CHEER-Ekman: Fine-grained Embodied Emotion Classification

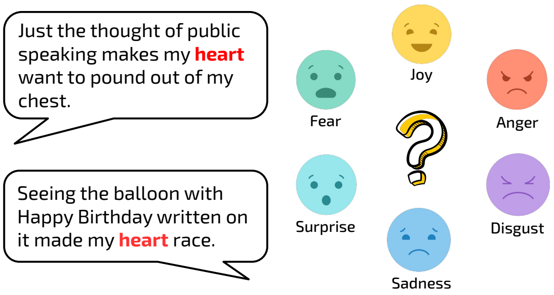
Phan Anh Duong, Cat Luong, Divyesh Bommana, Tianyu Jiang
University of Cincinnati
`{duongap, luongcn, bommandh}@mail.uc.edu, tianyu.jiang@uc.edu`

Abstract

Emotions manifest through physical experiences and bodily reactions, yet identifying such embodied emotions in text remains understudied. We present an embodied emotion classification dataset, CHEER-Ekman,[1] extending the existing binary embodied emotion dataset with Ekman's six basic emotion categories. Using automatic best-worst scaling with large language models, we achieve performance superior to supervised approaches on our new dataset. Our investigation reveals that simplified prompting instructions and chain-of-thought reasoning significantly improve emotion recognition accuracy, enabling smaller models to achieve competitive performance with larger ones.

1 Introduction

Emotions are not merely abstract mental states; they are deeply intertwined with somatic experiences. When we feel joy, our faces light up with smiles; when we are scared, our hearts race and our hands tremble. These physical reactions are more than just side effects—they are part of how we experience and express emotions. This concept, known as embodied emotion, suggests that our bodies play a key role in how we feel, perceive, and understand emotions (Lakoff and Johnson, 1999; Niedenthal, 2007). In natural language, these connections surface as descriptions of physiological reactions (e.g., "my stomach churned in disgust") or unintentional physical actions (e.g., "she stomped her feet in frustration")—phenomena termed *embodied emotions*. Recognizing such expressions is pivotal for understanding implicit emotional cues in narratives. While recent advances in NLP have focused more on explicit emotion classification (Mohammad et al., 2018) or sentiment analysis (Rosenthal et al., 2017), the subtler task of identifying embodied emotions remains less explored, despite its psychological grounding and practical relevance.



Just the thought of public speaking makes my **heart** want to pound out of my chest.

Seeing the balloon with Happy Birthday written on it made my **heart** race.

Figure 1: Illustration of embodied emotions classified into six categories.

The CHEER dataset (Zhuang et al., 2024) filled a gap in this field by providing a collection of sentences where body parts are used to express emotions. The dataset includes 7,300 human-annotated sentences containing body part references, proposing a binary classification task which we will refer to as "embodied emotion detection." However, one limitation of this work is that it does not distinguish between different types of emotions—for instance, whether a racing heart signals fear or excitement. The framework of Ekman's (1992) basic emotions offers a potential solution to this limitation. By linking embodied expressions to these specific emotions, we can build systems that better understand human emotional experiences.

To achieve this goal, we extend the CHEER dataset by annotating all its 1,350 positive samples with six Ekman emotion labels (*Joy*, *Sadness*, *Anger*, *Disgust*, *Fear*, and *Surprise*), creating a new dataset, **CHEER-Ekman**, as illustrated in Figure 1. For clarity, we refer to the novel classification task produced by this dataset as "embodied emotion *classification*," as compared to the binary task discussed above. We adopt the automatic best-worst scaling (BWS) technique (Kiritchenko and Mohammad, 2017; Bagdon et al., 2024) with large

[1] `https://github.com/menamerai/cheer-ekman`

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1118–1131
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

language models (LLMs) to tackle the task. Our experiments show that using Llama 3.1 8B with BWS significantly outperforms zero-shot prompting. The best BWS experiment achieved 50.6 F1-score, surpassing supervised BERT (49.6) and beating zero-shot approaches by around 20 points. Building on Zhuang et al.'s (2024) investigation of LLMs' capability for embodied emotion detection, we further explore prompting techniques that enhanced the detection task. Our experimental results reveal that LLMs can make better recognition when instructions are rephrased in plain, easily understood language, which boosts F1 by nearly 30 points compared to technical definitions. Moreover, chain-of-thought reasoning enables an 8B parameter model to nearly match a 70B model, closing the performance gap to within 7 F1 points. In summary, our contributions are three-fold:

1. We present CHEER-Ekman, an extension of the CHEER dataset that enriches embodied emotion expressions with fine-grained Ekman emotion labels, addressing a critical gap in understanding how specific emotions manifest through bodily expressions. Our dataset is available at: `https://github.com/menamerai/cheer-ekman`.

2. We demonstrate that the automatic best-worst scaling technique enables LLMs to perform emotion classification without any task-specific training, achieving performance that exceeds supervised approaches.

3. We reveal that counterintuitively, simplified everyday language in prompts dramatically outperforms technical definitions for embodied emotion tasks, and that structured reasoning through chain-of-thought can allow smaller language models to perform at a level closer to larger language models on our task.

2 Related Work

Emotion recognition in natural language processing has been extensively studied, with researchers focusing more on explicit emotion using datasets like SemEval (Strapparava and Mihalcea, 2007; Mohammad et al., 2018) and GoEmotions (Demszky et al., 2020). Recent research has expanded to explore nuanced aspects of emotion expression (Li et al., 2021), including emotion intensity prediction (Mohammad, 2018; Bagdon et al., 2024) and the detection of subtle emotional cues in dialogues (Poria et al., 2019; Ghosal et al., 2020; Li et al., 2022). The advent of large language models (LLMs) has catalyzed significant advances in emotion understanding capabilities (Lee et al., 2024; Sabour et al., 2024; Zhao et al., 2024; Liu et al., 2024).

While these works contribute to a deeper understanding of emotion detection in text, the embodied nature of emotions—how physical sensations and actions encode affective states—has received comparatively less attention. The concept of embodied emotion is rooted in cognitive science, particularly in the works of Lakoff and Johnson (1999) and Niedenthal (2007), which suggest that emotional experiences are closely tied to bodily states and actions. Despite the psychological grounding of embodied emotions, computational approaches to capturing them in text remain limited. Zhuang et al. (2024) introduced the CHEER dataset, which provides a collection of sentences where body parts are explicitly used to express emotions. Our work extends theirs by incorporating Ekman's six basic emotions into embodied emotion recognition, offering a more fine-grained classification system.

Another relevant line of work is the study of emotion taxonomy. Although recent advances in psychology have offered newer granular categories of emotions such as 27 emotions by Cowen and Keltner (2017), which has been adopted in both textual emotion datasets (Demszky et al., 2020) and visual emotion dataset (Kosti et al., 2019), we follow the vast majority of existing emotion datasets (Strapparava and Mihalcea, 2007; Mohammad et al., 2018; Poria et al., 2019) by utilizing the six basic emotions (*Joy*, *Sadness*, *Anger*, *Disgust*, *Fear*, and *Surprise*) proposed by Ekman (1992), which remain foundational due to their universality and simplicity. Future research may explore integrating alternative taxonomies into embodied emotion classification to enhance both granularity and coverage.

3 Methods

Our methodological approach comprises three key components that build upon and extend the work of Zhuang et al. (2024). First, we explore prompting strategies to enhance LLMs' capability to detect embodied emotions. Second, we introduce CHEER-Ekman, a refinement of the original CHEER dataset that adds fine-grained emotion labels. Finally, we adopt the BWS framework for emotion classification that leverages comparative judgments to improve classification accuracy.

3.1 Prompting LLMs for Embodied Emotion Detection

To address the gap in prompt design within the Zhuang et al.'s (2024) framework, we first sought to enhance the embodied emotion detection task using state-of-the-art LLMs and explore the impact of prompt engineering on performance. Our approach centers on two key strategies: prompt simplification to mitigate linguistic complexity and chain-of-thought (CoT) prompting.

Prompt Simplification. We investigated the effects of linguistic and domain complexity by conducting experiments with the base Llama-3.1 (Grattafiori et al., 2024) and the recently released DeepSeek-R1 distilled version (DeepSeek-AI et al., 2025). Specifically, we compared two prompts: the base prompt used in Zhuang et al. (2024) and a simplified prompt, which reduces syntactic and lexical complexity to minimize potential comprehension barriers for LLMs.

Chain-of-Thought Prompting. We further explored eliciting reasoning from the model by implementing chain-of-thought (CoT) prompting. Based on Zhuang et al.'s (2024) annotation criteria for embodied emotion detection, we developed three CoT variants: a 2-step variant that evaluates emotional causation and purposeless expression, a 3-step variant that adds body part identification, and a simplified 2-step variant with reduced linguistic complexity. These variants allowed us to examine how explicit causal reasoning affects both the model's emotion detection performance and its understanding of body-emotion relationships.

3.2 Dataset Creation

While embodied emotion detection identifies emotional expressions through bodily movements, understanding the specific emotions conveyed requires more fine-grained annotation. To address this need, we propose **CHEER-Ekman**, a refined dataset extending the original CHEER corpus (Zhuang et al., 2024) by annotating its 1,350 positive embodied emotion instances with Ekman's (1992) six basic emotions (*Joy*, *Sadness*, *Anger*, *Disgust*, *Fear*, and *Surprise*). Our adoption of Ekman's basic emotions taxonomy balances granularity with practical considerations, as recent research by Liu et al. (2024) demonstrates that finer-grained emotion taxonomies often face sparsity issues, even in bigger datasets like GoEmotions (Demszky et al., 2020).

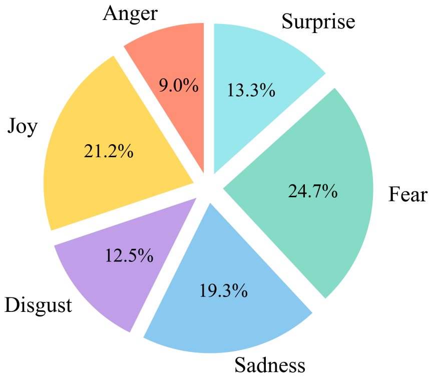


Figure 2: CHEER-Ekman dataset distribution of emotions.

Emotion	Sentence
Joy	... watched the fireflies with a loving look on his **face**.
Sadness	... frowning and scuffing his **feet** along the floor.
Fear	Marty nervously runs his **fingers** through his hair...
Anger	... makes me want to hit my **head** against the wall.
Disgust	Dean snorted incredulously, shaking his **head** in disbelief.
Surprise	... my **eyes** almost fell out of my head.

Table 1: Examples in our CHEER-Ekman datasets.

This approach also maintains consistency with Zhuang et al.'s (2024) methodology, which utilized emotion-associated adverbs derived from these basic emotions for weak supervision. We have also elected not to include the weakly labeled positive samples from the CHEER dataset, prioritizing our final dataset quality and reliability over quantity.

We recruited two annotators to label the 1,350 embodied emotion sentences from the original CHEER dataset. For each sentence, we provide the annotators with the sentence, the relevant body part, and up to three preceding sentences for context. We then ask the annotators to select one of the six emotions that best match the physical experience described through the body part. The pairwise inter-annotator agreement by Cohen's Kappa is 0.64, indicating good agreement. Finally, the annotators adjudicated their disagreements to produce the final gold labels. We show some example sentences with their annotated emotions in Table 1.

Figure 2 illustrates the emotion distribution of sentences in our newly constructed CHEER-Ekman dataset. Specifically, *Fear* is the most prevalent emotion (24.7%), followed by *Joy* (21.2%), *Sadness* (19.3%), *Surprise* (13.3%), and *Disgust* (12.5%). *Anger* appeared last at 9.0%.

3.3 BWS for Emotion Classification

To address the fine-grained emotion classification task proposed with CHEER-Ekman, we first tested zero-shot LLM prompting. However, the model often fails to adhere to the instructions and is prone to erroneous behaviors, leading to incorrect outputs. Inspired by recent work on automatic emotion intensity annotation using LLMs (Bagdon et al., 2024), we adopted the same best-worst scaling technique. The methodology involves presenting the LLM with tuples of four different sentences, instructing it to identify the body part instances that most and least represent a specific Ekman emotion. Then, the equation $\frac{\#Best-\#Worst}{\#Total}$ is used to calculate a score per sentence per emotion. Finally, we choose the emotion that receives the highest score to be the prediction for the sentence.

To examine how the number of comparisons affects classification accuracy, we tested a broader range of tuple counts, from $2N$ to $72N$, increasing by 50% at each step of expansion (where N is the number of instances to be classified). While Bagdon et al. (2024) found that more tuples improved accuracy in their experiments up to 12N, we expanded this investigation to 72N to further explore performance gain behaviors from scaling comparative rounds.

4 Evaluation

We conduct experiments to tackle both the embodied emotion detection and emotion classification tasks. The embodied emotion detection CHEER dataset contains 7,300 sentences, and our CHEER-Ekman dataset contains 1,350 sentences. We explored various strategies and models, and reported their F1-scores on both datasets.

4.1 Embodied Emotion Detection

Simple Prompting Analysis. To obtain the binary classification results, we directly compare the logit probabilities of "*True*" and "*False*" tokens instead of using text generation. This approach ensures deterministic outputs by avoiding the randomness inherent in sampling-based decoding methods, while also preventing potential output format violations that can occur during free-form generation. Table 2 shows that simplified prompts led to substantial performance improvements for the 70B parameter models. The F1-score increased by 29.5 points for Llama-3.1-70B, and by 41.6 points for DeepSeek-R1-70B (distilled on Llama), surpassing GPT 3.5 results reported in Zhuang et al. (2024).

Model	Macro F1	EE			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
Llama$_{\text{base}}$	37.2	21.5	99.6	35.3	99.6	24.2	39.0
Llama$_{\text{simple}}$	66.7	37.5	89.3	52.8	96.9	68.9	80.6
DeepSeek$_{\text{base}}$	32.6	20.3	99.4	33.7	99.3	18.7	31.5
DeepSeek$_{\text{simple}}$	74.2	51.2	69.3	58.9	93.1	86.2	89.5
GPT 3.5$_{\text{base}}$	70.2	44.0	68.3	53.5	92.5	81.9	86.9
BERT	83.5	73.2	72.1	72.6	94.2	94.5	94.4

Table 2: Results comparison for Embodied Emotion Detection. Llama: Llama-3.1-70B. DeepSeek: DeepSeek-R1-Distilled-Llama-70B. GPT 3.5 and fine-tuned BERT numbers are from Zhuang et al. (2024). The *base* and *simple* subscripts indicate the type of prompts, which can be found in Table 5.

Model	Macro F1	EE			Neutral		
		Pre	Rec	F1	Pre	Rec	F1
Llama$_{\text{2-step}}$	53.4	26.2	80.8	39.6	93.0	52.7	67.2
Llama$_{\text{3-step}}$	54.8	24.9	53.4	34.0	87.3	66.5	75.5
Llama$_{\text{2-step-simple}}$	60.1	31.5	44.5	36.9	87.4	79.8	83.4
DeepSeek$_{\text{2-step}}$	52.2	26.4	90.8	40.9	96.1	47.3	63.4
DeepSeek$_{\text{3-step}}$	57.4	27.9	62.0	38.5	89.4	66.7	76.4
DeepSeek$_{\text{2-step-simple}}$	67.5	40.1	65.2	49.7	91.7	79.8	85.3

Table 3: CoT results for Embodied Emotion Detection. Llama: Llama-3.1-8B. DeepSeek: DeepSeek-R1-Distilled-Llama-8B. The *2-step*, *2-step-simple*, and *3-step* subscripts indicate the type of prompt accompanying the model in that run. Prompt details are in Table 6.

Chain-of-Thought (CoT) Analysis. Table 3 shows that CoT prompting enhanced performance to competitive levels with larger models in the experiments of Table 2, particularly benefiting distilled reasoning models like DeepSeek-R1-8B (distilled on Llama). The DeepSeek 8B model using simple 2-step prompts (DeepSeek$_{\text{2-step-simple}}$) achieved results within 6.7 F1-points of its larger 70B counterpart (DeepSeek$_{\text{simple}}$) and 2.7 F1-points of GPT 3.5. Deeper reasoning processes proved more effective, with 3-step CoT consistently outperforming 2-step variants across both models. Finally, simplified prompting substantially improved CoT performance, yielding F1-score increases of 6.7 (Llama$_{\text{2-step-simple}}$ vs. Llama$_{\text{2-step}}$) and 15.3 (DeepSeek$_{\text{2-step-simple}}$ vs. DeepSeek$_{\text{2-step}}$).

Error Analysis. To investigate model failures in the zero-shot experiments using the *simple* prompt setting with Llama-3.1-70B, we analyzed incorrect predictions. We found a pronounced false-positive bias, accounting for 93.3% of all errors. A manual inspection of 100 false-positive cases revealed

Model	F1	F1-J	F1-Sa	F1-F	F1-A	F1-D	F1-Su
Llama	31.6	39.4	43.6	26.6	32.2	19.1	28.5
DeepSeek	28.4	43.3	35.7	33.1	23.1	14.8	20.2
BWS$_{4N}$	41.8	62.3	57.7	37.9	28.1	30.1	33.9
BWS$_{12N}$	44.6	67.1	59.2	44.3	38.6	19.0	39.3
BWS$_{36N}$	**50.6**	66.7	**64.7**	48.0	**53.2**	22.0	**48.9**
BWS$_{48N}$	49.8	68.0	62.8	48.2	51.3	24.8	43.8
BWS$_{72N}$	49.5	**68.5**	64.5	46.2	51.6	20.5	45.6
BERT	49.6	68.2	57.5	**50.1**	30.2	**56.1**	35.7

Table 4: Results for Emotion Classification. Llama: Llama-3.1-8B. DeepSeek: DeepSeek-R1-Distilled-Llama-8B. BWS: Automatic BWS with Llama-3.1-8B. The first column F1 is the macro-averaged score, followed by F1-score F1-x, where J - *Joy*, Sa - *Sad*, F - *Fear*, A - *Anger*, D - *Disgust*, and Su - *Surprise*.

three main patterns. First, 17% of cases involved referenced body parts that were present in the experience or expression without acting, as in "*tears falling down the* ***face***." Second, 42% of errors stemmed from body parts performing functional or physiological roles within emotional contexts, such as eyes closing when "*blackness crept across his* ***eyes***," a natural physiological reaction associated with the character passing away within the context. Finally, 41% of errors involved metaphorical or idiomatic expressions. These included cases where emotional embodiment was implied but not explicitly stated ("*I couldn't believe my* ***eyes***," implying widened eyes in surprise but not explicitly describing this action), expressions symbolically referring to emotional states without literal physical embodiment ("*a straw that broke my* ***back***"), or purely locational expressions involving body parts without any action ("*thoughts racing through my* ***head***"). These nuanced distinctions highlight the model's challenges in accurately interpreting metaphorical, symbolic, and non-embodied references.

4.2 Embodied Emotion Classification

Our experimental results demonstrate notable performance disparities across prompting strategies and model architectures. We use the Llama-3.1-8B as the LLM interpreter for best-worst scaling (BWS). In Table 4, the first section shows the performance of zero-shot large language models, including Llama-3.1-8B and DeepSeek-R1-8B. The second section shows BWS results with different numbers of tuples. And the last section shows a fine-tuned BERT model for comparison (details in Appendix C). We see that BWS exhibits superior performance even with smaller tuple configurations ($4N$), exceeding Llama-3.1-8B by 10.2 points. Performance improves consistently as the number of tuples increases from $2N$ to $36N$ (40.2 to 50.6), suggesting enhanced classification from expanded pairwise comparisons. We hypothesize that this expansion helps the model better weigh emotional significance in text, improving classification accuracy. Notably, the best result comes from the expansion to $36N$ tuples, with the F1-score beating the supervised method BERT by 1 point. Our experimental results show that as we keep increasing N, the performance will reach a plateau as evidenced by $48N$ and $72N$ (see Appendix B).

Error Analysis. To better understand model limitations, we conducted a qualitative error analysis on the misclassified cases. We identified several consistent failure modes, including the model's difficulty in interpreting emotionally complex inputs or making reliable distinctions between closely related emotional states. In one case, the model predicted *Joy*, even though the embodied expression *"Ryan ducks his* ***head*** *down to his notebook"* signaled *Fear*. This misclassification likely resulted from the influence of nearby positive context, such as *"Brendon waves and smiles"*, which distracted the model from the emotion-relevant phrase. In another case, the vivid scene *"the age-old rock tradition of holding up lighters spread across the 28,000 person deep crowd . . . lighting up the entire audience . . . the hair on my* ***arms*** *started to raise"* was misclassified as *Joy*, despite strong physiological cues such as raised arm hair that more closely reflect *Surprise*. This highlights the model's tendency to prioritize surface-level celebratory language over conflicting embodied cues.

5 Conclusion

This work advances embodied emotion recognition through three main contributions. First, we created a new dataset called CHEER-Ekman by extending the CHEER dataset with Ekman emotion labels to better understand the connection between bodily expressions and emotional states. Second, we demonstrated that best-worst scaling outperforms both prompted LLMs and fine-tuned BERT, showing the potential for emotion classification without task-specific training. Finally, we found that simplified language and chain-of-thought reasoning significantly improve LLM performance in embodied emotion detection, enabling smaller models to achieve competitive results.

Limitations

While our approach demonstrates a promising advancement in embodied emotion detection using LLMs and the best-worst scaling technique, several limitations warrant consideration.

First, a key observation in our embodied emotion detection task was that simplifying prompts significantly improved model performance. While these findings may suggest enhanced efficiency through linguistic streamlining, they simultaneously introduce concerns about potential overfitting to these simplified phrasings. Simplified prompts may inadvertently prioritize more explicit expressions of embodied emotion over subtler or more figurative language, meaning the models might learn to recognize patterns specific to the prompt structure rather than generalizing to a wide variety of natural language expressions.

Second, the CHEER-Ekman dataset is relatively small, consisting of only 1,350 sentences. This limited size stems from our decision to annotate only the sentences already identified as containing embodied emotions in the original CHEER dataset. This selective annotation was intended to efficiently focus our efforts on instances most relevant to embodied emotion, but it may introduce a bias towards positive examples.

Finally, when addressing emotion classification via best-worst scaling, the scalability and computational overhead of this methodology present challenges: while higher tuple quantities lead to higher accuracy, they also impose significant computational costs. Due to time and computational constraints, we utilize a smaller model, which may lead to suboptimal results for higher-order tuples. Additionally, the limited context window prevents us from effectively implementing a few-shot setting, further impacting performance in scenarios requiring extended context understanding.

Acknowledgments

We thank the CincyNLP group for their constructive comments. We also thank the anonymous ACL reviewers for their valuable suggestions and feedback.

References

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

Alan Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, 114.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, and Peiyi Wang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3).

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020 (Findings of EMNLP 2020)*.

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

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11).

George Lakoff and Mark Johnson. 1999. *Philosophy in the flesh: the embodied mind and its challenge of western thought*. Basic Books. OCLC: 302020239.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the 36th AAAI conference on artificial intelligence (AAAI 2022)*.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

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 (KDD 2024)*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

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 (SemEval 2018)*.

Paula M. Niedenthal. 2007. Embodying Emotion. *Science*, 316(5827):1002–1005.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007)*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. In *Findings of the Association for Computational Linguistics ACL 2024 (Findings of ACL 2024)*.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2024. My heart skipped a beat! recognizing expressions of embodied emotion in natural language. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

A LLM Prompt for Embodied Emotion Detection & Embodied Emotion Classification

Table 5-7 presents the complete prompt templates employed in our experimental methodology. Table 5 details the prompts used for zero-shot embodied emotion detection experiments. The *Base* prompt closely replicates the methodology of Zhuang et al. (2024), with the sole modification being the use of "True" and "False" as decision tokens rather than "Yes" and "No". The *Simple* prompt maintains the fundamental logic while employing more straightforward language and structure to reduce cognitive complexity.

Table 6 presents our chain-of-thought (CoT) prompt variants. The *2-step* implementation incorporates the dual criteria from the Base prompt as explicit reasoning steps. The *3-step* variant augments this with an initial body movement identification phase, designed to establish concrete context and facilitate more comprehensive reasoning. The *2-step simple* variant examines the effectiveness of linguistic simplification within the CoT framework.

Table 7 outlines the large language model (LLM) prompt utilized for emotion classification experiments with Llama-3.1-8B and DeepSeek-R1-8B models.

Throughout our prompt templates, we employ the following placeholder semantics:

- "`<sentence|>`" denotes the target sentence containing the body part for evaluation.
- "`<bdypart|>`" indicates the specific body part instance within the sentence.
- "`<preceed|>`" represents up to three preceding context sentences, when available.

This placeholder convention remains consistent across all experimental tasks presented in this research.

B Best-Worst Scaling

Experiment Setup. Best-worst scaling (BWS) is a comparative annotation method where annotators select the best and worst items from a given set, typically a 4-tuple. This approach efficiently derives pairwise comparisons, as selecting the best and worst items provides information about most item relationships within the set. A single annotation with best-worst scaling is equivalent to an annotation with 6 pairwise comparisons. Hence, using BWS allows for fewer inferences with the same result. This is particularly beneficial for identifying emotion intensity or emotion classification.

These 4-tuples are assembled from the test data and then are presented to LLMs using the prompt in Table 8. The model then picks one instance that most represents and one instance that least represents some property (in our case, this would be one of the six Ekman's (1992) emotions).

Once multiple 4-tuples are annotated, a simple counting procedure generates numerical scores, allowing items to be ordered according to their relevance to the given property. The score is calculated using the formula $\frac{\#Best - \#Worst}{\#Overall}$, where the $\#Best$ and $\#Worst$ represent the number of times a sentence is ranked Best or Worst, respectively; and $\#Overall$ denotes the number of occurrences of the sentence across all 4-tuples. This approach captures a continuous measure, reflecting the relative intensity of a sentence within the given category.

With this method, the LLMs can perform accurate annotations. The results from the annotations will be calculated to get a BWS intensity score across 6 Ekman emotions. The emotion with the highest intensity score will be chosen as the predicted label. This classification process can be represented by the following expression:

$$\hat{e} = \arg\max_{e_i \in E} S(e_i)$$

where e_i corresponds to each emotion, E is the set of all emotions, and $S(e_i)$ is the intensity score w.r.t. to such emotion. The resulting predictions are compared with the labels from the CHEER-Ekman dataset to assess performance using several metrics, with a particular focus on the F1-score. The approach of increasing the number of tuples to enhance performance was proposed by Bagdon et al. (2024).

Along with the embodied emotion classification prompt, we also incorporate two additional placeholder semantics:

- "`<textid|>`" denotes the unique instance ID from the dataset. This id helps the model easily pick out its answer from the sentence tuple when inferencing.
- "`<emo|>`" denotes the specific emotion required for ranking.

This placeholder convention remains consistent across all experimental tasks presented in this research.

Performance Plateau. Figure 7 illustrates the relationship between tuple count and F1-score performance in our BWS experiments. The results demonstrate significant performance improvements up to $24N$ tuples, reaching optimal performance at $36N$. Beyond this threshold, performance degradation is observed, with decreased F1-scores at both $48N$ and $72N$ tuple configurations. This pattern suggests a clear upper bound for effective tuple scaling in BWS implementations.

C BERT Experiment Setup

For the embodied emotion classification task discussed in Section 3.3, in addition to BWS, we fine-tuned BERT as a reference benchmark (Devlin et al., 2019). Inputs were constructed by concatenating the preceding context, main text, and referenced body part. We set the maximum sequence length to 512 and use a batch size of 16. The model was trained with the AdamW optimizer, a learning rate of 2e-5, and evaluated over 15 training epochs. Cross-entropy loss was used as the objective, with tokenization performed using truncation and padding. All runs were conducted with 5 seed values starting from 41 to 45 for reproducibility, and final results were averaged across these five runs.

D Data and Results Analysis

Figure 3 - 6 present a more detailed analysis of the CHEER-Ekman dataset and evaluation of models' performance. Figure 3 illustrates the frequency of the top 10 body parts associated with each emotion, where the size of the bubble reflects the co-occurrence of the body part and emotion pair. Notably, the body parts *face*, *eye*, *head*, *hand*, and *throat* appear consistently across the top 10 body parts across all emotions, with the highest frequency observed in *face*, *eye*, and *head*.

Figure 4 illustrates the classification performance of the 10 most frequent body parts, with their frequency and corresponding accuracy across the three language models: Llama-3.1-8B, DeepSeek-R1-8B, and fine-tuned BERT. As expected, the fine-tuned BERT model consistently outperforms both Llama and DeepSeek for the most frequent body parts. Generally, the fine-tuned BERT model outperforms both zero-shot Llama and Deepseek, achieving an average accuracy increase of 11.7 over Llama and 14.8 over DeepSeek across the top 10 frequent body parts.

Figure 5 and 6 present confusion matrices comparing the models' predicted emotions against the ground truth emotion for Llama and DeepSeek, respectively. When comparing the two figures, a notable pattern emerges. In Figure 5, strong activations across the diagonal indicate Llama's attempt to predict emotions accurately without bias towards any one emotion. In Figure 6, however, we observe a prominent concentration in *Joy* in the DeepSeek model.

Task	Prompt Template
Base	Please determine if a body part is involved in any embodied emotion. Specifically, a body part is involved in some embodied emotion if both conditions below are satisfied: 1) The physical movement or physiological arousal involving the body part is evoked by emotion. 2) The physical movement, if there is any, has no other purpose than emotion expression. Answer "True" if the body part is involved in any embodied emotion, and "False" otherwise. Preceding Context: `<preceed|>` Sentence: `<sentence|>` Body part: `<bdypart|>` Answer:
Simple	Decide if a body part is used purely to express emotion. Ask: - Did emotion cause the body part's movement/response? - Was the movement ONLY for expressing emotion (no other reason)? If both are true, say "True." Else, say "False." Preceding Context: `<preceed|>` Sentence: `<sentence|>` Body part: `<bdypart|>` Answer:

Table 5: Zero-shot templates for different tasks.

Setting	Prompt Template
2-Step	Please determine if a body part is involved in any embodied emotion. First, answer Condition 1: Is the body part's movement/arousal caused by emotion? Then, answer Condition 2: Does the movement lack non-emotional purposes? If both of those conditions are true, answer "True." Otherwise, answer "False." Please reason step-by-step for your answer. Here is the question: Preceding Context: `<preceed|>` Sentence: `<sentence|>` Body part: `<bdypart|>`
3-Step	Please determine if a body part is involved in any embodied emotion. Specifically, a body part is involved in some embodied emotion if both conditions below are satisfied: Before answering, reasoning step-by-step 1. Identify the body part mentioned. 2. Check if emotion directly caused its movement/arousal. 3. Verify if the movement has no functional purpose. Only if all of the above are true, answer "True." Otherwise, answer "False." Here is the question: Preceding Context: `<preceed|>` Sentence: `<sentence|>` Body part: `<bdypart|>`
2-Step Simple	Decide if a body part is used purely to express emotion. Ask: - Did emotion cause the body part's movement/response? - Was the movement ONLY for expressing emotion (no other reason)? If both are true, say "True." Else, say "False." Before answering, give your reasoning step-by-step. Preceding Context: `<preceed|>` Sentence: `<sentence|>` Body part: `<bdypart|>`

Table 6: Chain of Thought (CoT) prompt templates for different settings.

Prompt Template

Classify the emotion expressed by the body part in a sentence into one of six categories: "Joy", "Sadness", "Anger", "Fear", "Surprise", or "Disgust".

Preceding Context: `<preceed|>`
Sentence: `<sentence|>`
Body part: `<bdypart|>`
Answer:

Table 7: Emotion classification prompt template.

Prompt Template

You are an expert annotator specializing in emotion recognition. Rank the following examples based on how much `<emo|>` the specified body part exudes in the text.

Instructions:

- Use only the Preceding Text for context.
- Identify which example conveys the MOST `<emo|>` and which conveys the LEAST `<emo|>` based on the body part mentioned.
- Do not repeat the text. Only provide the Example numbers in the specified format.

Example: `<textid|>`
Preceding Context: `<preceed|>`
Sentence: `<sentence|>`
Body part: `<bdypart|>`

Example: `<textid|>`
Preceding Context: `<preceed|>`
Sentence: `<sentence|>`
Body part: `<bdypart|>`

Example: `<textid|>`
Preceding Context: `<preceed|>`
Sentence: `<sentence|>`
Body part: `<bdypart|>`

Example: `<textid|>`
Preceding Context: `<preceed|>`
Sentence: `<sentence|>`
Body part: `<bdypart|>`

Format your response as:
Most `<emo|>` Example:
Least `<emo|>` Example:

Table 8: BWS-Emotion classification prompt template.

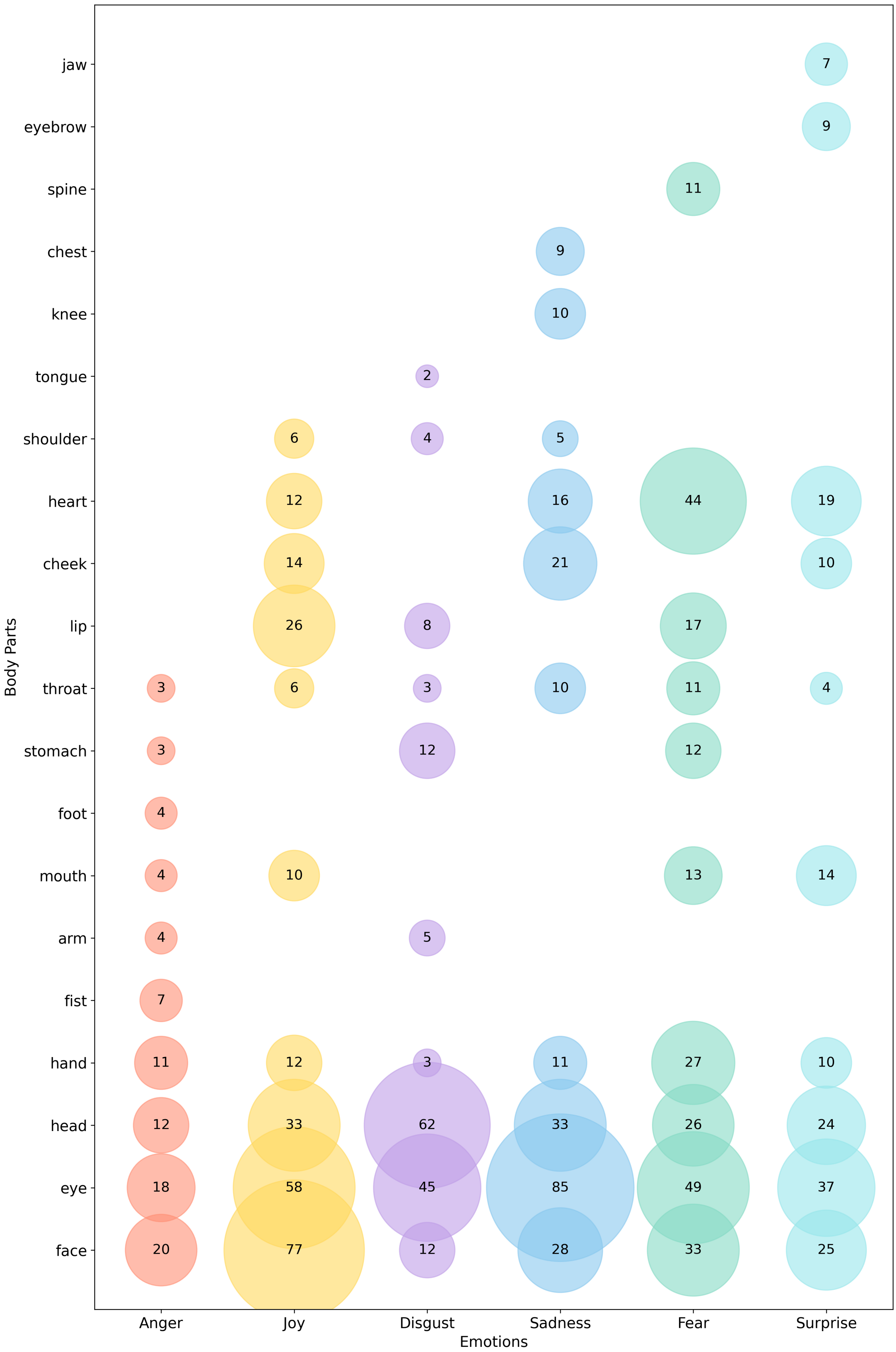


Figure 3: Frequency of top 10 body parts for each emotion.

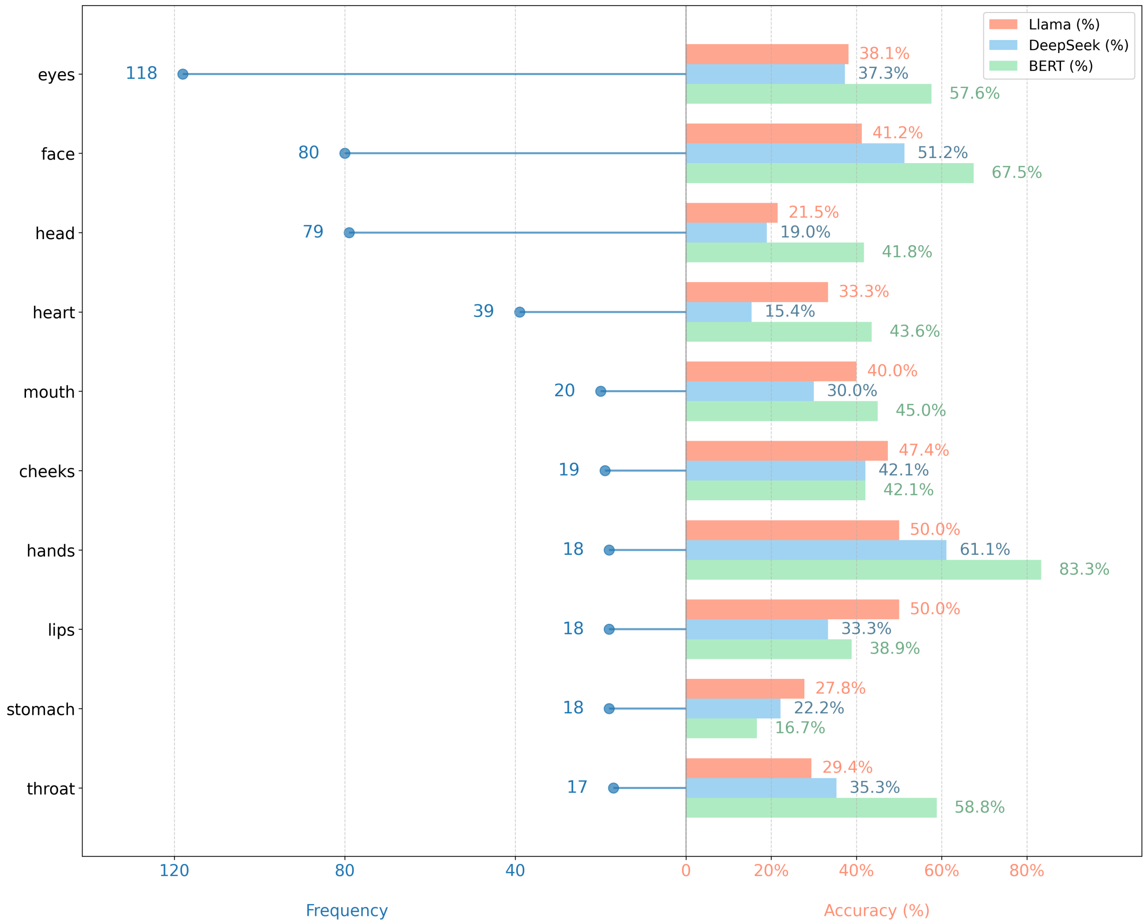
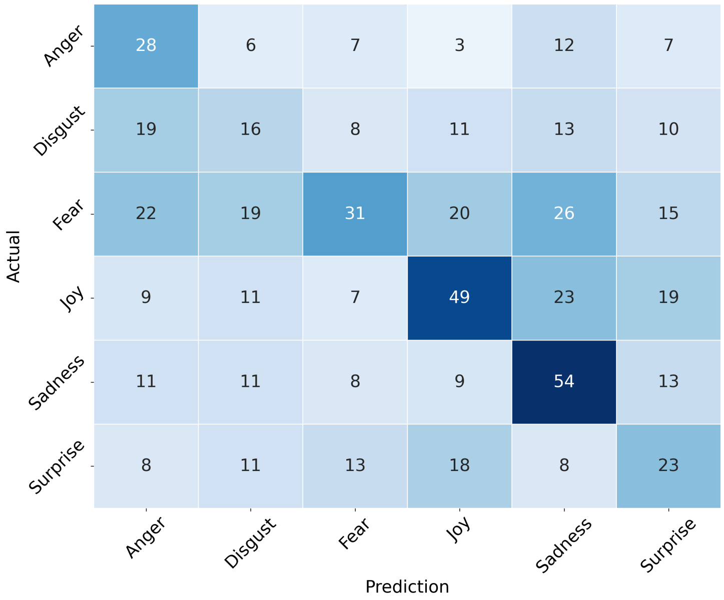


Figure 4: Embodied emotion classification of top 10 frequent body parts with frequency (left) and accuracy (right) comparing zeroshot with Llama, DeepSeek and finetuned BERT.



	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger	28	6	7	3	12	7
Disgust	19	16	8	11	13	10
Fear	22	19	31	20	26	15
Joy	9	11	7	49	23	19
Sadness	11	11	8	9	54	13
Surprise	8	11	13	18	8	23

Figure 5: Confusion matrix of Llama predictions.

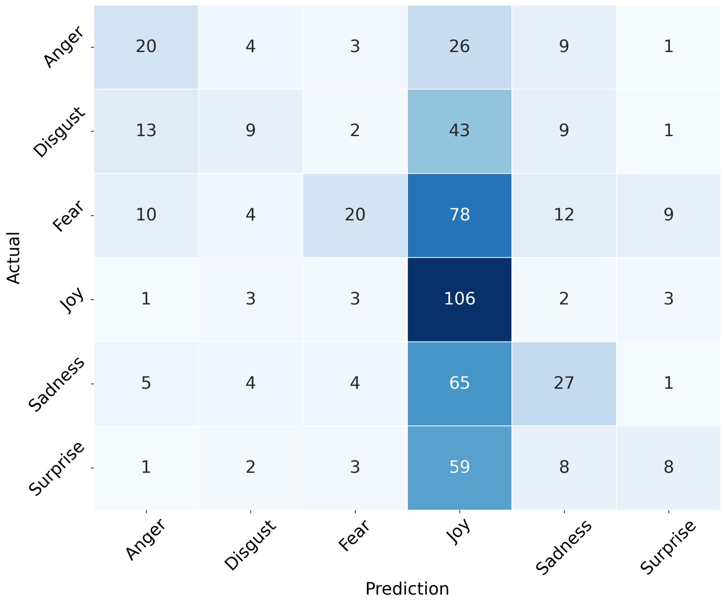


Figure 6: Confusion matrix of DeepSeek predictions.

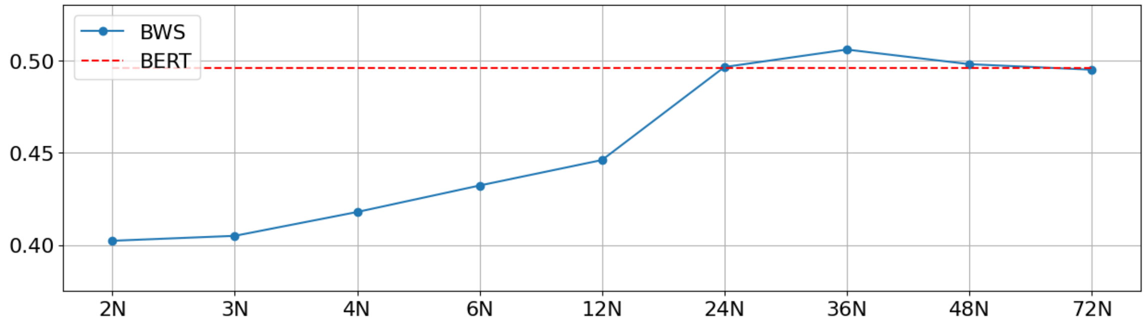


Figure 7: F1-score trends for BWS when increasing the number of tuples from 2N to 72N (where N is the total number of instances to be classified). BERT's performance is shown as a reference baseline.