model to provide the stance of a given text. We also noticed that using the Arabic translation of the target names would improve the performance. In this setup, we believe that we minimized any issues that might arise when asking the models to generate Arabic text, and we limit our reliance on their capabilities in Arabic to comprehension only. Table2 shows the Arabic translation of the target topics. Appendix A includes some of the prompts we tested. The final prompt is as follows:

*You are an expert in analyzing people's opinions. You are an expert in Arabic. You will be given an Arabic sentence as input. Your task is to identify the stance towards the topic or subject discussed in the sentence. Your task is to identify whether the sentence is in favor of the topic, against it, or neutral. Your output should be one of the following: favor, against, or neutral. You should not provide any further information. Your answer should be in English.*
الجملة: `input_sentence`
الموضوع: `target_topic_in_Arabic`
*What is the stance in the given sentence? [favor, against, neutral].*

All models were evaluated using the official metric for the task, which is the average macro F1 score over the *favor* and *against* classes:

$$F1 = \frac{1}{2}(F_1^F + F_1^A) \tag{1}$$

Where $F_1^F$ and $F_1^A$ are the $F_1$ scores over the *favor* and *against* classes respectively.

## 4 Results and Discussion

### 4.1 Results on the evaluation set

Table 3 shows the F1 and macro F1 scores for the models on the validation and the test sets. It is worth noting that we relied on the results on the validation set to select the model to be used for the official submission. The results on the tests were calculated after the gold labels were made public and are included for comparative purposes. From the table, it is clear that MARBERT performs best among the BERT/BART-based models with a macro F1 score of 0.770 and an F1 score of 0.570. For the LLMs in a zero-shot setup, their performance is comparable with the BERT/BART-based models, and some of them are on par with MARBERT. However, Command-R model achieves the best performance with an F1 score of 0.820, which shows the strong performance of the new larger LLMs.

Generally, Command-R and LLAMA 3(70b) models demonstrated notably higher performance in identifying the stance of Arabic sentences compared to smaller models. This finding indicates the advantage of larger model architectures in handling complex linguistic tasks. Secondly, variability in performance across models highlights the significance of model-specific tuning and context adaptation for improving stance detection outcomes. In fact, the outcome of the post-evaluation testing, i.e., testing the performance of the models on the released labeled test set, confirms the findings.

| | Validation Set | | Test Set | |
|---|---|---|---|---|
| Model | F1 | macro F1 | F1 | macro F1 |
| AceGPT | 0.550 | 0.713 | 0.523 | 0.670 |
| Gemma 7B | 0.473 | 0.671 | 0.444 | 0.624 |
| Llama 3 70B | 0.636 | 0.770 | 0.659 | 0.785 |
| Llama 3 8B | 0.544 | 0.684 | 0.561 | 0.699 |
| Command-R | **0.661** | **0.820** | **0.670** | **0.804** |
| Command-R+ | 0.609 | 0.733 | 0.616 | 0.728 |
| WizardLM-2 7B | 0.428 | 0.559 | 0.391 | 0.523 |
| Mistral 7B | 0.509 | 0.654 | 0.497 | 0.612 |
| Mixtral 8x7B | 0.357 | 0.428 | 0.413 | 0.514 |
| AraBART | 0.413 | 0.609 | 0.429 | 0.616 |
| AraBERTv0.2 | 0.530 | 0.729 | 0.523 | 0.738 |
| AraBERTv2 | 0.459 | 0.668 | 0.447 | 0.641 |
| mBERT | 0.411 | 0.617 | 0.4296 | 0.633 |
| CAMeLBERT-DA | 0.433 | 0.631 | 0.460 | 0.664 |
| CAMeLBERT-Mix | 0.490 | 0.709 | 0.494 | 0.695 |
| MARBERT | **0.570** | **0.770** | **0.568** | **0.761** |
| MARBERTv2 | 0.524 | 0.756 | 0.487 | 0.709 |
| QARiB | 0.540 | 0.711 | 0.508 | 0.702 |

Table 3: Results achieved by models on validation and Test sets.

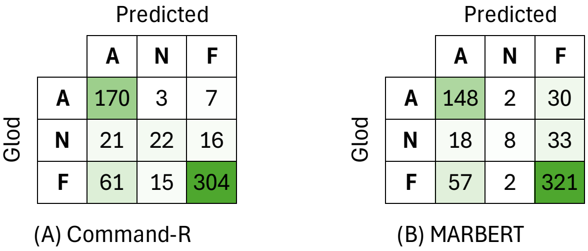
### 4.2 Official submission

Based on the results on the validation set, the best model is Command-R. Thus, it was used for the official submission. Our team, SMASH, was ranked **4th** with a macro F1 score of **0.8041**.

### 4.3 Discussion

From the results in Table 3, it is clear that fine-tuned models are achieving relatively good performance, with MARBERT being the best with a macro F1 of 0.770. This would probably be due to the fact that it was trained on tweets, which matches the nature of the task at hand. It is worth noting that MARBERT showed a stable good performance on multiple social media-related NLP tasks (Abdul-Mageed et al., 2021; Abu Farha and Magdy, 2021). For zero-shot models, their performance was variable, depending on the amount of Arabic text present in their training data. The best model was Command-R, with a macro F1 of 0.820.

The confusion matrix in Figure 1, shows the detailed performance of Command-R and MARBERT, the top two models. From the figure, it is noticeable that Command-R is better at identifying the 'against' and 'neutral' classes, while MARBERT is better at identifying the 'favor' class. Generally, it seems that MARBERT has a tendency to classify 'neutral' cases as 'favor', which might be due to the imbalance in the dataset. To address this, we balanced the training data by upsampling the 'against' class. Following this adjustment, the model achieved an F1 score of 0.543 and a macro F1 score of 0.757. These results are slightly lower than the results without data balancing. This gap in performance shows the promising capabilities of LLMs as they achieve excellent performance despite being multilingual.


Predicted
Glod
Predicted
Glod

| | A | N | F |
|---|---|---|---|
| A | 170 | 3 | 7 |
| N | 21 | 22 | 16 |
| F | 61 | 15 | 304 |

(A) Command-R

| | A | N | F |
|---|---|---|---|
| A | 148 | 2 | 30 |
| N | 18 | 8 | 33 |
| F | 57 | 2 | 321 |

(B) MARBERT

Figure 1: Confusion matrix of the predicted labels with the gold stance for Command-R and MARBERT.

## 5 Conclusion

This study explores various models for the Stance Detection in the Arabic Language (StanceEval) 2024 shared task, focusing on both BERT-based and new large language models (LLMs). Our experiments demonstrated that while MARBERT excelled among the BERT-based models, new larger models such as Command-R achieved the highest overall performance. These findings significantly highlight the potential of LLMs in capturing the nuanced semantic and contextual elements necessary for stance detection in Arabic. However, it is worth noting that all of these models are multilingual, which explains why some of them achieved relatively low performance. This signifies the importance of developing Arabic-specific LLMs, which would help improve the performance on Arabic NLP tasks.

## 6 Limitations

A major limitation of our work is that all of the models we used are either English models or multilingual models. Using Arabic-specific models would provide better performance and ability to handle Arabic, its dialects, and cultural context. We tried using JAIS, but we weren't able to load the model using the hardware available to us. Another limitation of this study is the application of a zero-shot learning approach. We would implement few-shot learning techniques in future works, as it has been shown that providing in-context examples enhances the performance of large language models.

## References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2021. A comparative study of effective approaches for arabic sentiment analysis. *Information Processing & Management*, 58(2):102438.

Tariq Alhindi, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. AraStance: A multi-country and multi-domain dataset of Arabic stance detection for fact checking. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 57–65, Online. Association for Computational Linguistics.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. Stanceeval 2024:the first arabic stance detection shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label Arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J. Jansen. 2020. Improving Arabic text categorization using transformer training diversification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 226–236, Barcelona, Spain (Online). Association for Computational Linguistics.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Google AI. 2024. Gemma.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization. In *Proceedings* *of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

AI at Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

## A Prompt Experiments

| ID | Prompt |
|---|---|
| 1 | You are an expert in analyzing people's opinions. You are an expert in Arabic. You will be given an Arabic sentence as input. Your task is to identify the stance towards the topic or subject discussed in the sentence. Your task is to identify whether the sentence is in favor of the topic, against it, or neutral. Your output should be one of the following: favor, against, or neutral. You should not provide any further information. Your answer should be in English.<br>الجملة: input_sentence<br>الموضوع: target_topic<br>What is the stance in the given sentence? [favor, against, neutral].} |
| 2 | You are an expert in analyzing people's opinions. You are an expert in Arabic. You will be given an Arabic sentence as input. Your task is to identify the stance towards the topic or subject discussed in the sentence. Your task is to identify whether the sentence is in favor of the topic, against it, or neutral. Your output should be one of the following: favor, against, or neutral. You should not provide any further information. Your answer should be in English.<br>الجملة: input_sentence<br>الموضوع: target_topic_in_Arabic<br>What is the stance in the given sentence? [favor, against, neutral]. |
| 3 | أنت مختص في اللهجات العربية والعربية الفصحى ، يجب أن تقوم بتصنيف الموقف من الموضوع المحدد في الجملة حيث يتم تحديد الموضوع ضمن السؤال. يجب عليك الإجابة بكلمة واحدة فقط من إحدى المواقف التالية (رفض ، حياد ، دعم) من دون تفصيل.<br>الجملة: input_sentence<br>الموضوع:target_topic<br>ما الموقف في هذه الجملة تجاه الموضوع المحدد؟ (رفض ، حياد ، دعم) |
| 4 | أنت مختص في اللهجات العربية والعربية الفصحى ، يجب أن تقوم بتصنيف الموقف من الموضوع المحدد في الجملة حيث يتم تحديد الموضوع ضمن السؤال. يجب عليك الإجابة بكلمة واحدة فقط من إحدى المواقف التالية (رفض ، حياد ، دعم) من دون تفصيل.<br>الجملة: input_sentence<br>الموضوع:target_topic<br>ما الموقف في هذه الجملة تجاه الموضوع المحدد؟ (رفض ، حياد ، دعم) |
| 5 | For the topic (target_topic), what is the stance of the following sentence: input_sentence [favor, against, neutral] |
| 6 | بالنسبة للموضوع (target_topic) ما الموقف تجاه الجملة: input_sentence (رفض، حياد، دعم) |

Table 4: Prompts used with LLMs.