# UT-NLP at SemEval-2025 Task 11: Evaluating Zero-Shot Capability of GPT-4o Mini on Emotion Recognition via Role-Play and Contrastive Judging

**AmirHossein Safdarian*** University of Tehran a.safdarian@ut.ac.ir

**Milad Mohammadi*** University of Tehran miladmohammadi@ut.ac.ir

**Hesham Faili** University of Tehran hfaili@ut.ac.ir

## Abstract

Emotion recognition in text is crucial in natural language processing but challenging in multilingual settings due to varying cultural and linguistic cues. In this study, we assess the zero-shot capability of GPT-4o Mini, a cost-efficient small-scale LLM, for multilingual emotion detection. Since small LLMs tend to perform better with task decomposition, we introduce a two-step approach: (1) Role-Play Rewriting, where the model minimally rewrites the input sentence to reflect different emotional tones, and (2) Contrastive Judging, where the original sentence is compared against these rewrites to determine the most suitable emotion label. Our approach requires no labeled data for fine-tuning or few-shot in-context learning, enabling a plug-and-play solution that can seamlessly integrate with any LLM. Results show promising performance, particularly in low-resource languages, though with a performance gap between high- and low-resource settings. These findings highlight how task decomposition techniques can enhance small LLMs' zero-shot capabilities for real-world, data-scarce scenarios.

## 1 Introduction

SemEval-2025 Task 11 (Muhammad et al., 2025b) addresses emotion recognition in text across multiple languages, ranging from high-resource to low-resource languages, which is a crucial area in NLP with far-reaching applications in social media analytics, customer service, and healthcare.

By providing a multilingual, multi-labeled dataset of 28 languages, the task highlights the challenges of building robust emotion detection systems under limited training data conditions.

In this paper, we describe our team's participation in SemEval-2025 Task 11, specifically in:

- **Track B (Emotion Intensity):** Predicting ordinal intensity (0–3) for emotions such as joy, sadness, fear, anger, surprise, and disgust.
- **Track C (Cross-lingual Emotion Detection):** Zero-shot emotion detection in a target language using only training data from a different language.

Our primary goal was to evaluate the zero-shot capability of a small-sized LLM, GPT-4o Mini, which is a more cost-efficient model in the GPT-4o family (OpenAI et al., 2024). As LLMs have demonstrated impressive zero-shot capabilities in recent years (Kojima et al., 2022), we did not fine-tune the model on any task-specific data but instead introduced a zero-shot approach: **role-play** and **contrastive judging**, illustrated in Figure 1.

1. **Role-Play Rewriting:** Prompt the model to rewrite a given sentence as if it inherently conveyed a target emotion—altering only minimal surface details while preserving meaning.
2. **Contrastive Judging:** Have the model compare the original and rewritten sentences for each of the emotions, reason through the differences (via chain-of-thought), and then produce a final emotion score or label.

We hypothesize that rewriting the text to inject various emotional tones helps the model disentangle subtle cues, while contrastive judging ensures a reasoned final decision. By relying solely on prompting and structured reasoning, we aimed to investigate whether a smaller-scale LLM could discern subtle emotional cues without specialized training.

* Equal contribution, ordered randomly.



*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2183–2189

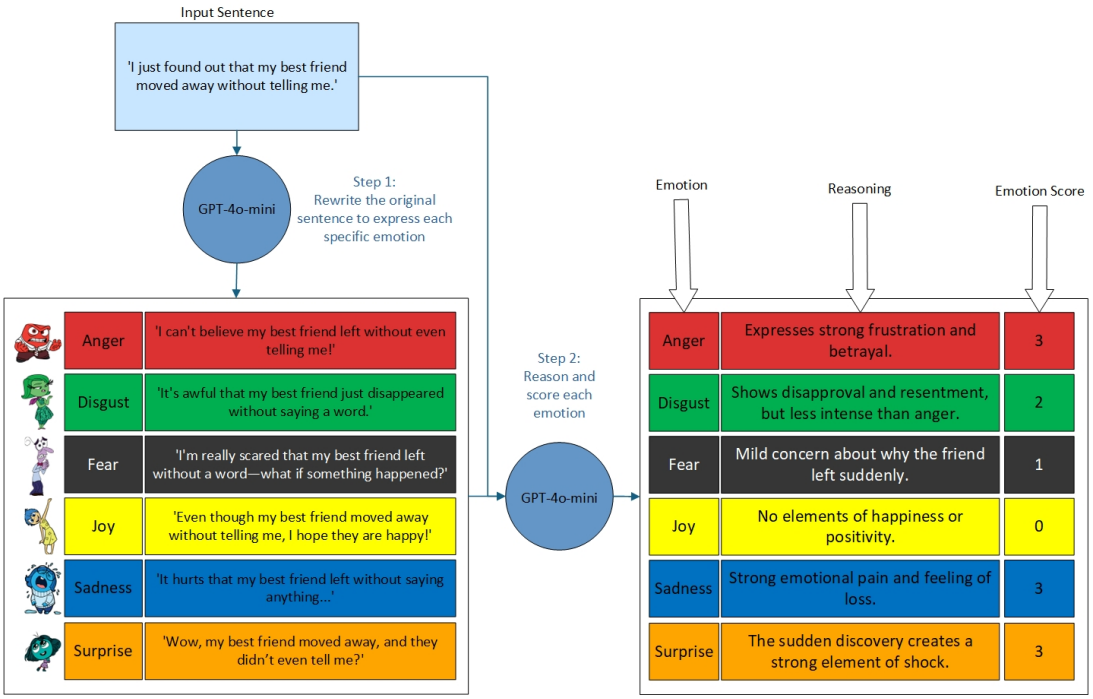


Figure 1: Brief overview of our system

Our zero-shot method, although not striving for state-of-the-art performance, yielded results around the baseline for both Track B and Track C. We observed that:

- The role-play mechanism helped the model clarify the emotional content in ambiguous texts.
- Contrastive judging offered explicit reasoning steps but was prone to occasional misclassifications, especially for nuances such as *fear* vs. *surprise*.
- Performance varied notably across languages, echoing the challenges of low-resource settings.

We have released our code, prompts, and intermediate outputs (rewritten sentences and model reasoning) on GitHub for reproducibility and further research GitHub Repository.

## 2 Background

### 2.1 Classic Approaches to Emotion Recognition

Early approaches to emotion recognition often relied on lexicons and feature-based machine learning. Lexicon-based methods drew on resources like WordNet-Affect and the NRC Word–Emotion Lexicon to match emotion words in a text (Nandwani and Verma, 2021). Although transparent and easy to use, such methods struggled with context and intensity (Nandwani and Verma, 2021). Feature-based supervised learning went further by encoding various cues (e.g., n-grams, emotion lexicon hits, negation) and training models like SVM or MaxEnt (Oberländer and Klinger, 2018; Nandwani and Verma, 2021).

### 2.2 Neural Networks and Transformers

Neural network approaches like LSTMs and CNNs automatically learn higher-level features. Studies showed bi-directional LSTMs outperformed linear models on emotion classification, while CNNs captured relevant n-grams (Oberländer and Klinger, 2018). Transformer-based architectures such as BERT advanced these gains by providing rich contextual representations. For instance, GoEmotions (58K Reddit comments with 27 emotion labels) showed a fine-tuned BERT achieving strong F1 scores, though still leaving room for improvement (Demszky et al., 2020). In dialogue scenarios, hierarchical models (e.g., DialogueRNN, HiGRU) incorporate utterance- and dialogue-level encoders to handle context across multiple turns (Zhu et al., 2021).

### 2.3 Large Language Models for Zero/Few-Shot Emotion Classification

LLMs like GPT-3.5 and GPT-4 can perform zero-shot classification via prompts without fine-tuning. However, prompt design is critical, as poorly phrased instructions can lead to suboptimal performance. (Kazakov et al., 2024). Cultural variance and domain mismatch pose further difficulties (Plaza-Del-Arco et al., 2024). While fine-tuned models often outperform LLM zero-shot prompts (Juan et al., 2024), some results show LLMs can close the gap on simpler tasks (Juan et al., 2024).

### 2.4 Our Work in Context

In this work, we explore a creative strategy for emotion detection leveraging the capabilities of GPT-4o Mini in a two-step process: a role-play rewriting step followed by a contrastive judging step.

In the first stage, the model is prompted to "role-play" – effectively rewriting or rephrasing the input text as if it were being expressed with a specific emotional stance. In the second stage, a contrastive evaluation is performed: the model (or a separate process) compares the original text to the emotion-specific paraphrases and judges which emotion's paraphrase best matches or explains the original.

This two-step approach is reminiscent of prompting techniques where the model is encouraged to reason or decompose the task before giving an answer (Bhaumik and Strzalkowski, 2024), that frames emotion detection as a generative question-answering problem – essentially asking the model to explain what might be happening or felt in the text before naming the emotion. This is analogous to our idea of role-play generation as a form of explanation. Moreover, the practice of using chain-of-thought (CoT) prompting for reasoning tasks has shown that LLMs can often improve accuracy by elaborating on the problem before answering (Wei et al., 2022).

By positioning our approach in this context, we aim to leverage both the generative flexibility and knowledge of an LLM and its ability to function as a classifier. There is little prior work that explicitly uses a role-play rewriting technique for emotion detection, so we believe this adds a fresh perspective to the toolkit of LLM-based emotion analysis. Our method aligns with the trend of using LLMs as reasoning engines that are more interpretable than classic methods.

## 3 System Overview

Our approach relies on a minimalistic two-step pipeline built using GPT-4o Mini (`gpt-4o-mini-2024-07-18`) as the processing core:

1. **Role-Play Rewriting**, where the model is prompted to rewrite the input sentence multiple times—once per candidate emotion—making minimal changes to reflect that emotion while preserving the original meaning.

2. **Contrastive Judging**, where a second call to GPT-4o Mini compares the original sentence to each of these rewrites and determines the best-matching emotion. (figure 1)

Both steps use OpenAI's structured output feature to parse the results, and no hyperparameter changes (e.g., temperature, max_tokens) were made. We leveraged the BRIGHTER dataset (Muhammad et al., 2025a) only for zero-shot inference—no training or fine-tuning was performed—and used the development set purely to refine our prompts.

Since our method is self-contained, switching LLM providers or models requires only updating the client and model name. The simplicity of the pipeline (two API calls, standardized prompt structure) and the absence of fine-tuning or other methods that require labeled data make our approach a plug-and-play solution.

## 4 Experimental Setup

We tested our pipeline on Track B and Track C of SemEval-2025 Task 11 using the official splits provided in BRIGHTER (Muhammad et al., 2025a) (train, dev, test). We crafted our prompts using the dev set (treating it only for evaluation and prompt iteration) and then ran the final prompts on the test data. This included languages of varying resource levels, from English and Chinese to Ukrainian and isiZulu.

We did not apply any explicit preprocessing or domain adaptation (e.g., removing special characters), relying on GPT-4o Mini's internal handling. For each test instance, we called the model twice: once per emotion candidate (for role-play rewriting) and once for the final contrastive judgment, both using the `Structured Output` built-in feature of the GPT-4o Mini model.

The prompts and schemas are provided in the appendix. We employed the OpenAI Python client library to run our prompts and used macro-averaged F1 (for emotion recognition) and mean absolute error (for intensity), aligned with the task's official evaluation metrics.

## 5 Results

Overall, our system provided near-baseline results on both tracks, reflecting the minimalist nature of our zero-shot approach. For Track B (figure 2), English performed best (0.5693 average Pearson $r$), followed by Ukrainian (0.3517), Hausa (0.3372), and Chinese (0.3068). We see that English data achieved higher scores for fear and joy, which aligns with the fact that GPT-4o Mini is primarily trained on extensive English corpora. Chinese exhibited the lowest consistency for surprise ($r = 0.133$), suggesting that the model struggled with subtle intensities in non-Latin scripts.

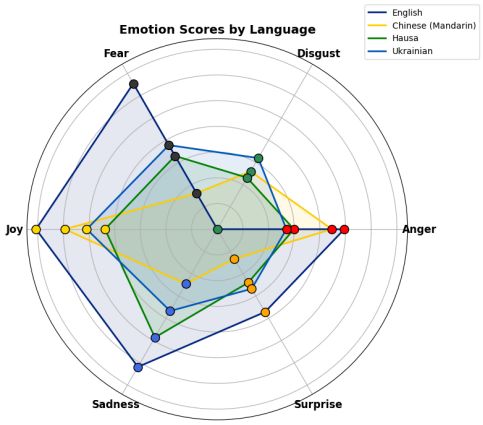


Figure 2: Track B Emotion Comparison Among Different Languages

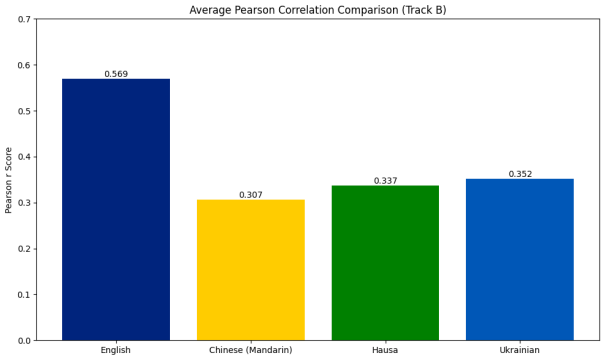


Figure 3: Track B Language Comparison

For Track C (figure 4), macro-F1 ranged from 0.5598 in English down to 0.1894 for isiZulu. Notably, languages with sparse resources—like isiZulu (0.1894) and Ukrainian (0.237)—lagged behind. Even within medium-resource languages (e.g., Indonesian at 0.5055 macro-F1), performance varied across emotions: fear and surprise were particularly challenging, possibly due to cross-lingual semantic gaps and fewer training signals in GPT-4o's domain knowledge. This discrepancy highlights how zero-shot performance can fluctuate significantly by language (figures 3, 5 ) and emotion category.

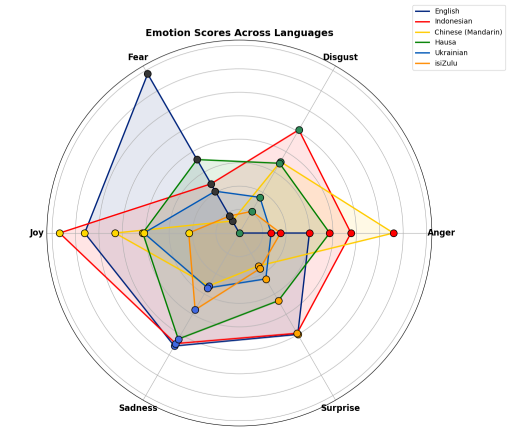


Figure 4: Track C Emotion Comparison Among Different Languages

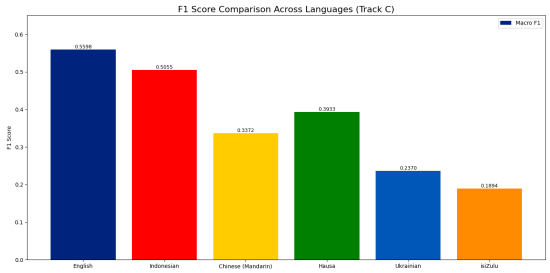


Figure 5: Track C Language Comparison

A closer look at error patterns revealed that many mistakes occurred when the Role-Play Rewriting step excessively modified or insufficiently modified the texts. When the rewrite did not accurately reflect the target emotion, the Contrastive Judging step struggled to pick a correct match. Conversely, when the rewrite was successful, the model often selected the correct emotion label. We suspect that isolating the contrastive judging mechanism (i.e., removing the rewriting stage) might help gauge how much rewriting errors degrade final predictions—an avenue for future systematic ablation.

In terms of competition ranking, our system was not among the top-scoring submissions. However,

| Languages | Emotions | | | | | | Average Pearson $r$ |
|---|---|---|---|---|---|---|---|
| | Anger | Disgust | Fear | Joy | Sadness | Surprise | |
| English | 0.4951 | - | 0.6545 | 0.7062 | 0.6186 | 0.3719 | 0.5693 |
| Chinese (Mandarin) | 0.4464 | 0.2598 | 0.1618 | 0.5944 | 0.2456 | 0.1330 | 0.3068 |
| Hausa | 0.2986 | 0.2317 | 0.3291 | 0.4375 | 0.4863 | 0.2403 | 0.3372 |
| Ukrainian | 0.2693 | 0.3195 | 0.3777 | 0.5087 | 0.3674 | 0.2678 | 0.3517 |

Table 1: Pearson correlation ($r$) scores for emotion intensity prediction across different languages in Track B.

| Languages | Emotions | | | | | | Macro F1 | Micro F1 |
|---|---|---|---|---|---|---|---|---|
| | Anger | Disgust | Fear | Joy | Sadness | Surprise | | |
| English | 0.2993 | - | 0.7834 | 0.6609 | 0.5561 | 0.4992 | 0.5598 | 0.5788 |
| Indonesian | 0.4769 | 0.5081 | 0.2422 | 0.7680 | 0.5442 | 0.4934 | 0.5055 | 0.5368 |
| Chinese (Mandarin) | 0.6587 | 0.3514 | 0.0595 | 0.5312 | 0.2604 | 0.1624 | 0.3372 | 0.3556 |
| Hausa | 0.3860 | 0.3436 | 0.3618 | 0.4128 | 0.5220 | 0.3333 | 0.3933 | 0.4061 |
| Ukrainian | 0.1358 | 0.1758 | 0.2060 | 0.4080 | 0.2705 | 0.2259 | 0.2370 | 0.2525 |
| isiZulu | 0.1761 | 0.1069 | 0.0841 | 0.2162 | 0.3780 | 0.1755 | 0.1894 | 0.2125 |

Table 2: Macro and Micro F1 scores for emotion classification in Track C.

it notably required no labeled data and relied on a lightweight LLM, making it cost-effective for low-resource use cases. The results indicate that while role-play rewriting and chain-of-thought reasoning can enhance emotion detection, further optimization—such as improved prompt engineering or partial fine-tuning—may significantly improve accuracy across languages.

## 6 Conclusion

We presented a novel zero-shot pipeline for multilingual emotion recognition using GPT-4o Mini, employing a role-play rewriting step followed by contrastive judging. Despite near-baseline official results, our method remains highly adaptable, requiring no labeled data and minimal computational cost.

Future work can explore replacing or refining the rewriting step, comparing this approach across different LLMs, and systematically evaluating each module (rewriting vs. judging) in isolation. We believe this line of research will open opportunities for more accessible, modular, and interpretable emotion detection solutions—especially valuable in low-resource and multilingual settings.

## References


Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a generative approach for emotion detection and reasoning.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Martin Juan, José Bucher, and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification.

Roman Kazakov, Kseniia Petukhova, and Ekaterina Kochmar. 2024. Petkaz at semeval-2024 task 3: Advancing emotion classification with an llm for emotion-cause pair extraction in conversations. pages 1127–1134.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35.

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, and 29 others. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint 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, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025b. SemEval-2025 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.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11:81.

Laura Ana Maria Oberländer and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card.

Flor Miriam Plaza-Del-Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, pages 5696–5710.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 1571–1582.

# Appendix

## Prompts and Schemas

### Rewrite Prompt

```
PROMPT_REWRITE = """
You are a helpful language model tasked with rewriting a given text to
convey specific emotional tones. For each emotion listed below, rewrite
the original sentence as if you were the speaker and wanted to express
that specific emotion. Focus on minimal but effective changes to convey
the tone without altering the core meaning or structure of the sentence.
Ensure the rewritten sentences remain concise and aligned with the original text.

Emotions to rewrite for:
- Joy
- Sadness
- Fear
- Anger
- Surprise
- Disgust

Please rewrite the text for each emotion. Do not extend or shorten the text
unnecessarily.
"""
```

The corresponding schema for the rewritten text is structured as follows:

```
class EmotionRewrittenText(BaseModel):
original_text: str
joy: str
sadness: str
fear: str
anger: str
surprise: str
disgust: str
```

### Scoring Prompt

```
PROMPT_SCORE = """
You are a helpful language model tasked with analyzing the emotional
intensity of an original sentence. Given a set of rewritten sentences
(one for each emotion), evaluate the original sentence to determine
how strongly it aligns with each emotion. The rewritten sentences
are clues to help guide your assessment, but your focus should
remain on the original sentence.

Instructions:
1. For each emotion (joy, sadness, fear, anger, surprise, disgust),
compare the original sentence to its corresponding rewritten version.
2. Assess how closely the original sentence aligns with the tone of
each rewritten sentence, considering subtle cues in language,
context, and implied sentiment.
3. Provide a brief reasoning for each emotion explaining the alignment.
4. Assign an intensity score for each emotion:
- 0: No emotion present
- 1: Low degree of emotion
- 2: Moderate degree of emotion
- 3: High degree of emotion

Provide your analysis and intensity scores for each emotion.
"""
```

The schema for the emotion analysis output is structured as follows:

```
class EmotionAnalysisNonNestedOutput(BaseModel):
    joy_reasoning: str
    joy_intensity: int
    sadness_reasoning: str
    sadness_intensity: int
    fear_reasoning: str
    fear_intensity: int
    anger_reasoning: str
    anger_intensity: int
    surprise_reasoning: str
    surprise_intensity: int
    disgust_reasoning: str
    disgust_intensity: int
```

These two prompts and their corresponding schemas illustrate our modular approach for role-play rewriting (`PROMPT_REWRITE`) and contrastive judging/scoring (`PROMPT_SCORE`). By calling the language model twice—once to obtain the rewritten text for each emotion, and once to assign intensities based on those rewrites—we maintain a clean separation between generation and evaluation steps, facilitating easy adjustments or substitutions of different large language models.