MGKM at StanceEval2024: Fine-Tuning Large Language Models for Arabic Stance Detection

Mamoun Alghaslan and Khaled Almutairy
King Fahd University of Petroleum and Minerals
{g200818720, g202204240}@kfupm.edu.sa

Abstract

Social media platforms have become essential in daily life, enabling users to express their opinions and stances on various topics. Stance detection is a task that identifies the viewpoint expressed in text toward a target subject. Despite the growing importance of Arabic tweets in shaping public opinion, there is a lack of research on stance detection in this domain. In this work, we evaluate the effectiveness of fine-tuning three Large Language Models (LLMs) in detecting target-specific stances in the MAWQIF dataset (Alturayeif et al., 2022). The LLMs assessed are ChatGPT-3.5-turbo, Meta-Llama-3-8B-Instruct, and Falcon-7B-Instruct. Our findings demonstrate that fine-tuning substantially enhances the stance detection capabilities of LLMs in Arabic tweets. GPT-3.5-Turbo exhibits the highest performance among the evaluated models, achieving a macro-F1 score of 82.93. Our work ranked second in the StanceEval2024 leaderboard, on a blind test set (Alturayeif et al., 2024).

1 Introduction

In the age of digital communication, social media platforms have become the norm in people's daily lives, not just as a medium for social interaction but also as a platform for expressing opinions and stances towards a wide array of topics, from politics and events to services and controversial issues. This amplified the need for advanced stance detection technologies, aimed at detecting whether the author of a text is in favor, against, or neutral towards a specific subject, which is a task that is becoming increasingly important for decision-making processes across various sectors such as businesses and public authorities (Alturayeif et al., 2022).

Stance detection operates primarily on analyzing textual input to predict the author's viewpoint, which may be explicitly or implicitly conveyed. Other author's social activities such as retweets and likes can be used to enhance model performance. The task is further categorized into target-specific, cross-target, and target-independent detection, each with its unique challenges and requirements, emphasizing the complexity of this field.

Initially, identifying viewpoints relied mostly on rules and traditional machine-learning methods. For instance, support vector machines (SVM) were highly regarded in the early stages (Mohammad et al., 2016; Walker et al., 2012; Anand et al., 2011). However, the emergence of deep learning models transformed this landscape significantly. These models excelled in processing vast amounts of data and uncovering intricate patterns within it (Zhang et al., 2019; Huang et al., 2018; Dey et al., 2018; Zarrella and Marsh, 2016; Wei et al., 2016). Subsequently, pre-trained language models, such as BERT emerged, aiding in a richer comprehension of text through contextual analysis (Kawintiranon and Singh, 2021; Li et al., 2021; Devlin et al., 2018).

Nowadays, LLMs such as OpenAI's ChatGPT and Meta AI's LLaMa-2 are revolutionizing natural language processing (NLP) (Touvron et al., 2023; Qin et al., 2023). These models, trained on extensive datasets, demonstrate a remarkable ability to mimic human language nuances with high accuracy (Yin et al., 2023; Zhao et al., 2023). Unlike older models, they can tackle different types of questions and grasp language nuances better, making them valuable for NLP tasks, including stance detection.

Refining LLMs such as GPT, LLaMa3, and Falcon for particular tasks through fine-tuning improves their accuracy and applicability for contextually sensitive stance detection on social media channels (Zhang et al., 2023d). By tailoring to the unique linguistic patterns and expressions prevalent in social media discourse, these models are empowered to exceed traditional methods, demonstrating remarkable proficiency in understanding sentiments and viewpoints.

This research examines the improved perfor-

Proceedings of The Second Arabic Natural Language Processing Conference, pages 816–822
August 16, 2024 ©2024 Association for Computational Linguistics

mance of fine-tuned LLMs with the Mawqaf dataset to analyze diverse user opinions. We demonstrate that fine-tuning greatly enhances the models' comprehension of user perspectives, providing a more profound understanding of online discussions. Our study emphasizes the benefits of fine-tuning in natural language processing (NLP), particularly for stance detection, and showcases its advantages over conventional, less customized approaches.

2 Related Work

2.1 Stance Detection

Stance detection is a well-studied task in natural language processing that involves identifying an entity's opinion about a specific target (Ng and Carley, 2022). Unlike sentiment analysis, which can operate independently of context, stance classification requires understanding the context and target. This task is significant across various fields, prompting the development of numerous benchmark datasets and methodologies. Historically, the focus has been on supervised machine learning models, such as Support Vector Machines, which performed well in the SemEval-2016 stance detection competition (Lai et al., 2018; Elfardy and Diab, 2016; Mohammad et al., 2016). Neural network-based models, including convolutional neural networks (Wei et al., 2016), recurrent neural networks (Zarrella and Marsh, 2016), and advanced architectures using textual entailment (Zhao and Yang, 2020) and data augmentation (Kawintiranon and Singh, 2021), are also widely used. Recent approaches have explored multi-task learning and transfer learning with transformer-based neural networks (Alturayeif et al., 2023; Yang et al., 2019; Zhao and Yang, 2020). Despite strong in-domain performance, these models often struggle to generalize to new data or targets (Ng and Carley, 2022; Alturayeif et al., 2023).

While supervised learning with human annotations dominates the field, unsupervised techniques are also explored. Unsupervised stance labeling uses language homogeneity for classification (Zhang et al., 2023c). For instance, graph neural networks analyze information from Twitter users to infer stances based on their past tweets and interactions (Zhang et al., 2023c). Another method involves label propagation within user interaction networks, mapping user relationships, and deriving stances from aggregated data within these networks (Pick et al., 2022; Weber et al., 2013). These methods do not require predefined stance labels but often rely on specific assumptions about user behavior and language.

Zero-shot methods, which allow models to classify items without prior examples, are also used in stance classification. (Allaway and McKeown, 2023) discuss zero-shot stance detection techniques and introduce adaptations of the SemEval-2016 dataset and their VAST dataset for zero-shot classification. They outline three main paradigms: topic, language, and genre. In these paradigms, the model is trained on all data except for one element reserved for testing. Although zero-shot models typically underperform compared to fully supervised models, they provide valuable insights into stance detection without prior exposure to specific cases (Allaway and McKeown, 2023).

2.2 Large Language Models for Stance Detection

LLMs excel in tasks such as reading comprehension and solving math problems. They are trained on extensive datasets and can evaluate sentences and generate responses based on given prompts. Recent research has explored their application in stance detection, which will be discussed in the section.

Despite the development of various LLMs, much of the research has concentrated on the GPT family (e.g., ChatGPT, GPT-3) (Achiam et al., 2023), yielding mixed results. For instance, (Zhang et al., 2022) found that ChatGPT, using an instruction-based prompt, outperformed supervised models on the SemEval2016 benchmark dataset. Conversely, (Aiyappa et al., 2023) noted that while ChatGPT shows performance improvements in stance detection, the reliability of these results might be compromised by potential data contamination from its extensive training data. (Mets et al., 2024) evaluated ChatGPT for zero-shot stance detection on a custom dataset concerning immigration topics in news articles across multiple languages. Their findings indicated that although ChatGPT's performance approached that of the best supervised models, it was still inferior for stance classification. Recent studies have introduced various Chain-of-Thought prompting techniques, demonstrating improved performance by leveraging LLMs' reasoning abilities (Zhang et al., 2023a,b). (Lan et al., 2023) achieved state-of-the-art results on benchmark stance datasets using a multistage Tree-of-

Thought-like prompt.

Given the overlap between stance and sentiment classification, (Kheiri and Karimi, 2023) examined several OpenAI models for sentiment analysis, concluding that GPT models, especially when fine-tuned, surpass other models. Despite these advances, the effectiveness of LLMs in stance classification, particularly with prompt engineering and without fine-tuning, remains uncertain.

Fine-tuning LLMs poses significant challenges due to their immense size, often requiring multiple high-end GPUs and extensive memory capacity. These models, containing billions of parameters, require vast amounts of data and extensive training time, leading to high costs that can be prohibit individual researchers or small organizations. In response, techniques like Low-Rank Adaptation (LoRA) have been developed to mitigate the resource intensity of fine-tuning LLMs. LoRA introduces trainable low-rank matrices that modify the pre-trained weights during the adaptation phase, rather than retraining all parameters. This significantly reduces the number of trainable parameters, lowering memory usage and decreasing computational demands, allowing these models to be adapted with fewer resources and making it feasible using consumer-grade GPUs. (Hu et al., 2021)

LLMs operate based on prompts—free-text inputs that instruct the model on the desired output. The field of prompt engineering has emerged to optimize these inputs for better outputs (Schmidt et al., 2023; White et al., 2023; Ramlochan, 2023). Prominent prompt engineering techniques include zero-shot prompting, where the LLM receives only the task description, and few-shot prompting, which includes a few examples within the prompt (White et al., 2023; Brown et al., 2020; Wei et al., 2023). Unlike fine-tuning, these examples do not involve adjusting model weights but provide context to aid understanding. Although effective, few-shot prompting can be unstable due to factors like the order of examples (Zhao et al., 2021; Lu et al., 2021). While these techniques have shown promise, the optimal approach for using LLMs in stance detection remains an open question in research.

3 Methods

3.1 Models

We fine-tuned different LLMs and assessed their performance in detecting the author's stance. We selected three models for evaluation: GPT-3.5-Turbo-0125, Meta-Llama-3-8B-Instruct, and Falcon-7B-Instruct. We choose the Instruct variants given their pre-trained nature towards instructions. Since the dataset includes Arabic tweets with special characters such as emojis, we pre-processed them using the AraBERT pre-processor before passing them to the models.

3.2 Dataset

The experiment is based on the MAWQIF dataset, which consists of 4,121 tweets in multi-dialectal Arabic. Each tweet is annotated with a stance (Favor, Against, None) toward one of three targets: COVID-19 vaccine, Digital Transformation, and Women Empowerment. Additionally, it includes annotations for Sentiment (Positive, Negative, Neutral) and Sarcasm (Yes, No) polarities (Alturayeif et al., 2022).

3.3 System Prompt

The following example applies to Meta-Llama-3-8B-Instruct and Falcon-7B-Instruct training. For GPT-3.5-Turbo-0125, we followed the guideline provided by OpenAI. However, the system prompt part is shared between all models. An example of the prompt with input and output used for training follows:

> You are an assistant that, given an Arabic tweet, detects the writer's stance (Favor, Against, or None). None means there is no evidence in the tweet to judge the author's stance, such as inquiries, or news that does not express any positive or negative position.
>
> Input: عشان يلمع صورته ويعنني تمكين المرأة ويصير ترند والحكومة هي اكثر من تقمع المرأة اخر شيء كل الي فرحانين بالقرار ودارسين قانون متوظفين كاشير في مول راتبهم 3 الاف
>
> Output: Against

Figure 1: Example of a Tweet against Women Empowerment

3.4 Evaluation Metrics

We followed the standard evaluation protocols in (Alturayeif et al., 2024). The primary evaluation metric used is the macro F1-score. The macro F1-score is calculated as the average of the F1-scores for the "Favor" and "Against" categories. The score for the "None" stance is ignored since it is undersampled in the dataset. This metric is computed for each target separately, and then the overall macro

Model	Target	F1-scores		
		Favor	Against	Average
Meta-Llama-3-8B-Instruct	Women Empowerment	0.8922	0.7342	0.8132
	COVID-19 Vaccine	0.8466	0.7248	0.7857
	Digital Transformation	0.9141	0.3478	0.6309
	Overall	0.7433		
Falcon-7B-Instruct	Women Empowerment	0.8013	0.0392	0.4203
	COVID-19 Vaccine	0.5840	0.1569	0.3704
	Digital Transformation	0.8623	0.0000	0.4312
	Overall	0.4073		
GPT-3.5-Turbo-0125	Women Empowerment	0.9266	0.8817	0.9042
	COVID-19 Vaccine	0.9172	0.8852	0.9012
	Digital Transformation	0.8571	0.8000	0.8286
	Overall	**0.8293**		

Table 1: Macro F1-scores for on the validation split.

F1-score is computed across all targets, which is given as follows:

$$F_{macro} = \frac{F_{favor} + F_{against}}{2}$$

3.5 Experimentation

We fine-tuned GPT-3.5-Turbo-0125 through OpenAI's web services, and the two open-source models locally. To reduce the model sizes to a manageable scale for the local training, we applied Low-Rank Adaptation (LoRA) (Hu et al., 2021) using the LitGPT (AI, 2023) framework. The dataset was divided into an 85-15% train-validation split with stratification based on stance.

All models were fine-tuned over 3 epochs using a single NVIDIA A100 PCIE with 40GB RAM. The hyperparameters for fine-tuning include LoRA rank of 32, α of 16, dropout rate of 0.05, batch size of 8, 10 warm-up steps, and a learning rate of 2×10^{-4}. Meta-Llama-3-8B-Instruct and Falcon-7B-Instruct required around 20 and 15 minutes to train respectively, whereas GPT-3.5-Turbo-0125 took around 30 minutes.

4 Results

Our results, shown in Table 1, GPT-3.5-Turbo demonstrated the highest overall performance, achieving a macro F1-score of 82.93%. It performed consistently well across all target categories: Women Empowerment (90.42%), COVID-19 Vaccine (90.12%), and Digital Transformation (82.86%). This indicates GPT-3.5-Turbo-0125's robust generalization capabilities and superior ability to handle multi-dialectal Arabic tweets.

Meta-Llama-3-8B-Instruct showed moderate performance with a macro F1-score of 74.33%. It performed well in detecting 'favor' stances but struggled with 'against' stances, particularly in the Digital Transformation category. This suggests limitations in the model's comprehension and classification capabilities.

Falcon-7B-Instruct had the lowest performance, with a macro F1-score of 40.73%. The model's performance varied significantly across different categories: Women Empowerment (42.03%), COVID-19 Vaccine (37.04%), and Digital Transformation (43.12%). The model struggled particularly with the 'against' class, achieving scores of 3.92% for Women Empowerment, 15.69% for COVID-19 Vaccine, and 0% for Digital Transformation. The exact reasons for Falcon-7B-Instruct's poor performance remain uncertain, but it seems the model struggles significantly with accurately classifying 'against' stances, which heavily impacted its overall effectiveness.

fine-tuning LLMs significantly enhances their ability to detect stances in Arabic tweets. GPT-3.5-Turbo-0125 emerged as the most effective model, while Meta-Llama-3-8B-Instruct and Falcon-7B-Instruct showed potential but need further optimization. Future research can focus on refining hyperparameters, and instruction sets, or integrating Retrieval-Augmented Generation (RAG) to ingest a few examples similar to the input query as context to the model to improve LLM performance.

5 Conclusion

In this paper, we conducted a detailed analysis of stance detection on the MAWQIF dataset, an Arabic-language corpus annotated for multiple opinion dimensions across various dialects. Our experiments demonstrate the effectiveness of LLMs in enhancing the accuracy of stance detection. We evaluated the performance of three different LLMs, GPT-3.5-Turbo-0125, Meta-Llama-3-8B-Instruct, and Falcon-7B-Instruct and observed notable differences in their ability to handle the complexities of multi-dialectal Arabic in social media texts.

The results indicate that fine-tuning LLMs significantly improves their ability to understand and detect stances in Arabic tweets. Notably, GPT-3.5-Turbo-0125 emerged as the top performer, achieving remarkable precision in identifying both 'favor' and 'against' stances, underscoring the potential of fine-tuned LLMs for language-specific applications. The effectiveness of fine-tuning is further validated by the significant improvement over baseline models. Additionally, the research highlights the challenges associated with fine-tuning LLMs, such as the substantial computational resources required and the complexities of adapting these models to specialized tasks. However, techniques like LoRA have proven effective in mitigating these challenges, facilitating more accessible and efficient fine-tuning processes.

As we move forward, the insights gained from this study can guide future research towards enhancing model robustness, exploring more diverse datasets, and refining computational techniques to better meet the evolving needs of natural language processing applications. The integration of stance detection models into practical applications promises to improve decision-making processes, social media monitoring, and public sentiment analysis, making significant strides towards more informed and responsive digital communication platforms.

References

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

Lightning AI. 2023. Litgpt. https://github.com/Lightning-AI/litgpt.

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on chatgpt? *arXiv preprint arXiv:2303.12767*.

Emily Allaway and Kathleen McKeown. 2023. Zero-shot stance detection: Paradigms and challenges. *Frontiers in Artificial Intelligence*, 5:1070429.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144.

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 The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 529–536. Springer.

Heba Elfardy and Mona Diab. 2016. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 434–439.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling: 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings 11*, pages 197–206. Springer.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 15–27. Springer.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2023. Stance detection with collaborative role-infused llm-based agents. *arXiv preprint arXiv:2310.10467*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2024. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *Plos one*, 19(4):e0302380.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Lynnette Hui Xian Ng and Kathleen M Carley. 2022. Is my stance the same as your stance? a cross validation study of stance detection datasets. *Information Processing & Management*, 59(6):103070.

Ron Korenblum Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. Stem: Unsupervised structural embedding for stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11174–11182.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *Preprint*, arXiv:2302.06476.

Sunil Ramlochan. 2023. What is prompt engineering? `https://www.promptengineering.org/what-is-prompt-engineering/`.

Douglas C Schmidt, Jesse Spencer-Smith, Quchen Fu, and Jules White. 2023. Cataloging prompt patterns to enhance the discipline of prompt engineering. *URL: https://www. dre. vanderbilt. edu/~schmidt/PDF/ADA_Europe_Position_Paper. pdf [accessed 2023-09-25]*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 290–297.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 384–388.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. Blcu_nlp at semeval-2019 task 7: An inference chain-based gpt model for rumour evaluation. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 1090–1096.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *Preprint*, arXiv:2305.18153.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Bowen Zhang, Daijun Ding, Liwen Jing, and Hu Huang. 2023a. A logically consistent chain-of-thought approach for stance detection. *arXiv preprint arXiv:2312.16054*.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023b. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2023c. Doubleh: Twitter user stance detection via bipartite graph neural networks. *arXiv preprint arXiv:2301.08774*.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023d. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *Preprint*, arXiv:2303.16199.

Guangzhen Zhao and Peng Yang. 2020. Pretrained embeddings for stance detection with hierarchical capsule network on social media. *ACM Transactions on Information Systems (TOIS)*, 39(1):1–32.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.