NYCU-NLP at SemEval-2025 Task 11: Assembling Small Language Models for Multilabel Emotion Detection and Intensity Prediction

Zhe-Yu Xu, Yu-Hsin Wu and Lung-Hao Lee*
Institute of Artificial Intelligence Innovation
National Yang Ming Chiao Tung University
*lhlee@nycu.edu.tw

Abstract

This study describes the design of the NYCU-NLP system for the SemEval-2025 Task 11 that focuses on multi-lingual text-based emotion analysis. We instruction-tuned three small language models: Gemma-2 (27B), Mistral-small-3 (22B), and Phi-4 (14B) and then assembled them as our main system architecture. Our NYCU-NLP system participated the English Track A for multilabel emotion detection and English Track B for emotion intensity prediction. Experimental results show our best-performing submission produced a macro-averaging F1 score of 0.8225, ranking second of 74 participating teams for Track A, and ranked second among 36 teams for Track B with a Pearson correlation coefficient of 0.8373 in the task official rankings.

1 Introduction

Emotion recognition is a well-known NLP task that focuses on identifying affective states from texts. People express their perceived feelings using commonly used language in highly variable ways even within the same culture or social groups (Wiebe et al. 2005, Mohammad and Kiritcheko 2018, Mohammad et al. 2018). How to detect multiple perceived emotions and predict their emotion intensities is still a challenging research problem.

SemEval-2025 Task 11 (Muhammad et al., 2025b) aims to determine what emotion most people would think the speaker may be feeling given a short text written by the speaker. This shared task consists of three tracks, including 1) Track A (multilabel emotion detection): Given a target text, predict the perceived emotions of the speaker by selecting whether each of the following emotions apply: joy, sadness, fear, anger, surprise or disgust. 2) Track B (emotion intensity): Given a target text and target perceived emotions, predict the intensity for each of the classes. The set of ordinal intensity includes: 0 (no emotion), 1 (low degree of emotion), 2 (moderate degree of emotion), and 3 (moderate degree of emotion). 3) Track C (cross-lingual emotion): given a text written in one of 32 involved languages, predict the perceived emotion labels of a new text in a different target language. The dataset in this track has the same format as in Track A. Participating teams can choose to join in one or more languages and tracks based on their preference.

This paper describes the **NYCU-NLP** (**N**ational **Y**ang Ming **C**hiao Tung **U**niversity, **N**atural **L**anguage **P**rocessing Lab) system for the SemEval-2025 Task 11. Given the promising results obtained by Large Language Models (LLM) for various NLP tasks, we aggregate several Small Language Models (SLM), which are essentially smaller versions of LLM counterparts for this text-based emotion analysis task. We participated in English Tracks A and B only. Our system explored the use of instruction-tuned SLMs, including Gemma-2 (27B) (Riviere et al., 2024), Mistral-small-3 (22B) and Phi-4 (14B) (Abdin et al., 2024) and then assembled the SLMs to detect multilabel emotions and predict their intensities for given text-based emotion analysis. Experimental results showed our best submission achieved a macro-averaging F1-score of 0.8225, ranking second of 74 participating teams for Track A, and produced a Pearson correlation coefficient of 0.8373, also ranking second of 36 teams for Track B.

The rest of this paper is organized as follows. Section 2 reviews recently related studies on emotion detection and intensity prediction. Section

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1129–1135

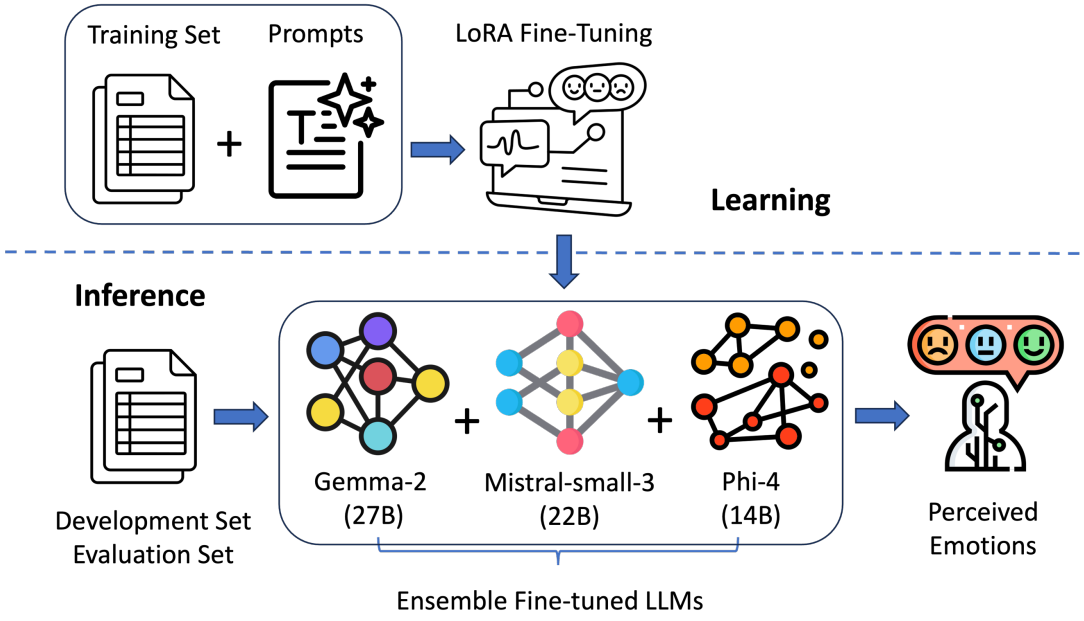

Figure 1: Our NYCU-NLP system architecture for the SemEval-2025 Task 11.

3 describes the NYCU-NLP system for this shared task. Section 4 presents results and performance comparisons. Conclusions are drawn in Section 5.

2 Related Work

Empirical evaluations showed that transformer-based language models usually outperformed conventional neural networks for emotion intensity prediction (Lee et al., 2022). Sentiment-enhanced RoBERTa transformers were used to predict emotion and empathy intensities (Lin et al., 2023). A transformer-based fusion model was proposed to integrate semantic representations at different degrees of linguistic granularity for emotional intensity predication (Deng et al., 2023). Recently, transformer-based large language models (LLM) have been used for emotion detection. EmoLLM (Liu et al., 2024) is a series of instruction-following LLMs for affective analysis based on fine-tuning various LLMs with instruction data. The LLM-GEm (Hasan et al., 2024) system was designed to use GPT 3.5 for empathy intensity prediction. EmoTrigger (Singh et al., 2024) was proposed to evaluate the ability of CPT-4, Llama-2-Chat-13B and Alpaca-13B to identify emotion triggers and consider their importances for emotion detection. An assembly of the Starling-7B and Llama-3-8B was fine-tuned to prediction cross-lingual emotion intensity (Lin et al., 2024).

Small language models (SLM) are smaller in scale and scope than their original large model counterparts, and typically include fewer than 70 billion parameters, as opposed to LLMs with up to trillions of parameters. SLM are thus usually compact and efficient with less memory and computational power. Given limited computation resources, we are motivated to explore systems based on SLMs for emotion detection and intensity prediction.

3 The NYCU-NLP System

Figure 1 shows our NYCU-NLP system architecture for the SemEval-2025 Task 11. We instruction-tuned several SLMs and then assembled them by averaging the predicted results for multi-label emotion detection (Track A) and emotion intensity prediction (Track B).

3.1 Small Language Models

The following SLMs were used to detect emotions and predict the corresponding emotion intensity.

(1) Gemma-2 (27B)

Gemma-2 (Riviere et al., 2024) with 27B parameters is a new addition to Google's Gemma family, and provides higher-performing and more efficient inference. It applies interleaving local-global attentions and group-query attention to offer a competitive alternative to models more than twice its size.

System Prompt:
You are an emotion classification assistant. Your task is to predict the intensity of emotions expressed in the input sentences for the following categories: 'Anger', 'Fear', 'Joy', 'Sadness', and 'Surprise'.
- The intensity levels are:
  - 0: No emotion,
  - 1: Low intensity,
  - 2: Moderate intensity,
  - 3: High intensity.
- Provide the intensity for each emotion as a number between 0 and 3.
- Always output the result in the format: 'Anger: X, Fear: X, Joy: X, Sadness: X, Surprise: X' where X is the predicted intensity for each emotion.
- Ensure your prediction reflects the perceived intensity of the input sentence for all emotions, even if some intensities are 0.

User Prompt:
Input sentence: {sentence}
Output format: Provide the intensity for each emotion in the following format: 'Anger: X, Fear: X, Joy: X, Sadness: X, Surprise: X', where X is a number between 0 and 3

Assistant Prompt:
{intensity}

Figure 2: Prompts used for instruction tuning.

(2) Mistral-small-3 (22B)

Mistral-small-3 set a new benchmark in the small LLMs below 70B, successfully converting a mixture-of-experts architecture into a single dense 22B parameter model.

(3) Phi-4 (14B)

Phi-4 (Abdin et al., 2024) is the latest SLM in Microsoft's Phi family, offering high quality results at a small size with 14B parameters. It outperforms larger models due to the use of high-quality datasets and post-training innovations.

3.2 Instruction Fine-tuning

We used instruction tuning (Wei et al., 2022) and LoRA (Hu et al., 2021) techniques with prompts shown in Fig. 2 to optimize the above-mentioned three pre-trained SLMs model for this task. The system was configured as an emotion classification assistant. We asked the SLM to classify a given sentence into four defined emotion intensities, including 0 for no emotion, 1 for low intensity, 2 for medium intensity and 3 for high intensity. We also guided the SLM to provide the intensity score for each emotion using the given output format.

3.3 Assembly Mechanism

During the inference phase, each SLM conducts an independent prediction for each testing instance. We then used an averaging-based assembly mechanism to determine the system output by averaging the predicted intensity scores for each emotion.

For the multilabel emotion detection subtask (Track A), if a testing instance obtained an average intensity value exceeding 0, we predicted perception of the emotion, otherwise no emotion.

For the emotion intensity prediction subtask (Track B), if a testing instance obtained an average intensity that is a non-integer value, we rounded the value to predict as its intensity score for each emotion.

4 Experiments and Results

4.1 Data

The datasets were mainly provided by task organizers (Muhammad et al., 2025a). Tracks A and B shared the same datasets, respectively including 2768, 117 and 2768 instances in the training, development and test sets. We only used the training set for instruction-tuning the SLMs without data augmentation. The average instance length is 15.5 tokens with about 1.5 emotion labels per instance. The English datasets used do not include the disgust emotion. The mostly common emotion was found to be fear (total 1,611 cases accounting for 58.20%), followed by sadness (878 cases/31.72%), surprise (839 cases/30.31%), joy (674 cases/24.35%), and anger (333 cases/12.03%).

4.2 Settings

All pre-trained models were downloaded from HuggingFace[1]. We continuously fine-tuned these models using only the training set provided by task organizers. All experiments were conducted on a server with four Nvidia V100 GPUs (Total 128GB memory). The hyperparameter values of our used LLMs were finally optimized as follows: epochs 10; batch size 4; optimizer paged AdamW (32 bit); learning rate 1e-4; LoRA r 16; LoRA alpha 32 and LoRA drop 0.01.

4.3 Metrics

For Track A on multilabel emotion detection, the macro-averaging F1 was used to measure the model performance based on predicted emotion labels and the ground truth.

[1] https://huggingface.co/google/gemma-2-27b-it
https://huggingface.co/NyxKrage/Microsoft_Phi-4
https://huggingface.co/mistralai/Mistral-Small-Instruct-2409

Model (#para)	Track A: Multilabel Emotion Detection (English/Development Set)						
	Anger	Fear	Joy	Sadness	Surprise	Micro F1	Marco F1
Gemma-2 (27B)	0.9143	0.8636	0.7925	0.7568	0.8125	0.8268	0.8279
Mistral-small-3 (22B)	**0.9375**	**0.8722**	**0.8214**	0.7671	**0.8750**	**0.8492**	**0.8546**
Phi-4 (14B)	0.8824	0.8615	0.7925	**0.8182**	0.8065	0.8348	0.8322
Assemble	0.9091	0.8722	0.7925	0.8056	0.8571	0.8475	0.8473

Table 1: Fine-tuned SLM results on the development set of Track A.

Model (#para)	Track B: Emotion Intensity (English/Development Set)					
	Anger	Fear	Joy	Sadness	Surprise	Average Pearson r
Gemma-2 (27B)	**0.9001**	**0.8157**	0.8336	0.8429	0.8114	0.8407
Mistral-small-3 (22B)	0.8887	0.7889	**0.8554**	0.8518	**0.8425**	**0.8455**
Phi-4 (14B)	0.8927	0.7747	0.8285	0.8651	0.7614	0.8245
Assemble	0.8834	0.8048	0.8285	**0.8881**	0.8202	0.8450

Table 2: Fine-tuned SLM results on the development set of Track B.

Model (#para)	Track A: Multilabel Emotion Detection (English/Evaluation Set)						
	Anger	Fear	Joy	Sadness	Surprise	Micro F1	Marco F1
Mistral-small-3 (22B)	**0.7741**	0.8845	0.8164	0.8076	0.7844	0.8308	0.8134
Assemble	0.7720	**0.8865**	**0.8318**	**0.8213**	**0.8010**	**0.8400**	**0.8225**

Table 3: Testing results on the evaluation set of Track A.

Model (#para)	Track B: Emotion Intensity (English/Development Set)					
	Anger	Fear	Joy	Sadness	Surprise	Average Pearson r
Mistral-small-3 (22B)	0.8247	0.8373	0.8406	0.8329	0.7725	0.8216
Assemble	**0.8332**	**0.8488**	**0.8591**	**0.8530**	**0.7923**	**0.8373**

Table 4: Testing results on the evaluation set of Track B.

For Track B on emotion intensity prediction, the Pearson correlation between the predicted intensity for each emotion and gold standard was used to evaluate performance.

4.4 Results

Tables 1 and 2 respectively show the evaluation results for Tracks A and B on the development sets. Among three independent SLMs, Mistral-small-3 (22B) outperformed the others on both tracks. Assembly models usually outperformed independent ones. However, assembly SLMs showed slightly reduced performance on both tracks. The small size of the development (only 117 instances) may have introduced bias. Each participating team was allowed to submit at most three submissions for evaluation and the last submission will be regarded as the official submission. We submitted the independent Mistral-small-3 (22B) and the assemble model as our final submission for official ranking.

Tables 3 and 4 respectively show the submission results for Tracks A and B on the evaluation set. The assemble SLMs usually outperformed Mistral-

small-3 (22B) in terms of both overall performance and individual emotion for both tracks. Our assemble SLM-based model respectively achieved a macro-averaging F1 of 0.8225 for Track A on multilabel emotion detection and a Pearson correlation coefficient of 0.8373 for Track B on emotion intensity prediction.

4.5 Rankings

Our final assemble submission ranked second for English Track A among a total of 74 participating teams and second for English Track B among all 36 official submissions.

4.6 Discussion

Due to limited time and computational resources, we did not use prompt engineering techniques to configure other prompts for optimization. Therefore, prompts used for instruction fine-tuning may need to be improved for performance enhancement.

We only used the training set for instruction-tuning the SLMs, and data augmentation techniques may further improve model tuning.

Since the SLMs were pre-trained using multi-lingual data, the distribution of emotion classes in small fine-tuned data may not affect model performance for individual emotion categories in our experiments.

We selected the SLMs based on the recent performance of the general benchmarks, which may not be appropriate for multilabel emotion detection and intensity prediction tasks.

The SLMs are multi-lingual so that they may be expanded to languages other than English with language-specific fine-tuned data for text-based emotion analysis task.

5 Conclusions

This study describes the NYCU-NLP system for the SemEval-2024 text-based emotion analysis task, including system design and performance evaluation. We instruction-fine-tuned the SLMs to effectively detect emotion categories and predict their emotion intensities. Experimental results indicate that our best submission is an assembly of the Gemma-2 (27B), Mistral-small-3 (22B) and Phi-4 (14B) models, achieving a macro-averaging F1 score of 0.8225 for the multilabel emotion detection track (ranking second out of seventy-four submissions) and a Pearson correlation coefficient of 0.8373 for the emotion intensity prediction track (ranking second of thirty-six).

This pilot study is our first exploration based on applying SLMs to text-based emotion analysis tasks. Future work will exploit other advanced SLMs to further improve performance.

Limitations

This work does not propose a new model to address this task for multilabel emotion detection and intensity prediction. Experiments were conducted with basic settings without other advanced explorations due to computational resource limitations.

Acknowledgments

This study was partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 111-2628-E-A49-029-MY3. This work was also financially supported by the Co-creation Platform of the Industry Academia Innovation School, NYCU.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sebastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *arXiv preprint*, arXiv:2412.08095v1. https://doi.org/10.48550/arXiv.2412.08905

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Towards transformer fusions for Chinese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*, 11:109974-109982. https://doi.org/10.1109/ACCESS.2023.3322436

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2214-2231.

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. *arXiv preprint*, *arXiv:2106.09685v2*. https://doi.org/10.48550/arXiv.2106.09685

Lung-Hao Lee, Jian-Hong Li, ad Liang-Chih Yu. 2022. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4): Article 65, 1-18. https://doi.org/10.1145/3489141

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 Empathy, Emotion, and Personality Shared Task: Perceived intensity prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 548-552. https://doi.org/10.18653/v1/2023.wassa-1.49

Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, and Lung-Hao Lee. 2024. NYCU-NLP at EXALT 2024: Assembling large language models for cross-lingual emotion and trigger detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 505-510. https://doi.org/10.18653/v1/2024.wassa-1.50

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487-5496. https://doi.org/10.1145/3637528.3671552

Saif Mohammad, and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1-17. https://doi.org/10.18653/v1/S18-1001

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 languages. *arXiv preprint*, arXiv:2502.11926. https://doi.org/10.48550/arXiv.2502.11926

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, and Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval Task 11: Bridging the gap in the text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation*.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena

Heuermann, Leticia Lago, Lilly McNealus et al. 2024. Gemma 2: Improving open language models at a practical size. arXiv preprint, arXiv:2408.00118v3. https://doi.org/10.48550/arXiv.2408.00118

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 603-614. https://doi.org/10.18653/v1/2024.naacl-short.51

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations*. arXiv:2109.01652v5. https://doi.org/10.48550/arXiv.2109.01652

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2005), 165-210. https://doi.org/10.1007/s10579-005-7880-9