Howard University-AI4PC at SemEval-2025 Task 11: Combining Expert Personas via Prompting for Enhanced Multilingual Emotion Analysis

Amir Ince **Saurav K. Aryal**
Howard University
Department of Computer Science and Electrical Engineering
amir.ince@bison.howard.edu, saurav.aryal@howard.edu

Abstract

For our approach to SemEval-2025 Task 11, we employ a multi-tier evaluation framework for perceived emotion analysis.Our system consists of several smaller-parameter large language models, each independently predicting the perceived emotion of a given text while explaining the reasoning behind its decision. The initial model's persona is varied through careful prompting, allowing it to represent multiple perspectives. These outputs, including both predictions and reasoning, are aggregated and fed into a final decision-making model that determines the ultimate emotion classification. We evaluated our approach in official SemEval Task 11 on subtasks A and C in all the languages provided.

1 Introduction

SemEval-2025 Task 11 (Muhammad et al., 2025b) focuses on detecting perceived emotion in a given text. Understanding emotion in natural language is an inherently complex task as the author not only expresses an emotion, but each reader may perceive a different emotion based on linguistic, cultural, and contextual factors. In natural language processing, emphasis is traditionally placed on the emotion explicitly expressed in the given text (Plaza-del Arco et al., 2024); however, perceived emotion detection aims to predict the emotion evoked in the audience, which may ultimately differ from the emotion portrayed by the author. These challenges are amplified in multilingual settings, where variations in word connotations, tone, and idiomatic expressions often lead to subjective and inconsistent interpretations of emotion (Havaldar et al., 2023).

To address these challenges, SemEval-2025 Task 11 introduces a multilingual perceived emotion detection task [1], to bridge the gap in NLP systems' ability to handle perceived emotion. This task consists of 3 tracks,

- Track A: Multi-label Emotion Detection
- Track B: Emotion Intensity
- Track C: Cross-lingual Emotion Detection

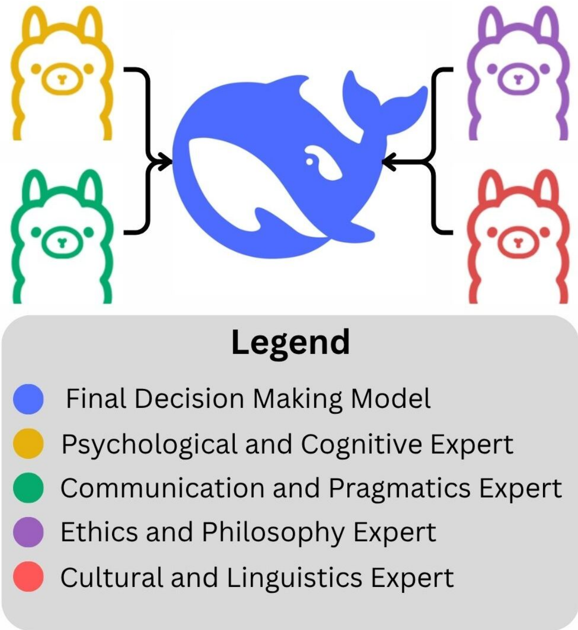

Figure 1: System Diagram

To address this problem, we developed a multi-tier evaluation framework (see Figure 1), inspired by collaborative strategies for Large Language Models (Lu et al., 2024). This ensemble-based approach uses smaller-parameter models to independently analyze a given text, providing both their predicted perceived emotion and the reasoning behind it. These models adopt carefully designed prompts that assign each a distinct expert persona—Cultural and Linguistic Expert, Psychological and Cognitive Expert, Communication and Pragmatics Expert, and Ethics and Philosophy Expert—guiding their analysis from different perspectives for a more comprehensive understanding of emotion. The outputs from these experts are then aggregated by a larger-parameter model for the final prediction.

[1]Task data available at: https://github.com/emotion-analysis-project/SemEval2025-Task11

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1645–1655
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

1.1 Novelty of Our Approach

Our framework extends existing work through several key innovations:

- **Specialized Expert Personas**: Assigning expert roles to the smaller models enables multi-faceted analysis across cultural, psychological, communicative, and ethical dimensions.
- **Reasoning-based Predictions**: Beyond simple classification, the smaller models offer reasoning alongside predictions, providing transparency into their decision-making.
- **Ensemble Aggregation**: A larger model aggregates and synthesizes the outputs from these specialized experts, enhancing prediction accuracy and nuance.
- **Cross-cultural Consideration**: Incorporating cultural and linguistic expertise directly addresses the challenges of emotion detection across diverse languages and cultures, as emphasized in recent studies.

We evaluate our framework using the dataset provided for SemEval-2025 Task 11 (Muhammad et al., 2025a; Belay et al., 2025).

2 Related Work

Detecting emotion in text has been a significant area of research in natural language processing (NLP). Previous studies have explored various approaches, including lexicon-based methods, machine learning techniques, and using deep learning models (Machová et al., 2023; Aryal et al., 2022a). For example, Mohammad (2018) developed the NRC Emotion Lexicon, a resource widely used for text emotion analysis. More recent approaches have harnessed the capabilities of transformer-based models such as BERT for emotion detection tasks, demonstrating improved performance in emotion detection across multiple languages (Machová et al., 2023; Aryal et al., 2023a).

Several studies have attempted to address the challenge of multilingual emotion detection, Bianchi et al. (2022) introduced XLM-EMO, a multilingual emotion classification model that performs well across 32 languages. However, researchers have highlighted that using machine translations in multilingual datasets can be problematic as it has the potential to overlook language-specific characteristics of emotion verbalization(Bianchi et al., 2022; Aryal and Adhikari, 2023; Sapkota et al., 2023).

Our work builds on recent advancements in collaborative strategies for Large Language Models, as explored by Lu et al. (2024). These approaches leverage the strengths of multiple perspectives to enhance overall performance and robustness in complex NLP tasks, particularly in emotion analysis (EA).

3 System Overview

Our multi-tier evaluation framework[2] for perceived emotion detection combines specialized expert models with a larger aggregator model, all locally hosted using Ollama[3]. This architecture is designed to capture nuanced emotional perceptions across diverse linguistic and cultural contexts.

Expert Models

We deploy four instances of the Llama 3.2 3B model (AI@Meta, 2024), using Q4_K_M quantization (4-bit precision with grouped scaling factors for optimized memory efficiency). Each instance is configured with a distinct expert persona through tailored system prompts:

- Cultural and Linguistic Expert
- Psychological and Cognitive Expert
- Communication and Pragmatics Expert
- Ethics and Philosophy Expert

These expert models provide diverse analytical perspectives. System prompts used for persona customization are detailed in Appendix A.

Aggregator Model

The outputs from the expert models are synthesized by a larger aggregator, the DeepSeek R1 32B model (DeepSeek-AI, 2025), also quantized using Q4_K_M. The aggregator integrates the expert responses to produce the final emotion prediction. The final prediction prompt is provided in Appendix B.

[2] Source code: https://github.com/amirince/SemEval-2025-Task-11

[3] Ollama website: https://ollama.com/

4 Experimental Setup

Technical Implementation

All models (Llama 3.2 3B (AI@Meta, 2024) and DeepSeek R1 32B (DeepSeek-AI, 2025)) are locally hosted using Ollama. We utilized the Ollama Python library for programmatic interaction with these models. Expert personas are implemented by modifying the prompts and roles assigned to each Llama 3.2 3B model instance.

Datasets

We utilized the SemEval-2025 Task 11 dataset, which contains multilingual text examples, each labeled with perceived emotions. The dataset was already divided into training, development, and test sets from the competition organizers.

4.1 Language Detection with Custom Identification Library

The first step in our pipeline is determining the language of the given text using a custom-built language identification library. Accurate language detection is crucial to ensure that subsequent analysis is properly contextualized for each language.

Our language identification library was developed using the training data provided by SemEval-2025 Task 11 to construct a corpus for each supported language. The process involved the following steps:

- **Text Extraction**: Words were parsed from the development set examples.
- **Data Cleaning**: Unwanted characters, such as emojis and punctuation, were removed to standardize the text.
- **Bag-of-Words Creation**: A bag of words was generated for each language present in the dataset.
- **Percentage Match**: To identify the language of an input text from the test set, the text was compared against each language's bag of words. The language with the highest percentage match was selected as the detected language.

4.2 Expert Analysis

For each example, the four expert models (Llama 3.2 3B (AI@Meta, 2024) variants) analyze the text independently of each other. Each expert provides a prediction of the emotion(s) and detailed reasoning for the prediction.

4.3 Intermediate Storage

The outputs from all expert models, including predictions, reasoning, and language detected, are stored in a CSV file. This allows for easy retrieval and analysis of intermediate results.

4.4 Aggregation Prompt Creation

We craft a comprehensive prompt that incorporates all the outputs of the expert models. This prompt provides the aggregator model with full context from the expert opinions and reasoning. One prompt is created for each example.

4.5 Final Prediction

The aggregated prompt is fed into the DeepSeek R1 32B (DeepSeek-AI, 2025) model. This model processes the collective expert insights and generates a final output.

4.6 Result Extraction

We parse the output from the DeepSeek model to extract the final emotion label for the given text.

5 Results and Analysis

Our multi-tier evaluation framework for perceived emotion detection demonstrated varying performance across different languages and emotions in Track A of the SemEval-2025 Task 11. Below is a detailed analysis of the results.

Note: Due to space constraints, the complete results are provided in Appendix C

5.1 Track A: Multi-label Emotion Detection

5.1.1 Overall Performance

The system's performance varied significantly across languages, with F1 scores ranging from 0.1284 (Makhuwa) to 0.6288 (Hindi). This wide range suggests that the effectiveness of the model is highly dependent on the language being processed.

5.1.2 Top Performing Languages

Table 1 presents the languages with the highest overall F1 scores, highlighting areas of strong model performance. Notably, Hindi and Marathi—both belonging to the Indo-Aryan language family—achieved top results, suggesting that the model may effectively leverage shared linguistic features within this group to enhance emotion detection.

Language	Avg. F1 Score
Hindi (hin)	0.6288
Russian (rus)	0.5896
Marathi (mar)	0.5838
Spanish (esp)	0.5696

Table 1: Top Performing Languages Track A Ranked by Average F1 Score Across Emotions

5.1.3 Low Performing Languages

Table 2 shows the languages in which the system performed poorly, likely due to limited training data or distinctive linguistic characteristics.

Language	Avg. F1 Score
Makhuwa (vmw)	0.1284
Yoruba (yor)	0.1357
Kinyarwanda (kin)	0.1657
Somali (som)	0.1601

Table 2: Lowest Performing Languages Track A Ranked by Average F1 Score Across Emotions

5.1.4 Emotion-Specific Performance

The performance of emotion detection varies across different languages. **Joy** demonstrates high performance across many languages, with particularly strong results in Swedish (0.7553) and Hindi (0.7834). **Anger** also performs well, especially in German (0.6922) and Chinese (0.7653). **Sadness** shows mixed results, with strong performance in Spanish (0.6576) but significantly weaker detection in Makhuwa (0.2523). **Fear** exhibits high variability, ranging from very low performance in Sundanese (0.0645) to very high accuracy in Russian (0.7426). **Surprise** generally has lower detection performance across languages, indicating difficulty in recognizing this emotion. Finally, **Disgust** consistently scores low, suggesting significant challenges in its detection across different linguistic contexts.

5.1.5 Language Family Trends

Overall, the performance of emotion detection varied across language families. **Indo-European languages**, such as Hindi, Spanish, and German, generally performed well. In contrast, **Afroasiatic languages**, including Somali and Hausa, exhibited mixed results. Meanwhile, **Niger-Congo languages**, such as Yoruba and Igbo, showed lower performance, indicating greater challenges in detecting emotions within these languages.

5.1.6 Implications

The framework performs well in widely spoken languages like Hindi, Russian, and Spanish but struggles with low-resource languages, highlighting the need for better data collection and fine-tuning. Strong results for joy and anger suggest universal markers, while poor detection of disgust indicates areas for improvement. Performance variability across language families also points to the potential of transfer learning between related languages.

5.2 Track C: Cross-lingual Emotion Detection

5.2.1 Overall Performance

The system's performance varied significantly across languages, with F1 scores ranging from 0.1397 (Amharic) to 0.5127 (Romanian), indicating language-dependent effectiveness in cross-lingual emotion detection.

5.2.2 Top Performing Languages

As shown in Table 3, the following languages achieved the highest overall F1 scores. The strong performance in Romanian and Hindi suggests that shared linguistic features may aid cross-lingual emotion detection within the Indo-European family.

Language	Avg. F1 Score
Romanian (ron)	0.5127
Hindi (hin)	0.5015
Algerian Arabic (arq)	0.4180
Javanese (jav)	0.3749

Table 3: Top Performing Languages Track C Ranked by Average F1 Score Across Emotions

5.2.3 Low Performing Languages

The system struggled the most with the languages listed in Table 4. These low scores suggest challenges in transferring emotion detection capabilities to Afroasiatic languages or reflect insufficient cross-lingual training data.

Language	Avg. F1 Score
Amharic (amh)	0.1397
Somali (som)	0.1634
Oromo (orm)	0.1760
Tigrinya (tir)	0.1791

Table 4: Lowest Performing Languages Track C Ranked by Average F1 Score Across Emotions

5.2.4 Emotion-Specific Performance

The framework exhibited consistently high performance in detecting **joy**, particularly in languages such as Romanian (0.7371) and Hindi (0.5981). **Sadness** was well detected in Javanese (0.5950) and Algerian Arabic (0.5571). **Fear** showed high variability, ranging from very low in Amharic (0.0600) to very high in Romanian (0.6775). **Anger** produced mixed results, with strong performance in Algerian Arabic (0.5193) but weaker performance in Mozambican Portuguese (0.1439). **Surprise** was generally difficult to detect across languages, suggesting challenges in cross-lingual transfer. Finally, **disgust** demonstrated inconsistent performance, ranging from very low in Kinyarwanda (0.0908) to moderate in Romanian (0.4167).

5.2.5 Language Family Trends

Indo-European languages, such as Romanian and Hindi, demonstrated the best performance. The **Austronesian language** Javanese also performed relatively well. In contrast, **Afroasiatic languages**, including Amharic, Somali, and Oromo, exhibited lower performance. Meanwhile, **Niger-Congo languages** like Swahili and Igbo showed moderate performance.

5.2.6 Implications for Cross-lingual Emotion Detection

Indo-European languages show strong potential in emotion detection, but cross-lingual transfer to Afroasiatic languages remains weak. Improving data collection and fine-tuning is essential for low-resource languages. The strength of the model in detecting joy and sadness suggests that these emotions have strong linguistic markers, which aid in transfer learning. However, consistently poor performance in detecting surprise highlights the need for better cross-lingual features. Leveraging linguistic similarities between related languages could further enhance performance.

These findings emphasize both the promise and challenges of cross-lingual emotion detection, with clear opportunities for improvement in linguistically diverse and low-resource languages.

6 Ethical Considerations

Bias and Fairness: Our framework showed varying performance across different languages, potentially leading to unequal treatment of speakers of different languages. This could result in bias against speakers of low-resource languages or languages not well-represented in the training data. Secondly, the reliance on pre-trained models like Llama 3.2 3B (AI@Meta, 2024) and DeepSeek R1 32B (DeepSeek-AI, 2025) may inherit biases present in their training data, potentially amplifying societal biases related to emotion expression across different cultures.

7 Conclusion

In this paper, we presented a multi-tier evaluation framework for perceived emotion detection in text, which demonstrated mixed performance across multiple tracks of the SemEval-2025 Task 11. Our system leverages a combination of specialized expert models based on Llama 3.2 3B (AI@Meta, 2024) and a powerful aggregator model using DeepSeek R1 32B (DeepSeek-AI, 2025), all locally hosted using Ollama. We showed that this approach can effectively capture nuanced emotional perceptions across diverse linguistic and cultural contexts, particularly excelling in Indo-European languages like Hindi and Romanian.

Our framework demonstrated strength in detecting emotions like joy and anger across multiple languages, suggesting these emotions may have more universal linguistic markers. The system's performance varied significantly across language families, with Indo-European languages generally outperforming others. This highlights our approach's potential for nuanced emotion detection while underscoring challenges in cross-lingual analysis. For future work, we aim to enhance the system's ability to handle linguistic diversity and improve performance on underrepresented languages and emotions. Addressing these challenges moves us closer to developing robust, multilingual emotion detection systems capable of capturing the complexities of human emotions across diverse cultural contexts.

8 Limitations

While our multi-tier evaluation framework for perceived emotion detection showed promise, it's important to acknowledge several limitations:

8.1 Limited Expert Model Customization

Our expert models were not fine-tuned for specific languages or emotions. Instead, their personas were varied via prompts. This approach, while flexible, means that languages or emotional contexts not well-represented in the original training data could lead to misclassifications or unintended biases.

8.2 Lack of Model Diversity

All our expert models were based on the same Llama 3.2 3B (AI@Meta, 2024) architecture. This uniformity may have limited the diversity of perspectives and could have amplified any inherent biases or limitations of the base model across all experts.

8.3 Incomplete Track Submissions

Due to computational constraints and time limitations, we were unable to submit results for Track B and only made a partial submission for Track C. This incomplete participation limits our ability to fully evaluate the system's performance across all aspects of the task.

8.4 Language Imbalance

The system's performance varied significantly across languages, with Indo-European languages generally outperforming others. This suggests a potential bias in the model towards more widely spoken or well-resourced languages.

8.5 Code-switched Text Evaluations

Often code-switching is a significant phenomenon in multilingual text where two more languages are utilized in a single sentence. In a globalizing world, these tertiary languages may require further analysis and evaluation (Aryal et al., 2023c,b, 2022b).

8.6 Emotion Detection Inconsistency

Certain emotions, particularly disgust and surprise, consistently showed lower performance across languages. This indicates a limitation in the model's ability to capture and transfer these emotional concepts across linguistic boundaries.

9 Acknowledgments

We would like to express our gratitude to the organizers of SemEval Task 11[4] for providing the competition framework that facilitated this research. Additionally, we acknowledge Muhammad et al. (2025a) and Belay et al. (2025) for their valuable multilingual emotion annotations, which were instrumental in our experiments. This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.

[4] https://github.com/emotion-analysis-project/SemEval2025-Task11

References

AI@Meta. 2024. Llama 3.2 model card. https://ai.meta.com/llama/.

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023a. Sentiment analysis across multiple african languages: A current benchmark. *arXiv preprint arXiv:2310.14120*.

Saurav K Aryal, Howard Prioleau, Surakshya Aryal, and Gloria Washington. 2023b. Baselining performance for multilingual codeswitching sentiment classification. *Journal of Computing Sciences in Colleges*, 39(3):337–346.

Saurav K Aryal, Howard Prioleau, and Gloria Washington. 2022a. Sentiment classification of code-switched text using pre-trained multilingual embeddings and segmentation. *arXiv preprint arXiv:2210.16461*.

Saurav K Aryal, Howard Prioleau, and Gloria Washington. 2022b. Sentiment classification of code-switched text using pre-trained multilingual embeddings and segmentation. *arXiv preprint arXiv:2210.16461*.

Saurav K Aryal, Howard Prioleau, Gloria Washington, and Legand Burge. 2023c. Evaluating ensembled transformers for multilingual code-switched sentiment analysis. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 165–173. IEEE.

Saurav Keshari Aryal and Gaurav Adhikari. 2023. Evaluating impact of emoticons and pre-processing on sentiment classification of translated african tweets. *ICLR Tiny Papers*.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. XLM-EMO: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 195–203, Dublin, Ireland. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. https://github.com/deepseek-ai/DeepSeek-R1.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models.

Kristína Machová, Martina Szabóova, Ján Paralič, and Ján Mičko. 2023. Detection of emotion by text analysis using machine learning. *Frontiers in Psychology*, 14.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

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.

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, 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 text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Hrishav Sapkota, Saurav Keshari Aryal, and Howard Prioleau. 2023. Zero-shot classification reveals potential positive sentiment bias in african languages translations. *ICLR Tiny Papers*.

A Expert Model Prompts

A.1 Cultural and Linguistic Expert Prompt

Task: Cultural and Linguistic Emotion Analysis
You are a cultural and linguistic expert specializing in analyzing emotions through the lens of language, cultural context, and sociolinguistic nuances. Your role is to identify and explain the emotions conveyed in the provided text while considering cultural nuances, idiomatic expressions, and the sociolinguistic factors that may influence emotional interpretation.
Instructions: 1. Analyze the text for emotional content, considering how cultural context and language usage shape emotional expression. 2. Identify emotions from the following list: {{possible_langs}}. 3. Note the language of the text: {{lang_id}}.
Text for Analysis: *"{{text}}"*
Deliverable: - Identify the emotions perceived in the text. - Provide a culturally sensitive explanation for each emotion identified, referencing idiomatic expressions or cultural factors where applicable. - Highlight any linguistic features (e.g., tone, word choice, syntax) that influenced your interpretation.
Note: The text may convey multiple emotions. Your analysis should be thorough and context-sensitive.

A.2 Psychological and Cognitive Expert Prompt

Task: Emotional Perception and Psychological Impact Assessment
You are a trained expert in psychology and cognitive science, specializing in the anal-

ysis of emotional tone, psychological responses, and cognitive processes that shape human perception. Your role is to assess the emotional tone of the given text, identify the emotions it evokes, and offer insights grounded in psychological theory.
Key Instructions: 1. Analyze the emotional tone of the provided text, considering both overt and subtle cues. 2. Identify and categorize the emotions conveyed, drawing on established psychological frameworks (e.g., the basic emotions theory, cognitive appraisal theory). 3. Explain the cognitive and psychological mechanisms that contribute to the perception of each identified emotion.
Emotions to consider: {{possible_langs}} (select all applicable emotions that fit the text).
Language of the given text: {{lang_id}}
Text for Analysis: *"{{text}}"*
In your response, explain: - Why you selected each identified emotion(s). - How the psychological or cognitive processes underlying these emotions might manifest in the text.
Note: The text may evoke a range of emotions. Feel free to identify and explain multiple emotions where applicable.

A.3 Communication and Pragmatics Expert

Task: Emotional and Pragmatic Response Analysis
You are an expert in communication, behavioral analysis, and natural language processing. Your task is to assess the emotional and pragmatic impact of the following text. Focus on how the language may influence the reader's emotions, behavioral responses, and overall interpretation.
Goals: - Identify the emotions conveyed by the text. - Evaluate how the text's language and tone might affect the reader's emotional state or behavior. - Consider implied meanings, subtext, and the potential impact of the text on the audience.
Emotions to consider: {{possible_langs}}
Language of the given text: {{lang_id}}
Text for Analysis: *"{{text}}"*
Explanation: - Provide a brief rationale for your choice(s) of emotion(s). - Highlight any subtext or implied meanings that influence emotional perception. - If the text has a mix of emotions, explain the potential shifts or contrasts in how a reader might emotionally react.
Note: The text can reflect multiple emotions or conflicting emotional cues.

A.4 Ethics and Philosophy Expert

Role: Examine the intentionality, ethical implications, and broader societal effects of the text's emotional expression.
Task: Perceived Emotion and Ethical Implication Detection
You are an expert in philosophy, language, and ethics. Your task is to analyze the given text by identifying the emotions it conveys, but with a deeper focus on the ethical dimensions and potential societal effects of these emotions.
In addition to recognizing the emotions in the text, consider the following:
Intentionality: What might the author intend to communicate with these emotions?
Ethical Implications: Are the emotions expressed fair, just, and morally sound? Do they align with standards of ethical communication?
Broader Societal Impact: How might these emotions influence the broader social context or affect the audience's understanding?
Emotions to consider: {{possible_langs}}
Language of the given text: {{lang_id}}
Text for Analysis: *"{{text}}"*
Ethical Considerations: - Provide a rationale for each emotion identified, specifically focusing on: - How the emotion aligns with moral standards. - The possible impact this emotion could have on social fairness or bias.
Note: The text may evoke multiple emotions; please explore the broader ethical context of each.

B Aggregator Model Prompt

B.1 Final Aggregator Prompt

Task: Final Emotion Determination
Review the Juror Assessments:
Carefully review the emotion assessments provided by the Jurors. Pay attention to the range of emotions identified, the frequency of specific emotions, and the level of confidence expressed by each Juror.
Consider the Following:
1. **Consensus:** - Identify emotions that have been consistently selected by multiple Jurors. - Prioritize emotions with strong consensus.
2. **Confidence Levels:** - Assess the confidence levels expressed by the Jurors. - Give more weight to emotions that have been identified with high confidence.
3. **Nuance and Complexity:** - Consider the possibility of multiple emotions or complex emotional states. - Look for subtle cues and underlying feelings that may not be explicitly stated.
For Context:
- This is the sample text the Jurors were asked to classify: *"{{text}}"* - The language of the above text is: {{lang_id}} - The possible emotions invoked by the text are: {{possible_emotions}}
Juror Assessments:
{{juror_assessment}}
Make a Final Decision:
Based on your analysis, determine the primary emotion(s) conveyed in the text.
Please only provide the final emotion(s) in your response. You do not need to explain your thought process.

C Complete Track Results

C.1 Track A Results

amh	anger	0.2955	hin	anger	0.5949	ptbr	anger	0.6285	swa	anger	0.232
	disgust	0.0032		disgust	0.3889		disgust	0.1008		disgust	0.0979
	fear	0.0366		fear	0.7208		fear	0.297		fear	0.0936
	joy	0.1856		joy	0.7834		joy	0.6116		joy	0.4392
	sadness	0.3333		sadness	0.586		sadness	0.4641		sadness	0.2483
	surprise	0.0699		surprise	0.6985		surprise	0.2136		surprise	0.2624
	average	0.154		average	0.6288		average	0.3859		average	0.2289
arq	anger	0.4772	ibo	anger	0.3197	ptmz	anger	0.2058	swe	anger	0.6003
	disgust	0.0284		disgust	0.0284		disgust	0		disgust	0.0167
	fear	0.392		fear	0.1164		fear	0.3139		fear	0.1739
	joy	0.265		joy	0.3759		joy	0.467		joy	0.7553
	sadness	0.4861		sadness	0.2168		sadness	0.4767		sadness	0.2
	surprise	0.1735		surprise	0.0522		surprise	0.2314		surprise	0.097
	average	0.3037		average	0.1849		average	0.2825		average	0.3072
ary	anger	0.4467	kin	anger	0.2851	ron	anger	0.4945	tat	anger	0.3569
	disgust	0.1818		disgust	0.0556		disgust	0.0565		disgust	0.0339
	fear	0.296		fear	0.086		fear	0.7245		fear	0.1848
	joy	0.5474		joy	0.2067		joy	0.7931		joy	0.4584
	sadness	0.309		sadness	0.3308		sadness	0.4157		sadness	0.4489
	surprise	0.1596		surprise	0.0299		surprise	0.1455		surprise	0.3243
	average	0.3234		average	0.1657		average	0.4383		average	0.3012
chn	anger	0.7653	mar	anger	0.6143	rus	anger	0.6507	ukr	anger	0.2113
	disgust	0.086		disgust	0.3306		disgust	0.4881		disgust	0.1709
	fear	0.2254		fear	0.6957		fear	0.7426		fear	0.5161
	joy	0.686		joy	0.6548		joy	0.7187		joy	0.5433
	sadness	0.4217		sadness	0.6072		sadness	0.4276		sadness	0.4149
	surprise	0.1709		surprise	0.6		surprise	0.51		surprise	0.2629
	average	0.3926		average	0.5838		average	0.5896		average	0.3532
deu	anger	0.6922	orm	anger	0.3225	som	anger	0.1903	vmw	anger	0.0923
	disgust	0.1314		disgust	0.014		disgust	0.0163		disgust	0
	fear	0.3187		fear	0.0713		fear	0.1639		fear	0.1553
	joy	0.6042		joy	0.3898		joy	0.2713		joy	0.1897
	sadness	0.468		sadness	0.1702		sadness	0.231		sadness	0.2523
	surprise	0.1939		surprise	0.0828		surprise	0.088		surprise	0.0811
	average	0.4014		average	0.1751		average	0.1601		average	0.1284
esp	anger	0.6495	pcm	anger	0.2296	sun	anger	0.2255	yor	anger	0.131
	disgust	0.2355		disgust	0.0671		disgust	0.1096		disgust	0
	fear	0.6557		fear	0.2973		fear	0.0645		fear	0.0713
	joy	0.688		joy	0.56		joy	0.6869		joy	0.1386
	sadness	0.6576		sadness	0.3535		sadness	0.5909		sadness	0.3933
	surprise	0.5314		surprise	0.2635		surprise	0.2632		surprise	0.08
	average	0.5696		average	0.2952		average	0.3234		average	0.1357
hau	anger	0.3274									
	disgust	0.2413									
	fear	0.2872									
	joy	0.2738									
	sadness	0.4889									
	surprise	0.29									
	average	0.3181									

C.2 Track C Results

	anger	0.1908	hin	anger	0.4058		anger	0.2936		anger	0.3829
	disgust	0.1703		disgust	0.3803		disgust	0.0908		disgust	0.4167
	fear	0.06		fear	0.4828		fear	0.1096		fear	0.6775
amh	joy	0.1284		joy	0.5981	kin	joy	0.2878	ron	joy	0.7371
	sadness	0.2379		sadness	0.5421		sadness	0.3623		sadness	0.4993
	surprise	0.051		surprise	0.6		surprise	0.0934		surprise	0.3629
	average	0.1397		average	0.5015		average	0.2062		average	0.5127
	anger	0.5193	ibo	anger	0.3181		anger	0.2877		anger	0.1319
	disgust	0.3426		disgust	0.308		disgust	0.2322		disgust	0.1914
	fear	0.3621		fear	0.1675		fear	0.0986		fear	0.1088
arq	joy	0.3535		joy	0.3548	orm	joy	0.2264	som	joy	0.2185
	sadness	0.5571		sadness	0.2885		sadness	0.1562		sadness	0.1793
	surprise	0.3736		surprise	0.0694		surprise	0.0549		surprise	0.1503
	average	0.418		average	0.251		average	0.176		average	0.1634
	anger	0.3824		anger	0.3755		anger	0.1439		anger	0.2064
	disgust	0.1012		disgust	0.1857		disgust	0.1021		disgust	0.1295
	fear	0.2179		fear	0.1667		fear	0.2387		fear	0.07
ary	joy	0.4681	jav	joy	0.5551	ptmz	joy	0.3636	swa	joy	0.3882
	sadness	0.3624		sadness	0.595		sadness	0.3944		sadness	0.2165
	surprise	0.2407		surprise	0.3714		surprise	0.1564		surprise	0.2592
	average	0.2955		average	0.3749		average	0.2332		average	0.2117
	anger	0.1892									
	disgust	0.2693									
	fear	0.0753									
tir	joy	0.2033									
	sadness	0.2078									
	surprise	0.1298									
	average	0.1791									