# Emotion Entanglement and Bayesian Inference for Multi-Dimensional Emotion Understanding

**Anonymous ACL submission**

## Abstract

Understanding emotions in natural language is inherently a multi-dimensional reasoning problem, where multiple affective signals interact through context, interpersonal relations, and situational cues. However, most existing emotion understanding benchmarks rely on short texts and predefined emotion labels, reducing this process to independent label prediction and ignoring the structured dependencies among emotions. To address this limitation, we introduce **Emo**tional **Scen**arios (**EmoScene**), a theory-grounded benchmark of 4,731 context-rich scenarios annotated with an *8-dimensional emotion vector derived from Plutchik's basic emotions*. We evaluate six instruction-tuned large language models in a zero-shot setting and observe modest performance, with the best model achieving a Macro F1 of **0.501**, highlighting the difficulty of context-aware multi-label emotion prediction. Motivated by the observation that emotions rarely occur independently, we further propose an *entanglement-aware Bayesian inference* framework that incorporates emotion co-occurrence statistics to perform joint posterior inference over the emotion vector. This lightweight post-processing improves structural consistency of predictions and yields notable gains for weaker models (e.g., **+0.051** Macro F1 for Qwen2.5-7B). EmoScene therefore provides a challenging benchmark for studying *multi-dimensional emotion understanding* and the limitations of current language models.

## 1 Introduction

Emotional intelligence (EI) refers to the ability to recognize, understand, and regulate emotions in oneself and others (Mayer et al., 2004). Emotions play a fundamental role in human cognition, influencing decision making, social interaction, and behavioral responses to environmental stimuli (Hieida et al., 2018). Because emotional signals are often implicit and context dependent, understanding them requires interpreting subtle linguistic cues, situational context, and interpersonal relationships. In natural language processing, this ability is typically evaluated through emotion classification tasks that aim to measure a system's capacity for *Emotion Understanding* (EU), namely the interpretation of emotions expressed within textual context (Sabour et al., 2024; Buechel and Hahn, 2017). Despite significant progress in affective computing, modelling emotions remains challenging because emotional expressions are subjective, context sensitive, and often intertwined with social dynamics (Izard, 2002; Perloff, 1997).

Recent advances in Large Language Models (LLMs) have led to remarkable improvements across a wide range of NLP tasks, including text generation, translation, and question answering (Hagos et al., 2024; Ren, 2024). As these systems become increasingly capable of producing fluent and contextually appropriate responses, a natural question arises: **Do LLMs genuinely understand emotions, or do they merely reproduce linguistic patterns associated with emotional language?** Existing emotion understanding benchmarks provide only a partial answer. Datasets such as SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018), GoEmotions (Demszky et al., 2020), ISEAR (Troiano et al., 2019), and EmotionLines (Chen et al., 2018) have enabled large-scale evaluation of emotion classification systems. However, many of these datasets suffer from key limitations. First, inputs are often short sentences or social media posts that lack sufficient context to disambiguate subtle emotional cues. Second, they rely on predefined emotion labels that may not align with established psychological theories of emotion. Third, most evaluation protocols treat emotions as independent categories, ignoring the fact that emotional states frequently co-occur and influence one another.

In this work we introduce **Emo**tional **Scen**arios

**Scenario:**

Sue sat at her desk, laptop open to Amazon's website. She clicked through several pages, determined to find a specific album she'd heard about from a friend. Her eyes flicked back to the screen repeatedly, fingers tapping the table lightly. After placing the order, she checked the status frequently, **eager to have the new music** in her library. As soon as it arrived, she unwrapped the package and played the album immediately, **letting the tunes fill her room**.

**Subject:** Sue

| Emotion | Anger | Anticipation | disgust | fear | Joy | sadness | surprise | trust |
|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |

Figure 1: An example scenario from the EmoScene Dataset

(**EmoScene**), a benchmark designed to evaluate emotion understanding in context-rich situations. Each example consists of a subject and a detailed scenario describing environmental cues, bodily reactions, and interpersonal interactions. Instead of using a predefined list of complex emotions, we ground the annotation scheme in Plutchik's theory of emotions[1] and represent each scenario using an **8-dimensional binary emotion vector corresponding to Plutchik's basic emotions** (ref. example from Figure 1. This formulation enables the representation of complex emotional states as combinations of fundamental emotional dimensions. To further capture the structure of emotional expression, we observe that emotions rarely occur independently. Motivated by this observation, we model emotion prediction as a structured inference problem and introduce a *Bayesian post-processing* framework that combines LLM predictions with a prior capturing statistical dependencies between emotions. This formulation allows us to perform joint inference over the emotion vector and correct the independence assumption commonly made by vanilla multi-label predictions.

We evaluate six instruction-tuned open-source LLMs in a zero-shot setting on the EmoScene benchmark to assess their ability to interpret complex emotional scenarios. Despite strong general language capabilities, models achieve only modest performance on multi-label emotion prediction, with the best model reaching a Macro F1 of **0.501**. To account for dependencies among emotions, we introduce an entanglement-aware Bayesian inference framework that performs joint posterior inference over the emotion vector. This lightweight post-processing improves prediction consistency and yields substantial gains for weaker models, including a **+0.051 increase in Macro F1** and a **6.27 reduction in Hamming Loss**. Our analysis further reveals systematic failure modes of current models, particularly when emotional signals arise from relational interactions and contextual cues rather than explicit affective language.

**Our contributions are summarized as follows:**

1. **EmoScene Benchmark.** A context-rich dataset of **4,731 scenarios** for evaluating emotion understanding in natural language.

2. **Theory-grounded emotion representation.** Emotion prediction is formulated as an **8-dimensional multi-label task** based on Plutchik's basic emotions.

3. **Emotion entanglement modeling.** A **Bayesian inference framework** that incorporates statistical dependencies between emotions for joint prediction.

4. **LLM evaluation.** A systematic zero-shot study of **six instruction-tuned LLMs** on multi-dimensional emotion understanding.

The code and data has been released for academic use at: https://github.com

## 2 Related Work

**Emotion theories and psychological grounding.** Emotions have long been studied in psychology as adaptive mechanisms that influence cognition,

[1] https://en.wikipedia.org/wiki/Emotion_classification

behavior, and social interaction (Damasio, 1998; Gross and Barrett, 2011). Emotional intelligence refers to the capacity to recognize, interpret, and manage emotional information in oneself and others (Mayer et al., 2004). A large body of work has explored how emotional experiences are structured and expressed in human cognition. For example, emotional granularity describes the ability to distinguish subtle affective states and has been linked to improved emotional regulation and social functioning (Hoemann et al., 2021). Psychological theories such as Plutchik's psychoevolutionary theory further suggest that complex emotional states arise from combinations of a small set of fundamental emotions (Plutchik, 1982, 1980). These theoretical perspectives highlight that emotions are inherently structured and often co-occur rather than appearing in isolation.

**Emotion understanding in NLP.** Emotion analysis has been widely studied within natural language processing as an extension of sentiment analysis. Early work focused on detecting affective states in text through lexicons and classification models (Mohammad and Turney, 2013). Subsequent research expanded this task to emotion recognition in conversations and context-dependent settings, emphasizing the importance of discourse structure and speaker interactions (Fu et al., 2023). Recent work has also explored multimodal approaches that combine textual, acoustic, and visual signals to improve emotion recognition and generation in dialogue systems (Liu et al., 2024; Chandraumakantham et al., 2024). These studies demonstrate that emotional signals are often distributed across multiple modalities and contextual cues.

**Emotion datasets and benchmarks.** The development of emotion understanding systems has been driven by the availability of annotated datasets. Early benchmarks such as the SemEval affective text tasks focused on identifying emotions in short texts such as tweets (Strapparava and Mihalcea, 2007; Mohammad et al., 2018). More recent datasets, including GoEmotions (Demszky et al., 2020), EmotionLines (Chen et al., 2018), and crowdsourced corpora derived from ISEAR (Troiano et al., 2019; Scherer and Wallbott, 1990), have introduced larger taxonomies of emotions and conversational contexts. Despite these advances, most datasets rely on short or decontextualized inputs and predefined emotion categories, which may limit their ability to capture complex emotional states. In addition, annotation practices vary widely across datasets, ranging from crowdsourcing to expert labeling, which can affect label consistency and reliability.

**Emotion understanding in large language models.** With the rise of large language models (LLMs), recent research has begun to examine their capacity for emotional intelligence. Studies have evaluated LLMs on emotion recognition, empathy, and affect-aware reasoning tasks (Sabour et al., 2024; Sorin et al., 2024). Other work has investigated how LLM architectures represent emotional information internally or handle ambiguous emotional cues (Hong et al., 2025; Reichman et al., 2026). These findings suggest that while LLMs demonstrate promising capabilities in affective language tasks, their performance remains sensitive to context and evaluation design.

**Limitations of existing benchmarks and our approach.** Despite substantial progress, existing emotion benchmarks exhibit several limitations. Many datasets rely on short or decontextualized texts that lack sufficient contextual information to disambiguate emotional states. In addition, most evaluation protocols treat emotion prediction as a collection of independent classification decisions, ignoring the structured dependencies among emotions suggested by psychological theories. As a result, it remains unclear whether models genuinely reason about emotional states or merely detect surface-level affective cues. To address these limitations, we introduce **EmoScene**, a benchmark of context-rich emotional scenarios grounded in Plutchik's theory of emotions. Each scenario is annotated with an 8-dimensional emotion vector representing combinations of fundamental emotions. We further model dependencies among emotions using a Bayesian inference framework that incorporates statistical co-occurrence patterns observed in the dataset, enabling emotion understanding to be studied as a structured multi-label inference problem.

## 3 EmoScene Dataset

Understanding emotions in natural language requires reasoning over contextual cues, interpersonal interactions, and physical states described within a scenario. However, many existing emotion benchmarks rely on short, decontextualized

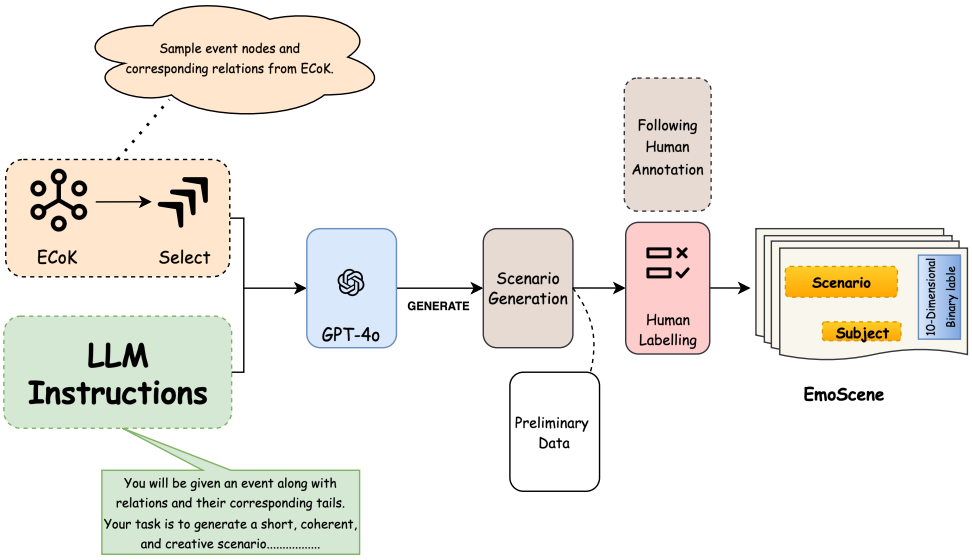



Figure 2: Overview of our steps to construct EmoScene dataset

texts, which makes it difficult to disambiguate subtle emotional signals and often reduces emotion prediction to a fixed set of predefined labels. To address this limitation, we introduce **Emo**tional **Scen**arios (**EmoScene**), a benchmark consisting of **4,731 context-rich scenarios** designed to evaluate multi-dimensional emotion understanding. Each example contains a **subject description** and a scenario describing the surrounding situation, and is annotated with an **8-dimensional binary emotion vector** grounded in Plutchik's psychoevolutionary theory of emotions (Plutchik, 1982, 1980). This representation corresponds to Plutchik's eight basic emotions—*joy, trust, fear, surprise, sadness, disgust, anger, and anticipation*—and allows multiple emotions to co-occur simultaneously, enabling the dataset to capture complex emotional states while supporting multi-label emotion prediction.

The dataset was constructed in two stages: **scenario generation** and **human annotation**. An overview of the pipeline is shown in Figure 2. We leverage the **Emotional Commonsense Knowledge Graph (ECoK)** (Wang et al., 2024) as the starting point for scenario creation. ECoK is a knowledge graph designed to represent emotional knowledge using `<head, relation, tail>` triples. The head entities describe emotional events, while the relations capture different aspects of emotional experience. The graph defines seven relation types: *group, causes, bodily symptoms, feelings, expression, action tendencies, and advice*.

For each head node in ECoK, we generate one scenario using GPT-4o. The model is prompted to construct a short narrative describing a subject experiencing an emotionally relevant situation, incorporating contextual cues such as environmental conditions, interpersonal interactions, and bodily responses. After generation, the scenarios were manually reviewed to remove duplicates and inconsistent outputs. This filtering process resulted in a final dataset of **4,731 scenarios**. The generation prompt used for this process is provided in Appendix E. Expert annotators have gone through every scenario and labeled the instance. Overall Fleiss kappa ($\kappa$) is 0.43 which falls in the moderate agreement range. More details on the Human Annotation is in the Appendix: B. Figure 3 shows the label count for each emotion dimension in the EmoScene.

## 4 Methodology

Traditional multi-label classification frameworks typically model each emotion as an independent Bernoulli variable, implicitly assuming that emotional states occur in isolation. However, this assumption is fundamentally misaligned with the nature of human affect, where emotions are rarely independent and often co-occur in structured, interdependent ways. For instance, emotions such as joy and surprise may reinforce each other, while others, like anger and sadness, may exhibit complex conditional relationships depending on context.

To address this limitation, we introduce the concept of Emotion Entanglement, which captures the inherent dependencies and interactions among emotions as a coupled system rather than a set of dis-

joint labels. Building on this perspective, we propose an Entanglement-Aware Bayesian Inference framework that explicitly models the joint distribution over emotions. By incorporating these dependencies into the inference process, our approach enables more coherent, context-sensitive predictions that better reflect the underlying structure of human emotional expression.

### 4.1 Modeling Emotion Entanglement

We define **Emotion Entanglement** as the phenomenon where the presence of one emotion in a text significantly alters the likelihood of another co-occurring. Formally, we hypothesize that for an input sentence $x$ and an emotion vector $E = (E_1, \ldots, E_{10})$, the joint distribution does not factorize:

$$P(E|x) \neq \prod_{i=1}^{10} P(E_i|x) \tag{1}$$

To quantify this, we analyze the structural non-independence of emotions using information-theoretic measures. Specifically, we compute the Mutual Information $I(E_i; E_j)$ between emotion pairs to establish that emotional expression is structurally dependent. For instance, in the EmoScene dataset, the probability of *Anger* is significantly higher given the presence of *Sadness* than its marginal probability, i.e., $P(\text{Anger} = 1|\text{Sadness} = 1) \gg P(\text{Anger} = 1)$.

### 4.2 Entanglement-Aware Bayesian Inference

To correct the independence assumption made by vanilla LLMs, we propose a post-processing step based on Bayesian inference. We treat the LLM output as a noisy likelihood and combine it with a data-derived prior that captures label entanglement.

Let $x$ be the input text and $E = (E_1, \ldots, E_{10})$ be the binary emotion vector where $E_i \in \{0, 1\}$. Our goal is to find the Maximum A Posteriori (MAP) estimate $\hat{E}$:

$$\hat{E} = \arg\max_E P(E|x) \propto \arg\max_E P(x|E) \cdot P(E) \tag{2}$$

#### 4.2.1 Estimating the Entanglement Prior

We model the label-only prior $P(E)$ using an **Ising Model** (a type of Markov Random Field) to find the least-committed distribution that matches observed marginals and pairwise correlations in the training data. The prior is defined as:

$$P(E) = \frac{1}{Z} \exp\left(\sum_i \theta_i E_i + \sum_{i<j} \theta_{ij} E_i E_j\right) \tag{3}$$

where $\theta_i$ represents the bias for individual emotions and $\theta_{ij}$ represents the coupling strength between emotion pairs. These parameters are estimated from the training set via maximum-entropy solutions:

$$\begin{aligned} \theta_i &= \log \frac{\hat{P}(E_i = 1)}{1 - \hat{P}(E_i = 1)}, \\ \theta_{ij} &= \log \frac{\hat{P}(E_i = 1, E_j = 1)}{\hat{P}(E_i = 1)\hat{P}(E_j = 1)} \end{aligned} \tag{4}$$

#### 4.2.2 Likelihood and MAP Inference

The likelihood $P(x|E)$ is derived from the LLM's log-probabilities for each emotion category. Assuming a Bernoulli likelihood for the text given an emotion configuration:

$$\log P(x|E) = \sum_i \left[E_i \log p_i^{(1)} + (1 - E_i) \log p_i^{(0)}\right] \tag{5}$$

Substituting the prior and likelihood back into the MAP objective, our final a-posteriori log-objective becomes:

$$\begin{aligned} \log P(E|x) = &\sum_i \left[E_i \log \frac{p_i^{(1)}}{p_i^{(0)}}\right] \\ &+ \alpha \left[\sum_i \theta_i E_i + \sum_{i<j} \theta_{ij} E_i E_j\right] \end{aligned} \tag{6}$$

Since we utilize a fixed space of 10 labels, the exact argmax is found by evaluating all $2^{10}$ combinations, ensuring a parameter-efficient post-processing step that explicitly models emotion co-occurrence fidelity.

## 5 Experimental Setup

In this section, we will list out the models and hyperparameters that we have used to run the experiments.

We have evaluated the performance of 6 open-source LLMs in zero-shot setting and letting the model explain it's answer along with confidence. Models that we have evaluted are: `Mistral-Nemo-Instruct-2407`,

[1] https://huggingface.co/models

| Model | Lexical Accuracy | | Vector Accuracy | | Hamming Loss | | F1-Macro | |
|---|---|---|---|---|---|---|---|---|
| | Baseline | BI (Δ) | Baseline | BI (Δ) | Baseline | BI (Δ) | Baseline | BI (Δ) |
| Llama-3.1-8B | 83.87 | 83.88 (+0.008) | 24.54 | 24.50 (-0.04) | 16.13 | 16.12 (+0.008) | 0.472 | 0.472 (0.000) |
| Llama-3.2-3B | 79.76 | 79.76 (0.000) | 17.40 | 17.40 (0.00) | 20.24 | 20.24 (0.000) | 0.451 | 0.451 (0.000) |
| Mistral-Nemo | 83.48 | 83.50 (+0.021) | 22.24 | 22.26 (+0.02) | 16.52 | 16.50 (+0.021) | 0.431 | 0.432 (+0.001) |
| Qwen2.5-7B | 75.95 | 82.22 (+6.275) | 14.06 | 21.48 (+7.42) | 24.05 | 17.78 (+6.275) | 0.428 | 0.479 (+0.051) |
| Gemma-2-2b | 70.93 | 78.79 (+7.858) | 10.27 | 15.96 (+5.69) | 29.07 | 21.21 (+7.858) | 0.437 | 0.474 (+0.037) |
| Gemma-2-9b | 83.37 | 83.48 (+0.111) | 24.88 | 25.07 (+0.19) | 16.63 | 16.52 (+0.111) | 0.500 | 0.501 (+0.001) |

Table 1: Model Performance Comparison on EmoScene. BI denotes our Bayesian Inference method, and Baseline denotes the zero-shot LLM results. Values in parentheses indicate the change (Δ) from the baseline. All the models in this table are instruction-tuned (instruct) models downloaded from HuggingFace.

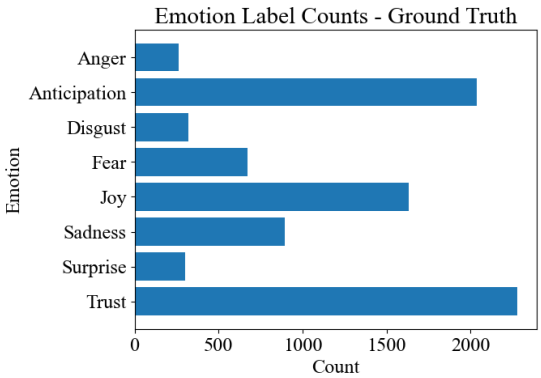



Figure 3: Emotion Label counts from the EmoScene Dataset.

`Llama-3.2-3B-Instruct`, `Gemma-2-2b-it`, `Qwen2.5-7B-Instruct`, `Gemma-2-9b-it`, `Llama-3.1-8B-Instruct`.

Across the MAP experiments, we have evaluated multiple alphas: [0.0, 0.1, 0.25, 0.5, 0.75, 1.0, 2.0, 5.0]. To get the final output we have extracted logit scores corresponding to Yes & No (just before LM head). All the models are downloaded from HuggingFace[1] and downloaded locally. We have use 2 Nvidia-A100 cards each 80GB to run the experiments.

## 6 Results & Analysis

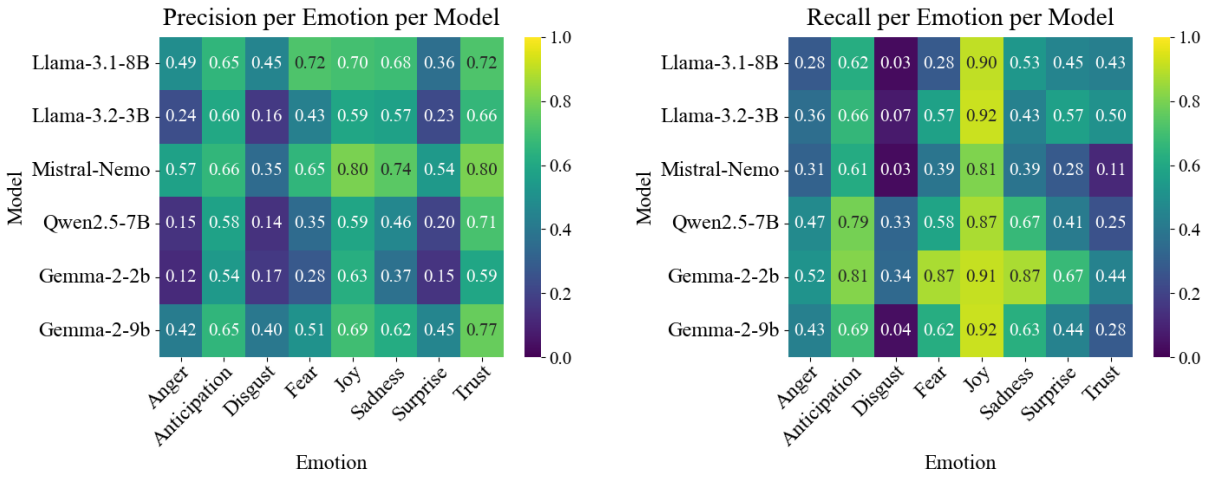
In this section, we discuss the evaluation of multiple instruction-tuned open-source LLMs on the EmoScene benchmark for multi-label emotion prediction using Scenario and Subject. We compare zero-shot performance with our proposed Bayesian Inference (BI) approach and report the results in Table 1. Performance is assessed using Lexical Accuracy, Vector Accuracy, Hamming Loss, and Macro F1-score, capturing label-wise correctness, exact set match, label errors, and balanced performance across emotions. Prompt that is used for evaluation is included in the Appendix 8

Our Bayesian Inference (BI) method consistently outperforms the zero-shot baseline across all models and metrics on the EmoScene dataset. Gemma-2-9b-it achieves the highest F1-Macro of 0.501 with lexical accuracy 83.48%, while Qwen2.5-7B-Instruct shows the largest gains (F1-Macro +5%, Hamming Loss, -6.275); improvements are most evident in smaller models such as Gemma-2-2b-it (lexical accuracy +7.858, Hamming –7.858), confirming BI boosts reliability and reduces per-emotion errors without affecting strong performers like Llama-3.1-8B-Instruct.

Vector accuracy remains modest (15–25%) despite high lexical scores because BI performs joint MAP inference over the full emotion vector $\mathbf{E}$, enforcing an entanglement prior via an MRF/Ising graph that models pairwise correlations (e.g., sadness co-occurs more with anger than joy). This corrects the independence assumption of zero-shot LLMs by re-weighting predictions to better match co-occurrence patterns and reduce Hamming Loss. As a result, the final vector may deviate from the original marginals, lowering exact-match accuracy. However, this trade-off consistently improves Hamming Loss and Macro F1, demonstrating BI's effectiveness for robust, scene-aware emotion classification.

### 6.1 Predicitive Analysis

Per-Emotion precision & Recall heatmaps can be found in the Figure 4. Joy and Anticipation emerge as the strongest categories overall, where Joy achieves high precision and recall across all models consistently, while Anticipation shows moderate precision and recall, reflecting their frequent occurrence in the dataset and the clear lexical cues that LLMs readily capture. In contrast, Disgust displays critically low recall (near-zero in Llama and Mistral variants), and Surprise similarly suffers low recall in larger models, indicating models might be




(a) Precision per Emotion per Model

(b) Recall per Emotion per Model

Figure 4: Heatmap comparison of Precision and Recall for Plutchik's 8 basic emotions across all evaluated models. Darker regions indicate lower performance, while lighter/greener regions indicate higher alignment.

adopting conservative strategies, predicting them only when evidence is overwhelming.

In this multi-label setting, precision reflects the reliability of positive predictions (avoiding false cues like incorrectly detecting Anger), while recall ensures coverage of the full emotional spectrum. Smaller model like Gemma-2-2b-it shows high recall for Anticipation, Fear, Joy, Sadness, but low precision on labels like Anger and Surprise, indicating over-prediction on rare classes (e.g., Disgust, Surprise). Larger models such as Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct are more balanced but remain conservative on these emotions. This motivates our Bayesian post-processing: the MRF/Ising-based entanglement prior, which enforces realistic co-occurrence patterns, improving Macro F1 and reducing Hamming Loss even for rare emotions without retraining. Below, we discuss a general case where our posterior formulation corrected the label from the baseline.

**Scenario [Anticipation, Sadness]**

Sonia scrolled through her Instagram feed, her eyes narrowing as she paused on Frank's latest post about his trip to Bali. Her stomach felt unsettled, and she noticed her muscles tensing as she read about his adventures. She couldn't help but compare her own routine life to his glamorous lifestyle. Trying to shake off the feeling, she decided to focus on her own upcoming weekend plans and the things she has achieved recently.

In the above scenario, the baseline incorrectly predicts *anger* along with *anticipation* and *sadness*. In general, sadness is a low-arousal negative emotion, whereas anger is a high-arousal negative emotion. The description of her stomach being "unsettled" and muscles "tensing" injects a high level of physical arousal into the text. The baseline model struggled to map this high-arousal discomfort to sadness, defaulting instead to anger, which is the most common high-arousal negative emotion in text classification. Our formulation removes this spurious label of anger driven by the learned prior, where $\theta_{ij}$ favours the co-occurrence of sadness and anticipation but penalises unlikely associations with anger.

## 6.2 Error Analysis

**Scenario 1**

Kim sat on the edge of her bed, scrolling through horror stories about dentist visits on her phone. Her hands were damp, and she could feel her heart racing as she thought about the appointment she had been putting off for weeks. Her sister, noticing her tense expression and the way she kept fidgeting with her hair, suggested they go together so Kim wouldn't have to sit in the waiting room alone. **Kim nodded slowly, grateful for the offer**, and promised herself to practice some deep breathing before they left.

To better understand the limitations of current LLMs in emotional reasoning, we systematically examined the scenarios in which LLMs have made

mistakes. We have examined the cases where the Bayesian post-processing is helping by correcting the baseline prediction and the cases where even after post-processing remained wrong. Specifically, we focused on Gemma-2-9b-it model's outputs, which achieved the best F1-Macro score among all the evaluated models.

In Scenario 1, reflects the large number of false negatives for Trust (1629) by the best-performing model. The emotional signal arises from the relational interaction, particularly the phrase "Kim nodded slowly, grateful for the offer," which indicates reliance on another person's support. However, the model likely focuses on earlier anxiety cues such as "hands were damp" and "heart racing," which describe stress rather than interpersonal trust. As a result, the relational aspect of the interaction is overlooked.

**Scenario 2**

In a bustling dance studio filled with music, William spontaneously shifted Raymond's hips during their routine, playfully adjusting his partner's stance. Raymond felt his **heart rate quicken** and **muscles tense**, unsure how to respond to the unexpected move. **Uncertain, his eyebrows raised slightly**, but he managed to maintain his balance with William's guidance. As the music continued, they both settled into the rhythm, a shared glance signaling an unspoken understanding.

In Scenario 2, the model likely relied on physiological cues such as "heart rate quicken" and "muscles tense," interpreting them as signs of anticipation. However, the phrase "Uncertain, his eyebrows raised slightly" indicates confusion rather than expectation of a future event.

## 7 Conclusion and Future Work

We introduced **EmoScene**, a context-rich benchmark for evaluating multi-dimensional emotion understanding in language models. Grounded in Plutchik's theory, EmoScene represents emotional states using an 8-dimensional emotion vector and contains 4,731 scenarios that require reasoning over contextual and relational cues. Our evaluation of six instruction-tuned LLMs in a zero-shot setting shows that even strong models achieve modest performance, highlighting the difficulty of context-aware multi-label emotion prediction. To address the independence assumption commonly made in emotion classification, we proposed an entanglement-aware Bayesian inference framework that incorporates emotion co-occurrence statistics to perform joint posterior inference, improving prediction consistency and yielding notable gains for weaker models. These results suggest that modelling dependencies among emotions is important for robust emotion understanding. In future work, we plan to extend EmoScene with additional modalities and languages and investigate architectures and reasoning strategies that explicitly model emotional structure in context. Further, we would also like to explore the utility of preference tuning strategies in improving the emotional awareness of large language models.

## Limitations

While EmoScene provides a context-rich benchmark for multi-dimensional emotion understanding, several limitations remain. First, the scenarios are generated using a language model and may not fully capture the diversity and complexity of naturally occurring emotional narratives. Second, the dataset focuses solely on textual descriptions and does not incorporate multimodal signals such as facial expressions, speech, or physiological cues that often accompany emotional expression. Finally, although our Bayesian inference framework models pairwise dependencies among emotions, it assumes relatively simple statistical relationships and may not capture higher-order emotional interactions present in real-world settings.

## Ethics Statement

All snippets and examples used in this work were collected or generated following ethical practices and, where applicable, under permissive licenses. The dataset does not contain personally identifiable information or sensitive data. All human annotators involved in constructing the EmoScene dataset participated voluntarily and provided informed consent prior to annotation. Annotators were compensated on an hourly basis in accordance with the labor regulations of their respective country, and the task was designed to minimize cognitive burden during labeling. The scenarios were generated using a language model and subsequently reviewed and annotated by human experts, with duplicates and inconsistent outputs removed during dataset curation. Additionally, generative AI tools were

used only to improve the clarity and grammar of the manuscript during writing; they were not used for literature search, research design, or idea generation. All scientific contributions, analyses, and conclusions presented in this work are the result of the authors' own research.

## References


Sven Buechel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Omkumar Chandraumakantham, N. Gowtham, Mohammed Zakariah, and Abdulaziz Almazyad. 2024. Multimodal emotion recognition using feature fusion: An llm-based approach. *IEEE Access*, 12:108052–108071.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. *Preprint*, arXiv:1802.08379.

Antonio R. Damasio. 1998. Emotion in the perspective of an integrated nervous system1Published on the World Wide Web on 27 January 1998.1. *Brain Research Reviews*, 26(2):83–86.

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*, pages 4040–4054, Online. Association for Computational Linguistics.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics (Basel)*, 12(22):4714.

James J Gross and Lisa Feldman Barrett. 2011. Emotion generation and emotion regulation: One or two depends on your point of view. *Emot. Rev.*, 3(1):8–16.

Desta Haileselassie Hagos, Rick Battle, and Danda B. Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *Preprint*, arXiv:2407.14962.

Chie Hieida, Takato Horii, and Takayuki Nagai. 2018. Deep emotion: A computational model of emotion using deep neural networks. *ArXiv*, abs/1808.08447.

Katie Hoemann, Lisa Feldman Barrett, and Karen S Quigley. 2021. Emotional granularity increases with intensive ambulatory assessment: Methodological and individual factors influence how much. *Front. Psychol.*, 12:704125.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Carroll E Izard. 2002. Translating emotion theory and research into preventive interventions. *Psychol. Bull.*, 128(5):796–824.

Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024. Speak from heart: an emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 international conference on multimedia retrieval*, pages 533–542.

John D Mayer, Peter Salovey, and David R Caruso. 2004. TARGET ARTICLES: "emotional intelligence: Theory, findings, and implications". *Psychol. Inq.*, 15(3):197–215.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Robert Perloff. 1997. Daniel goleman's emotional intelligence: Why it can matter more than IQ. *Psychol. Manag. J.*, 1(1):21–22.

Robert Plutchik. 1980. A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION. In *Theories of Emotion*, pages 3–33. Elsevier.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Soc. Sci. Inf. (Paris)*, 21(4-5):529–553.

Benjamin Reichman, Adar Avasian, Samuel Webster, and Larry Heck. 2026. Emotion is not just a label: Latent emotional factors in llm processing. *Preprint*, arXiv:2603.09205.

Mengchao Ren. 2024. Advancements and applications of large language models in natural language processing: A comprehensive review. *Applied and Computational Engineering*, 97(1):55–63.

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 (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Klaus R. Scherer and Harald G. Wallbott. 1990. International survey on emotion antecedents and reactions (isear). *Journal of Cross-Cultural Psychology*.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: systematic review. *Journal of medical Internet research*, 26:e52597.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Zhunheng Wang, Xiaoyi Liu, Mengting Hu, Rui Ying, Ming Jiang, Jianfeng Wu, Yalan Xie, Hang Gao, and Renhong Cheng. 2024. ECoK: Emotional commonsense knowledge graph for mining emotional gold. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8055–8074, Bangkok, Thailand. Association for Computational Linguistics.

## A Dataset Statistics

Table 2 summarizes the statistics of the EmoScene dataset. The scenarios contain rich contextual descriptions with an average length of **91 words**. Approximately **44.9%** of the scenarios contain more than one emotion, highlighting the prevalence of multi-dimensional emotional states in the dataset. The average label cardinality is **1.54**, indicating that most scenarios express one or two emotions simultaneously. Figure 3 shows the distribution of emotion labels across the dataset.

## B Human Annotation

All scenarios were annotated by **three expert annotators**. The annotators are English teachers with prior annotation experience. Each annotator independently labeled every scenario using the 8-dimensional emotion vector. Multiple emotions could be selected if the scenario expressed mixed emotional signals.

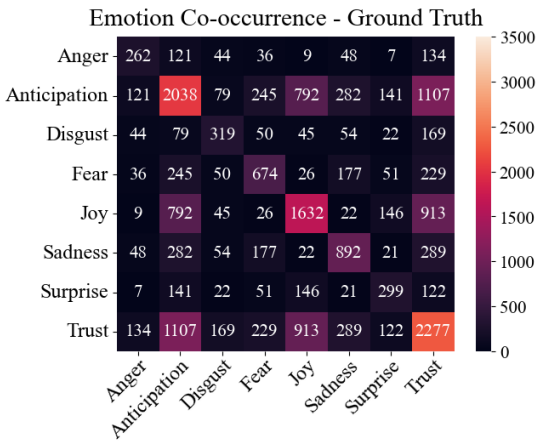
The annotation was conducted using a spreadsheet-based interface. Annotators reported that labeling each scenario typically required approximately **2–5 minutes**, depending on the complexity of the emotional cues present in the scenario. Annotation guidelines are clearly explained and shared to all the annotators. Guidelines are also uploaded in the codebase that is shared.

| Statistic | Value |
|---|---|
| Total number of scenarios | 4,731 |
| Avg. scenario length (words) | 91 |
| Max scenario length (words) | 131 |
| Min scenario length (words) | 59 |
| Single-emotion scenarios | 1,165 |
| Multi-label scenarios | 44.90% |
| Average label cardinality | 1.54 |

Table 2: Statistics of the EmoScene dataset.

The final label for each scenario was determined using **majority voting**. To assess annotation consistency, we compute pairwise agreement between annotators using the Kappa. The agreement is **Moderate (pairwise F1 ≈ 0.43)**, indicating moderate consistency across annotators.

## C EmoScene Emotion Entanglement

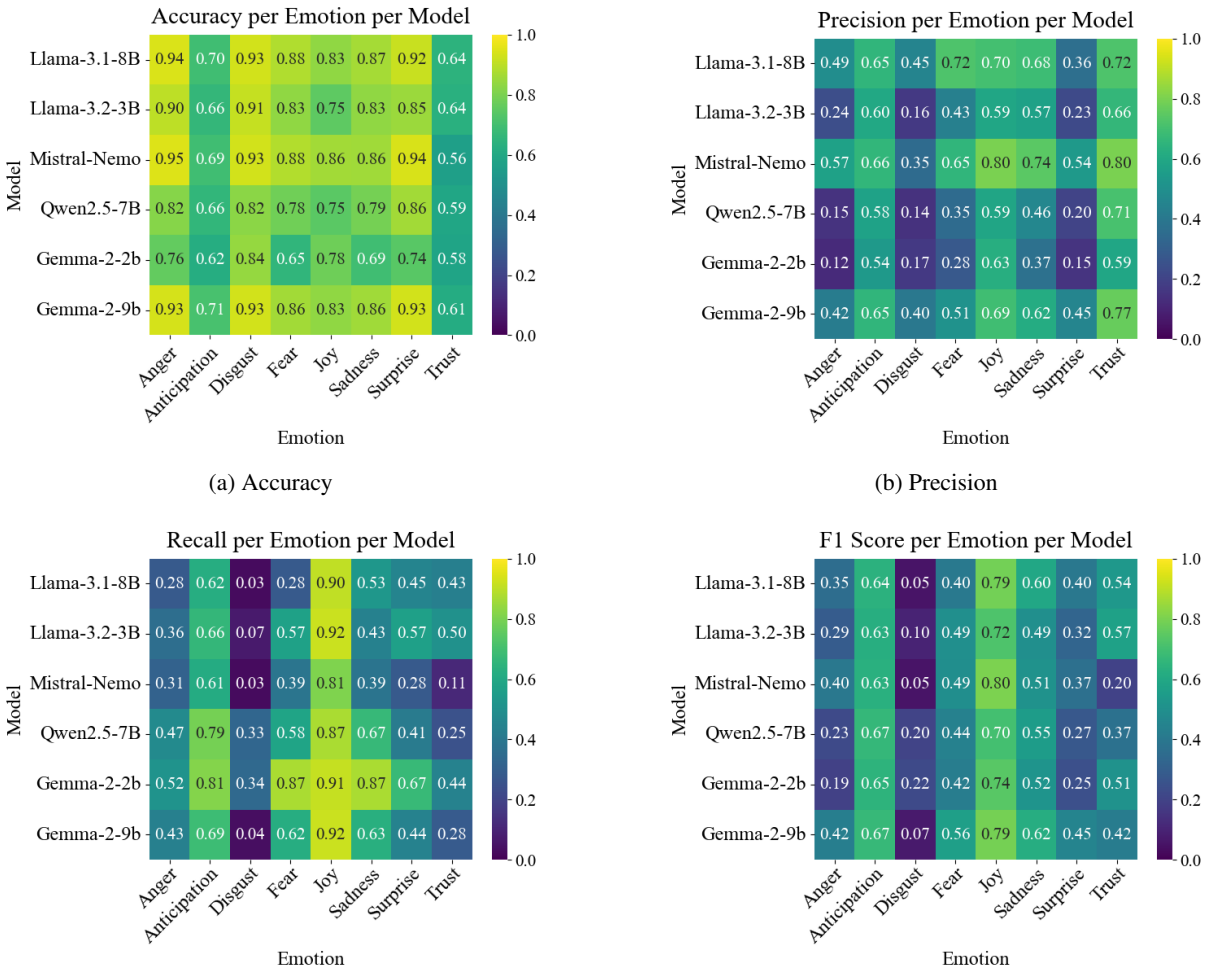



Above, we see the co-occurrence of emotion labels in EmoScene dataset. The frequency of two given emotions being assigned to the same Scenario. It appears that anticipation, joy, and trust co-occur quite frequently, meaning our models may confuse between them. Their co-occurrences might also account for their high individual counts in our dataset overall.

## D Predicted Labels Analysis

The following section contains the analysis of the predictions made by the models.

Looking at the heatmaps in Figure 5, we see some "vertical banding" across precision, recall, and F1 score, meaning that these metrics are relatively consistent per emotion rather than model. This implies that specific emotions are challenging




(a) Accuracy (b) Precision

(c) Recall (d) F1-score

Figure 5: Performance comparison across models for different emotions in terms of accuracy, precision, recall, and F1-score.

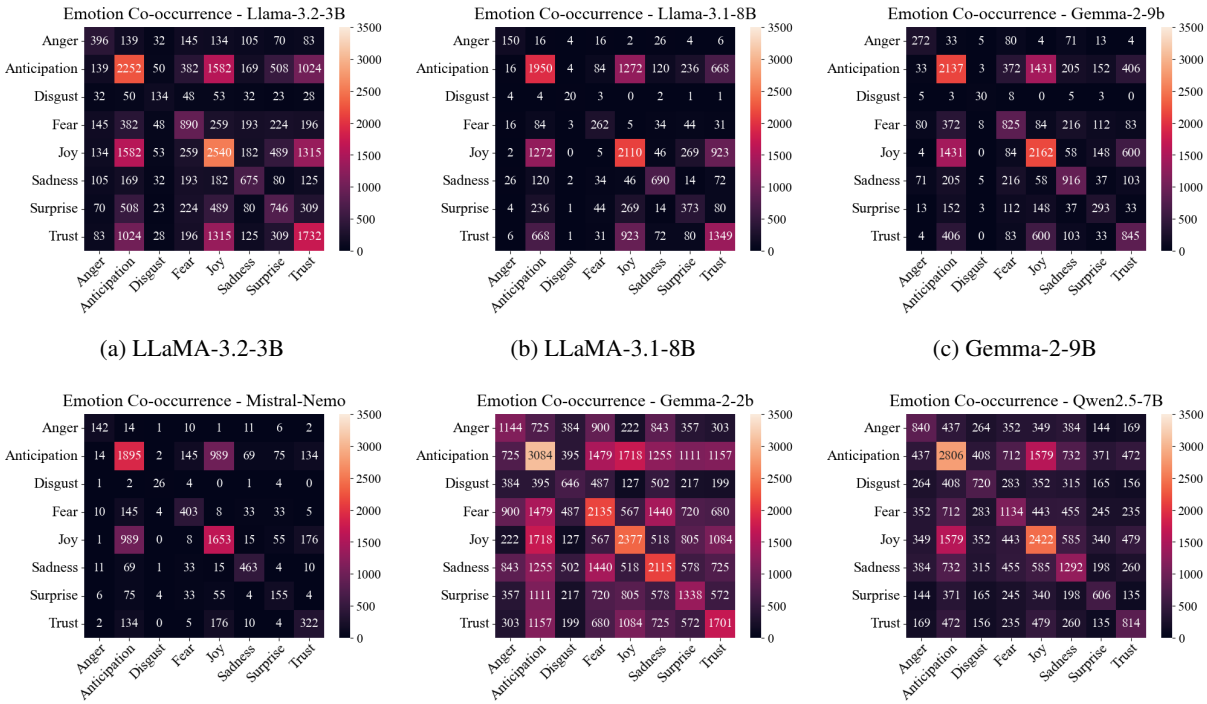
across models.

The label co-occurrences can be found in the Figure: 6. When looking at label co-occurrences of predictions, two models stand out. In Figure 6, we see a much "brighter" matrix (compared to the ground truth) for gemma-2-2b-it and Qwen2.5-7B-Instruct, meaning that these models think that emotions co-occur much more than they actually do. Thus gemma-2-2b-it and Qwen2.5-7B-Instruct over-predict emotions and have many more false positives per emotion. We see a much "darker" matrix (compared to the ground truth) for Mistral-Nemo-Instruct-2407, meaning that this model thinks that emotions co-occur much less than they actually do. Thus Mistral-Nemo-Instruct-2407 under-predicts emotions and has more false negatives per emotion.

## E Scenario Generation Prompt

The exact system and user prompts used to instruct the language model for synthetic scenario generation are detailed in Figure 7. We have prompted GPT-4o to generate the scenarios. We have used a temperature of 0.9, top_p=1.0 & max-tokens=200.

The evaluation prompt that we have used to evaluate the language models are provided in the Figure 8.




(a) LLaMA-3.2-3B (b) LLaMA-3.1-8B (c) Gemma-2-9B

(d) Mistral-Nemo (e) Gemma-2-2B (f) Qwen2.5-7B

Figure 6: Label co-occurrence patterns across different models.

**Scenario Generation Prompt**

## System Message

You are an expert at transforming structured knowledge graph information into vivid, emotionally resonant, and realistic scenarios. A knowledge graph is represented as (head, relation, tail) triplets. Events describing emotions are regarded as head entities, while frame relations constitute edge-type relations, with various corresponding attributes serving as tails. Note that a single event (head) can have multiple relations and tails. These relations are often interdependent, and the tails should be selected carefully to maintain narrative coherence and realism. You do not need to use every relation or tail—choose the ones that best support a believable and naturalistic scenario.
You will be given an event along with relations and their corresponding tails. Your task is to generate a concise, vivid, and realistic scenario (4-5 sentences MAXIMUM) that logically connects the head, relation, and tail. Avoid stylized, poetic, or abstract descriptions. **Do not use any feeling-words from the "Feelings" relation directly in your sentences**—the emotion should be implied through actions, tone, or imagery. You may add at most one extra character if needed.
Procedure for Scenario Generation:

1. Examine all available (relation, tail) pairs for the event.
2. Identify and group correlated or complementary relation-tail pairs
3. Enumerate different meaningful subsets of relation-tail pairs that can lead to a coherent narrative.
4. Choose the subset that logically connects to the event and allows for a realistic and emotionally suggestive scenario.
5. Construct a 4–5 sentence scenario using only the selected subset. Ensure the scenario has:
    - A clear flow of events
    - A realistic and natural naritive
    - Strong coreference resolution
    - Emotion shown through physical, verbal, or contextual cues

Rules:

- Use **only** the provided relations/tails—*do not* invent new details.
- Write a coherent, vivid, and emotionally expressive scenario in 4–5 sentences.
- Use **simple, everyday English**. Avoid complex sentence structures, rare words, or overly formal or literary tone. The scenario should sound like something a regular person would say or observe.
- Do not use metaphors, figurative expressions, or abstract concepts—be literal and grounded.
- You must not use any word or phrase from the "Feelings" relation directly—convey the emotion through actions, tone, and imagery instead.
- Pronouns must refer clearly to the correct entity; maintain good coreference resolution.

## User Message

Event: `head`
Do NOT use the following words in any form. These are emotion words from the Feelings relation: `feelings_list`.
Relations and Tails: `rel_tail_string`
Use only clear, simple, and natural English. Avoid complex phrases.
A concise, realistic scenario (4–5 sentences) using the most relevant relation-tail pairs. Emotion should be shown, not named.
Begin:

Figure 7: Scenario generation prompt used to instruct the language model.

**EmoScene Evaluation Prompt**

## System Message

You are an expert in emotional analysis and natural language processing. Your task is to answer whether the subject might feel the particular emotion with a yes or no. Yes indicates that the subject experiences the emotion, while no indicates that the subject doesn't experience the emotion. Think step by step to arrive at the final answer.

In your response, you must follow this exact sequence:

1. Provide a brief, step-by-step reasoning for your analysis.
2. State your confidence level in your prediction on a scale of 1 to 5 using the rubric below, enclosed within <confidence></confidence> tags. DO NOT provide an explanation for your confidence level, just the number.
3. Provide your final prediction (yes or no) between <answer></answer> tags. A "yes" indicates that the subject experiences the emotion, while a "no" indicates that the subject doesn't experience the emotion.

Confidence Rubric:

- <confidence>1</confidence> : The scenario gives almost no clues about the emotion; it is highly ambiguous.
- <confidence>2</confidence> : The scenario gives weak or unclear hints that could easily indicate another emotion.
- <confidence>3</confidence> : The scenario contains typical conversational cues for the emotion but no strong evidence.
- <confidence>4</confidence> : Most elements in the scenario clearly suggest this emotion, with little ambiguity.
- <confidence>5</confidence> : The emotion is explicitly stated or the evidence is completely clear.

## User Message

**Scenario**:
text
**Subject**:
subject
**Question**:
Does subject feel emotion?
**Note**:
Based on the above-provided information, answer the question. Please note that your output must follow the exact sequence outlined in the instructions: provide your brief reasoning, your 1-5 confidence score wrapped in <confidence></confidence> tags, and your final yes/no answer wrapped in <answer></answer> tags. DO NOT provide any explanation for your confidence score.

Figure 8: EmoScene Evaluation Prompt.