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*. 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 does not require any parameter updates and improves the structural consistency of predictions, and yields overall gains of 2.24% Lexical Accuracy without any additional cost. 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 behavioural 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 NLP, 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 a 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). As the existing NLP systems become increasingly capable of producing fluent and contextually appropriate responses, it becomes more important to know if LLMs genuinely understand emotions, or do they merely reproduce linguistic patterns associated with emotional language (Sabour et al., 2024)?

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. Recent work such as (Sabour et al., 2024) attempts to make emotion understanding and application tasks more implicit; however, it is limited to single-label settings and does not account for the co-occurrence of multiple emotions. Additionally, the proposed dataset is relatively small in scale, comprising only 200 instances.

In this work, we introduce Emotional Scenarios (EmoScene), a large-scale dataset comprising 4.7K instances, designed to evaluate emotion understanding in complex, context-rich settings while

Scenario:

Pamela sat cross-legged on the carpet, surrounded by scattered photo albums and a glowing laptop. As she flipped through the images, she stumbled upon a picture of herself with Jason from years ago, unexpectedly buried within a digital folder. Her heart skipped a beat; her palms became clammy. She chuckled softly, glancing over at her phone before sending the photo to Jason with a playful note. Not long after, her phone buzzed with his response, and they found themselves catching up after what felt like a long time.

Subject: Pamela

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

Figure 1: An example scenario from the EmoScene Dataset. The ground truth labels Joy, Surprise, and Trust must be inferred from behavioral and situational cues. The unexpected discovery ("stumbled upon", "heart skipped a beat") signals Surprise, the soft chuckle and playful outreach signal Joy, and the willingness to reconnect with Jason after a long gap signals Trust. Attending only to surface cues ("heart skipped a beat", "clammy palms") alone risks predicting Fear; the surrounding behavioral and relational context is what disambiguates the true affect.

explicitly accounting for the co-occurrence of multiple emotions. Each instance consists of a subject paired with a detailed scenario that captures environmental cues, physiological responses, and interpersonal interactions. We ground the annotation scheme in Plutchik's theory of emotions (Plutchik, 1980a) 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 modelling of complex emotional states as combinations of fundamental emotional dimensions, which is essential for multi-label settings. In contrast to (Sabour et al., 2024), which uses Plutchik's theory for developing a taxonomy that collapses multiple emotions into a single label, our approach preserves the compositional structure of emotional experiences. Furthermore, recognising that emotions rarely occur in isolation, we formulate emotion prediction as a structured inference problem. To this end, we propose an entanglement-aware Bayesian inference framework that jointly infers the emotion vector's posterior. This lightweight approach ($\sim 6 * 10^{-4}$ sec overhead per scenario) requires no parameter updates and explicitly captures dependencies among emotions, thereby addressing the independence assumptions commonly made in standard multi-label classification. As a post-processing step, it improves prediction consistency and yields an overall gain of 2.24% in lexical accuracy (CoT.

Our contributions are summarized as follows:

1. **EmoScene Benchmark.** A context-rich dataset of **4,731 scenarios** for evaluating emotion understanding in natural language grounded on Plutchik's basic emotions, where emotion prediction is formulated as an **8-dimensional multi-label task**

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

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

The code and data have been released for academic use at: EmoScene-Codes.

2 Related Work

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

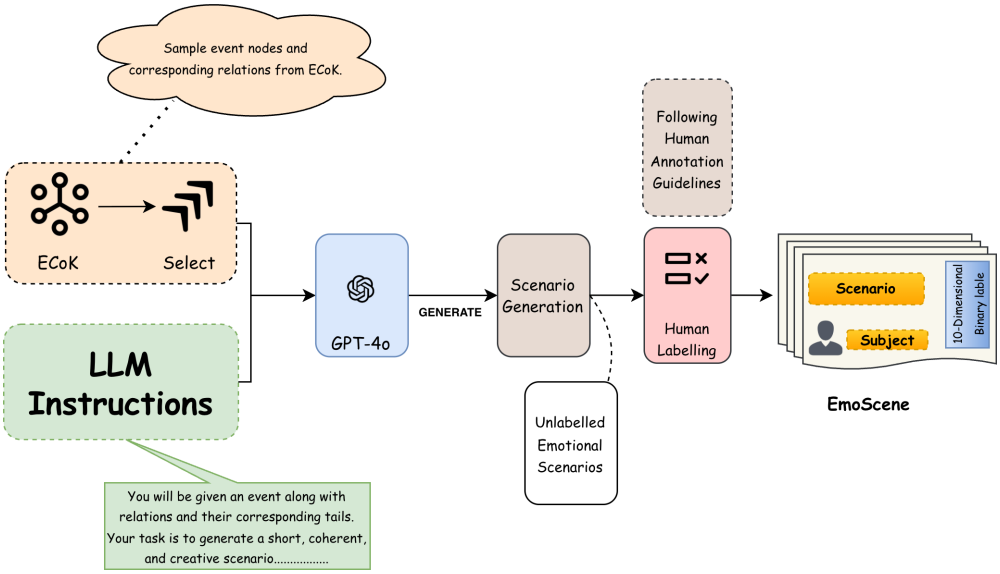


Figure 2: Overview of our steps to construct EmoScene dataset

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, 1980b); this model has been a commonly used grounding in recent work. 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 evolved from lexicon-based classification (Mohammad and Turney, 2013) to context-dependent, perspective-aware, and multimodal recognition (Fu et al., 2023; Yeo and Jaidka, 2023, 2025; Liu et al., 2024; Chandraumakantham et al., 2024). To capture interconnected affective signals, recent frameworks emphasize multi-label settings and structured relationships, such as Plutchik-grounded models (Belay et al., 2025; Xue et al., 2025; Chang, 2024). Concurrently, the rise of large language models (LLMs) has shifted research toward evaluating their broader capacity for emotional intelligence, including affect-aware reasoning and empathy (Sabour et al., 2024; Sorin et al., 2024). While studies indicate LLMs can internally represent and process ambiguous emotional cues (Hong et al., 2025; Reichman et al., 2026), their performance remains highly sensitive to contextual nuances and evaluation design, highlighting the need for more structured benchmarks, especially those that highlight the complexities arising from co-occurrences of emotions.

Emotion datasets and benchmarks. Early benchmarks focused on identifying emotions in short texts like tweets (Strapparava and Mihalcea, 2007; Mohammad et al., 2018). Subsequent datasets, such as GoEmotions (Demszky et al., 2020), EmotionLines (Chen et al., 2018), and ISEAR (Troiano et al., 2019; Scherer and Wallbott, 1990), expanded taxonomies into conversational contexts. While some work has explored long-form multi-class analysis (Liu et al., 2019) and evolving emotions (Wemmer et al., 2024), the majority of datasets still rely on short, decontextualized inputs and rigid, predefined categories that struggle to capture complex emotional states. Furthermore, varied annotation practices—ranging from crowdsourcing to expert labeling—often compromise label consistency and reliability.

Limitations of existing benchmarks and our approach. Existing benchmarks often treat emotions independently, ignoring the structured dependencies of human affect and obscuring whether models genuinely reason about emotions. Further, *to the best of our knowledge, there is no emotion understanding benchmark that captures bodily symptoms, action tendencies, while being multilabel*. To address these limitations, we introduce **EmoScene**, a benchmark of context-rich scenarios annotated with an 8-dimensional emotion vector grounded in

Plutchik's theory. We explicitly model these dependencies using a Bayesian inference framework that incorporates empirical co-occurrence patterns, thereby framing emotion understanding as a structured multi-label inference problem.

3 EmoScene Dataset

The EmoScene benchmark consists of $4,731$ scenarios designed to evaluate multi-dimensional emotion understanding. Each instance consists of a subject with a detailed scenario describing environmental cues, physiological responses, and interpersonal interactions. Following Plutchik's psychoevolutionary theory (Plutchik, 1980b), we annotate each scenario with an **8-dimensional binary emotion vector** where each dimension corresponds to Plutchik's basic emotions and allows co-occurrence of multiple labels. By preserving the compositional structure of emotional experiences, EmoScene enables the evaluation of complex emotional states as multi-label prediction tasks.

The dataset was constructed in two stages: **scenario generation** and **human annotation** (Figure 2). For scenario generation, we leverage the **Emotional Commonsense Knowledge Graph (ECoK)** (Wang et al., 2024), which consists of `<head, relation, tail>` triples. The head entities describe emotional events, while seven relation types (e.g., *causes, bodily symptoms, action tendencies*) capture different facets of the emotional experience. We leverage ECoK relations, specifically feelings (for subjective experiences), bodily symptoms (for physiological responses) and facial expressions (for the behavioural manifestation), which is crucial to appropriately describe subtle cues required for accurate emotion understanding. Specifically, this method of scenario generation ensures that its structured schema embeds situational, physiological, and interpersonal cues, yielding richer and more realistic context than short-text or social media datasets.

Using these triples as conceptual seeds, we prompt GPT-4o to generate the context-rich scenarios (the full generation prompt is provided in Appendix G). A comprehensive analysis of the dataset—including dataset statistics, topic diversity, and label distributions—is detailed in Appendices A and B.

Annotation and inter-annotator agreement Three expert annotators (English instructors with prior annotation experience) independently labelled every scenario by answering eight yes/no questios, one per emotion. Final Labels were assigned by strict majority vote. Before annotation, a training session has been conducted individuall to all annotators which involves explaining the guidelines and providing definitions of various emotions. Inter-annotator agreement (Fleiss' κ) varies by emotion, yielding a mean of 0.43. This is consistent with prior popular subjective-emotion benchmarks such as GoEmotions (Demszky et al., 2020) and falls in the moderate range described by McHugh (2012). Notably, Trust is the most prevalent emotion in EmoScene (48.1%), yet it proves the most difficult to label consistently and is frequently missed by LLMs (see §6.1). Annotators are compensated accordingly.

4 Methodology

Traditional multi-label frameworks typically model emotions as independent Bernoulli variables, failing to capture the structured interdependencies nature of human affect, where emotions are rarely independent and often co-occur in structured, interdepended ways. To address this limitations, we introduce Emotion Entanglement, a perspective that treats emotional states as a coupled system rather than disjoint labels. Building on this perspective, we propose an Entanglement-Aware Bayesian Inference framework that explicitly models the joint distribution over emotions, incorporating co-occurrence dependencies into the inference process to produce more coherent, context-sensitive predictions without requiring expensive parameter updates.

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 $E = (E_1, \ldots, E_8)$, the joint distribution does not factorize:

$$P(E|x) \neq \prod_{i=1}^{8} P(E_i|x) \tag{1}$$

To quantify this, we analyse 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

Model	Lexical Accuracy		Vector Accuracy		Hamming Loss		F1-Macro	
	Baseline	BI	Baseline	BI	Baseline	BI	Baseline	BI
Chain-of-Thought (CoT) Prompting								
Llama-3.1-8B	83.87	83.88	24.54	24.50	16.13	16.12	47.2	47.2
Llama-3.2-3B	79.76	79.76	17.40	17.40	20.24	20.24	45.1	45.1
Llama-3.1-70B	81.93	83.27	22.15	23.95	18.07	16.73	50.6	51.7
Mistral-Nemo	83.48	83.50	22.24	22.26	16.52	16.50	43.1	43.2
Qwen2.5-7B	75.95	82.22	14.06	21.48	24.05	17.78	42.8	47.9
Gemma-2-2b	70.93	78.79	10.27	15.96	29.07	21.21	43.7	47.4
Gemma-2-9b	83.37	83.48	24.88	25.07	16.63	16.52	50.0	50.1
Overall	79.89	82.13 (**+2.24**)	19.36	21.51 (**+2.15**)	20.11	17.87 (**−2.24**)	46.07	47.51 (**+1.44**)
Zero-Shot Prompting								
Llama-3.1-8B	77.92	82.04	14.61	20.10	22.08	17.96	50.5	53.0
Llama-3.2-3B	83.43	83.53	24.12	24.33	16.57	16.47	48.3	48.4
Llama-3.1-70B	75.57	80.91	10.95	17.42	24.43	19.09	50.2	53.7
Mistral-Nemo	83.54	83.54	21.69	21.69	16.46	16.46	44.0	44.0
Qwen2.5-7B	83.32	83.36	22.09	22.13	16.68	16.64	44.7	44.8
Gemma-2-2b	77.21	77.57	13.89	14.23	22.79	22.43	42.8	43.1
Gemma-2-9b	81.67	82.68	21.71	22.87	18.33	17.32	50.6	51.1
Overall	80.38	81.95 (+1.57)	18.43	20.39 (+1.96)	19.62	18.05 (−1.57)	47.3	48.3 (+1.00)

Table 1: Model Performance Comparison on EmoScene. We compare Baseline outputs against our Entanglement-Aware Bayesian Inference (BI) post-processing step across both Chain-of-Thought and Zero-Shot settings.

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 $E = (E_1, \ldots, E_8)$ 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 θ_i represents the bias for individual emotions and θ_{ij} represents the coupling strength between emotion pairs. These parameters are estimated from the training set via maximum-entropy solutions:

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

Critically, these learned couplings θ_{ij} empirically recover the dyadic structure of Plutchik's wheel, such as strong mutual exclusion of bipolar opposites (see Appendix C.1 for more details)

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:

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

Since we utilize a fixed space of 8 labels, the exact argmax is found by evaluating all 2^8 combinations, ensuring a parameter-efficient post-processing step that explicitly models emotion co-occurrence fidelity.

5 Experimental Setup

In this section, we describe the models and hyperparameters that we have used to run the experiments. Details about confidence extraction is provided in the Appendix D

Models Details. We evaluate seven open-source instruction-tuned LLMs from HuggingFace—`Mistral-Nemo-2407-Instruct`, `Llama-3.2-3B-Instruct`, `Gemma-2-2B-it`, `Qwen2.5-7B-Instruct`, `Gemma-2-9B-it`, and `Llama-3.1-8B-Instruct`, `Llama-3.1-70B-Instruct`—run locally on 4× NVIDIA A100 80GB GPUs.

Evaluation Metrics. We evaluate performance using **Lexical Accuracy (LA)** (fraction of correctly predicted emotion cells), **Vector Accuracy (VA)** (exact-match fraction of 8-dimensional vectors), and **Hamming Loss (HL)** (1 − LA). Our primary metric is **Macro F1**, which averages the F1 scores across all eight emotions equally, ensuring performance on rare emotions (e.g., Disgust, Surprise) is weighted equally to common ones (Mohammad and Turney, 2013).

Confidence Extraction and Bayesian Inference. The Bayesian Inference (BI) step relies on a Bernoulli likelihood $P(x|E)$, populated by soft-maxing the model's `yes`/`no` logits at the answer position. The Ising prior (θ_i, θ_{ij}) is derived from the full EmoScene label matrix. We perform exact inference by enumerating all $2^8 = 256$ label configurations and sweep the hyperparameter $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 1.0, 2.0, 5.0\}$ to find the per-model F1-optimal value (see §E.2).

Model	α	Baseline F1	Linear F1	Full-prior F1
Llama-3.1-8B	0.10	47.15	47.15	47.16
Llama-3.2-3B	0.00	45.13	45.13	45.13
Llama-3.1-70B	0.50	50.60	51.50	51.75
Mistral-Nemo	0.25	43.10	43.15	43.17
Qwen2.5-7B	0.75	42.81	46.15	47.89
Gemma-2-2B	0.75	43.68	46.17	47.36
Gemma-2-9B	0.50	50.02	49.98	50.07

Table 2: Effect of ablating the pairwise term θ_{ij} at each model's F1-optimal α.

6 Results & Analysis

The main results in Table 1 contain the evaluation results for both prompting strategies: CoT & ZS. For both prompts we compare the LLM's baseline prediction against our Bayesian Inference (BI) post-processing. Evaluation prompt can be found in Figure 9 & Figure 10. More details on Prompting Strategies, Significance of α and confidence calibration are provided in the Appendix E

Across the six open-source models, Bayesian Inference (BI) delivers its largest gains on the baselines that most over-predict rare emotions: **Qwen2.5-7B** (+5.1 F1, +6.27 LA) and **Gemma-2-2B** (+3.7 F1, +7.86 LA). BI's marginal prior θ_i pulls these over-predictions toward the empirical zero-rate, recovering F1 without retraining. For the four better-calibrated baselines (baseline LA $>$ 79%), gains are smaller as there is less over-prediction to correct; on Llama-3.2-3B, the optimal α is 0 and BI has no effect. Llama-3.1-70B achieves the highest baseline Macro F1 (**50.6**) narrowly beating the best open-source model, Gemma-2-9B which has 50.0 MacroF1 confirming EmoScene's difficulty.

Despite high lexical scores, Vector Accuracy remains modest (15–25%). Because BI performs joint MAP inference over the full emotion vector **E**, it enforces an entanglement prior via an MRF/Ising graph that models pairwise correlations. This corrects the LLM independence assumption by reweighting predictions to match real co-occurrence patterns. BI consistently improves Hamming Loss and Macro F1, demonstrating BI's effectiveness for robust, scene-aware emotion classification.

Where BI does and does not help? For baselines that are already well-calibrated to the emotion marginals (Llama-3.1-8B, Gemma-2-9B), per-emotion precision (refer Appendix E.1) is already high and BI has little room to operate consistent

with the small overall gains in Table 1. For baselines whose lexical-accuracy is dragged down by over-prediction on rare emotions (Qwen2.5-7B, Gemma-2-2B), BI recovers $+5$ to $+8$ LA points specifically by pruning these false positives.

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 θ_{ij} favours the co-occurrence of Sadness and Anticipation but penalises unlikely associations with Anger.

Efficiency Analysis To evaluate BI's inference overhead, we measure the additional latency using the largest model, `Llama-3.1-70B`. BI adds only 6×10^{-4} seconds per scenario on average, indicating negligible inference overhead.

6.1 Error Analysis

To better understand where LLMs fail on EmoScene, we systematically rank all candidate error patterns by their frequency across the six open-source models. We consider two types of errors: *single-emotion errors* (each of the 8 emotions, as a false negative or false positive: 16 candidates in total) and *emotion-pair co-occurrence violations* (the 28 emotion pairs where both labels are present in the ground truth but the model predicts only one of them). The top two patterns from each category give us the four dominant error modes **P1–P4** (refer Table 5 in Appendix). The counts in the table are aggregated across all six open-source models; the illustrative scenarios that follow are drawn from the best-performing baseline, Gemma-2-9B (Macro F1 = 0.500).

Single-emotion errors (P1, P2). The most common single-emotion failure is *Trust under-prediction*. Every model misses Trust in at least 1,133 of the 4,731 scenarios, and this pattern holds across model families and sizes. This is consistent with the low inter-annotator agreement we observed for Trust (Fleiss $\kappa = 0.272$): Trust is at the same time the most prevalent emotion in EmoScene, the hardest emotion to label consistently, and the most-missed emotion by LLMs. The most common over-prediction is *Anticipation*: models often read forward-looking language as Anticipation even when the underlying emotion is something else. The boxes P1 and P2 below show one representative example of each.

P1: Trust under-prediction

True: Trust. *Baseline:* (none).

Keisha noticed the baby starting to cry in its stroller, and **without a moment's hesitation, she bent down with a caring gaze**. As she gently spoke soothing words, her brows furrowed slightly, scanning the small face for any signs of distress. The mother, looking flustered and juggling bags, caught Keisha's eye. **Keisha offered to hold the baby**, suggesting the mother take a quick break to collect herself, which she gratefully accepted.

The model's own reasoning describes the actions accurately as "caring and helpful towards the distressed baby and the flustered mother," but then concludes "there's no direct indication that Keisha feels trust." The model recognises the caring behaviour, but treats Trust as something that needs to be explicitly named in the text. In Plutchik's framework, however, this kind of warm relational gesture – offering help, attending to a stranger's distress – is precisely what Trust looks like.

P2: Anticipation over-prediction

True: Fear. *Baseline:* Fear, Anticipation, Sadness.

Sylvia sat quietly in the corner of the bustling café, stirring her coffee absentmindedly and **avoiding eye contact with the lively group nearby**. She watched them laugh and chat effortlessly, **feeling her palms dampen** as she considered **what to say if she joined them**. **Memories of a recent awkward exchange resurfaced**, making her glance away and fidget with the napkin in her lap. She knew she needed to practice more, so she decided to start small, smiling softly at the barista as she left her tip.

The dominant emotion here is Fear – social anxiety, indicated by the dampening palms, the avoided eye contact, and the memory of an awkward exchange. The model correctly picks up Fear, but it also predicts Anticipation, and the model's own reasoning makes the confusion explicit: it states that "Sylvia is considering joining the group and feels anxious about it, which suggests she might be anticipating the interaction." But anxious imagining is not anticipation. Sylvia is not planning a future event; she is paralysed by social fear and turning the situation over in her head. The model conflates the conditional "what to say if she joined them" with eager planning. At the emotion-pair level, the top two co-occurrence violations (P3: Anticipation+Trust, P4: Joy+Trust) both involve Trust as the missed label. Representative examples and detailed analysis are provided in Appendix E.5

Takeaway: the Trust convergence. In EmoScene, Trust is the most prevalent emotion (48.1% of scenarios) yet has the *lowest* inter-annotator agreement (Fleiss $\kappa = 0.272$), reflecting genuine annotator disagreement on what counts as Trust. In Figure 6, Trust shows the largest gap between its prevalence and its per-model recall. In the error analysis above, three of the four dominant error patterns (P1, P3, P4) are Trust-related. The implication is the same in all three places: Trust is signalled *relationally and behaviourally* in human language – offering help, listening attentively, easing posture – but LLMs treat it as something that has to be explicitly named. The entanglement prior recovers Trust precisely where it co-occurs with cues the model *does* catch (Joy, Anticipation), which is the mechanism behind BI's gains on Qwen2.5-7B and Gemma-2-2B.

6.2 Linear-only Prior Ablation

The BI Ising prior combines a *linear* (marginal) term $\sum_i \theta_i E_i$ and a *pairwise* (entanglement) term $\sum_{i<j} \theta_{ij} E_i E_j$. To isolate the entanglement contribution, we ablate the pairwise dependencies ($\theta_{ij} = 0$) at each model's F1-optimal α (Table 2). For the models where BI is most effective, this pairwise term accounts for 32–34% of the total absolute BI gain (+1.7 F1 for Qwen2.5-7B; +1.2 for Gemma-2-2B). Relative to the linear-only baseline, incorporating entanglement boosts F1 gains by **48–55%** (from +3.3 to +5.1 for Qwen2.5-7B, and +2.5 to +3.7 for Gemma-2-2B). This confirms that while the marginal term drives the bulk of the improvement, the full Ising prior is critical for maximizing performance.

7 Conclusion and Future Work

We introduced **EmoScene**, a context-rich benchmark for evaluating multi-dimensional emotion understanding in language models. 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 chain-of-thought & 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 open-source 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 LLMs.

Limitations

While EmoScene provides a context-rich benchmark for multi-dimensional emotion understand-

ing, 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

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.

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.

Edward Y. Chang. 2024. Behavioral emotion analysis model for large language models. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 549–556.

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.

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.

Chen Liu, Muhammad Osama, and Anderson de Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pages 6293–6298, Hong Kong, China. Association for Computational Linguistics.

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.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

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. 1980a. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Robert Plutchik. 1980b. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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*, 21(3):307–322.

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.

Eileen Wemmer, Sofie Labat, and Roman Klinger. 2024. EmoProgress: Cumulated emotion progression analysis in dreams and customer service dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677, Torino, Italia. ELRA and ICCL.

Jieying Xue, Phuong Minh Nguyen, Minh Le Nguyen, and Xin Liu. 2025. JNLP at SemEval-2025 task 11: Cross-lingual multi-label emotion detection using generative models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 20–27, Vienna, Austria. Association for Computational Linguistics.

Gerard Yeo and Kokil Jaidka. 2023. The PEACE-reviews dataset: Modeling cognitive appraisals in emotion text analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2822–2840, Singapore. Association for Computational Linguistics.

Gerard Christopher Yeo and Kokil Jaidka. 2025. Understanding annotator perception: Modeling psychological inference from first-and third-person annotations (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29548–29550.

A Dataset Statistics

Table 3 summarises the macro statistics of EmoScene. Scenarios contain rich contextual descriptions with an average length of **82 words** across roughly 4 sentences each, drawing on a vocabulary of **13,869 unique tokens**. About **60.9%** of scenarios carry more than one emotion label, and 8.3% are labelled with no emotion (the annotators judged that no Plutchik basic emotion was clearly conveyed). The mean label cardinality is **1.77**. Figure 3 shows the marginal label count per emotion.

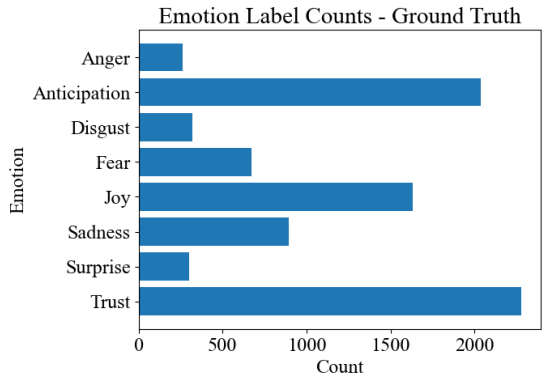


Figure 3: Emotion Label counts from the EmoScene Dataset.

Statistic	Value
Total number of scenarios	4,731
Avg. scenario length (words)	82
Median scenario length (words)	82
Max scenario length (words)	129
Min scenario length (words)	49
Avg. sentences per scenario	4.24
Vocabulary size (unique tokens)	13,869
Zero-emotion scenarios	395 (8.3%)
Single-emotion scenarios	1,455 (30.7%)
Multi-label scenarios (≥ 2)	2,881 (60.9%)
Average label cardinality	1.77

Table 3: Statistics of the EmoScene dataset.

B Dataset Analysis

To ensure the robustness of the benchmark, we performed extensive post-hoc analyses on topic distribution and label skew. We verified thematic breadth by clustering scenarios. Additionally, we analyze the prevalence of Anticipation and Trust in our dataset, attributing their frequency to the inherent temporal and interpersonal nature of the source ECoK events.

B.1 Topic diversity.

To verify that EmoScene covers a diverse range of everyday situations rather than a narrow topical band, we embed all 4,731 scenarios using SBERT framework (Reimers and Gurevych, 2019) with the `all-MiniLM-L6-v2` encoder and cluster their 384-dimensional embeddings into eight groups via KMeans. The resulting clusters are roughly balanced in size (258 to 746 scenarios) and span clearly distinct themes: food and dining, music and concerts, office and work, family and social warmth, physical wellness, illness and rest, outdoor everyday life, and studying. We interpret these clusters deterministically by extracting class-conditional TF-IDF (c-TF-IDF) keywords, excluding common stop-words and proper nouns (tokens capitalised $> 70\%$ of the time). The resulting cluster keywords and dominant emotions are detailed in Table 4, with a 2-D UMAP projection of the embeddings shown in Figure 4.

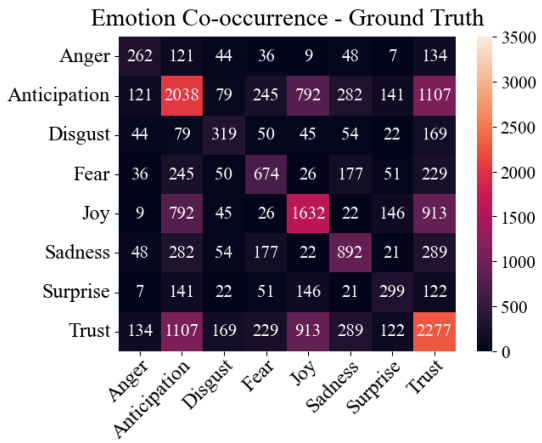
B.2 Anticipation and Trust over-representation.

Figure 3 for the label counts in EmoScene. Two emotions are over-represented in the marginal label distribution: Anticipation (43.1%) and Trust (48.1%). This reflects the structure of the underlying scenes rather than annotation drift. ECoK's source events are inherently temporal: each event ties together a cause, an outcome, and an action tendency, so scenarios derived from them naturally contain forward-looking narrative elements. Empirically, scenarios labelled Anticipation contain forward-looking lexical markers (*soon, will, plan, hope, prepare,* etc.) at roughly **1.7**$\times$ the rate of non-Anticipation scenarios. Trust over-representation similarly tracks ECoK's emphasis on bodily-symptom and feeling relations, which place scenarios in interpersonal contexts.

Cluster	Size	Top c-TF-IDF keywords
0 – Music / concerts	201	concert, song, band
1 – Celebrations	255	warmly, celebrate, genuine
2 – School / study	241	desk staring, cluttered desk, teacher
3 – Driving / transport	170	road, bus, steering wheel
4 – Social media	182	feed, clicked, tweet
5 – Sports / injury	141	gym, teammates, injury
6 – Emotional talk	172	tears, silence, talk
7 – Cooking / food	147	meal, aroma, food
8 – Meetings / work	292	meeting room, talk, sensing
9 – Outdoors / nature	43	grass, afternoon sun, tea
10 – Office / productivity	260	cluttered desk, papers, keyboard
11 – Technology	232	typed, tech, forum
12 – Errands / shopping	150	store, shopping, bills
13 – Rest / fatigue	158	sleep, eyelids, alarm
14 – Health / illness	192	pain, doctor, sleep
15 – Humor / bonding	251	genuine, laugh, funny
16 – Fitness	79	gym, rubbed temples, cheering
17 – Parenting / children	72	kids, felt tightness, desk staring
18 – Fashion / appearance	185	dress, stood mirror, fabric
19 – Pets	78	dog, cat, crouched
20 – Weather / outdoors	162	rain, sunscreen, clouds
21 – Cooking (detail)	129	ingredients, meal, aroma
22 – Travel / commute	240	train, steering wheel, wheel
23 – Emergency / medical	235	manager, doctor, emergency
24 – Entertainment / media	213	credits, episode, baby
25 – Academic / discourse	177	conference, audience, discussion
26 – Nighttime / leisure	74	forward slightly, popcorn, night sleep

Table 4: Topic clusters (k=27, elbow-optimised) of EmoScene scenarios derived from KMeans on SBERT embeddings, labelled by c-TF-IDF top keywords. The high optimal k confirms fine-grained topical diversity across the corpus.

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.

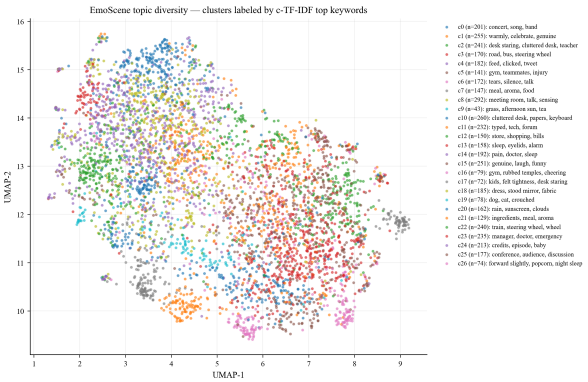


Figure 4: UMAP projection of SBERT embeddings of all 4,731 EmoScene scenarios, coloured by KMeans cluster ($k = 27$, Elbow optimized). Clusters separate cleanly into distinct topical regions.

C.1 Empirical Validation of Entanglement Prior against Plutchik's Theory

Figure 5 visualises the learned θ_{ij} matrix. The structure of the prior is not an arbitrary statistical artifact of EmoScene – it operationalises Plutchik's wheel. The strongest positive couplings are exactly the pairs Plutchik identified as primary dyads: *Anger* + *Disgust* ($\theta_{ij} = +0.91$, his "contempt" dyad), *Joy* + *Trust* ($+0.15$, the "love" dyad), *Anticipation* + *Trust* ($+0.12$), and *Fear* + *Sadness* ($+0.33$). Symmetrically, the strongest negative couplings fall on Plutchik's opposite axes: *Joy* ↔ *Sadness* (-2.64), *Anger* ↔ *Joy* (-2.31), *Fear* ↔ *Joy* (-2.19). The prior is therefore Plutchik-consistent by data, not by construction: the wheel structure emerges directly from EmoScene's annotation co-occurrence. This is the empirical link between our theoretical positioning (§3) and the inference machinery – the prior θ_{ij} is precisely what makes the Bayesian step "entanglement-aware" in Plutchik's sense, and an ablation removing it loses 48–52% of BI's F1 gain on the models where BI is most effective (Appendix 6.2).

D Extended Experimental setup

This section contains the details about Prompting Strategies that are used in the experiments.

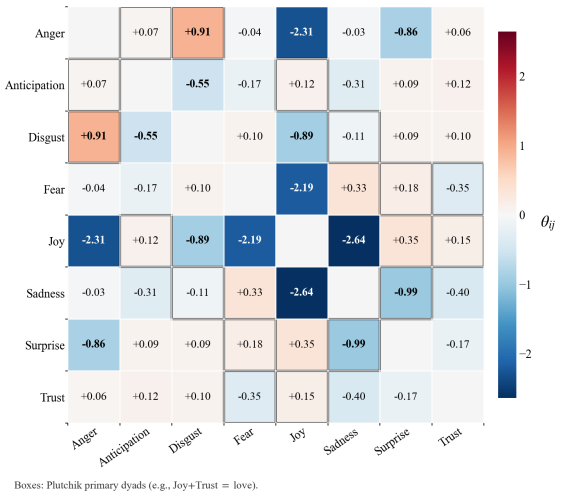


Figure 5: Learned pairwise couplings θ_{ij} on EmoScene. Red cells indicate positive couplings (the pair co-occurs more often than under independence) and blue cells indicate negative couplings (mutual exclusion). The eight cells outlined in black correspond to Plutchik's primary dyads (adjacent on his wheel; e.g., Joy+Trust = "love"). Five of the eight primary dyads carry positive couplings, including the strongest overall (Disgust+Anger = "contempt", $+0.91$), and the largest negative coupling (Joy+Sadness, -2.64) coincides with Plutchik's primary bipolar axis. The remaining three primary dyads ("submission", "disapproval", "remorse") appear with negative sign, reflecting corpus-specific deviations from Plutchik's symmetric prediction.

Prompting and Decoding. We evaluate two configurations: a **chain-of-thought (CoT)** prompt that elicits step-by-step reasoning and a self-reported confidence score (1–5) before outputting a yes/no label, and a **zero-shot (ZS)** prompt that requires only the label. For open-source models, we use greedy decoding ($T = 0$) with a max limit of 512 tokens for CoT and 16 tokens for ZS.

E Extended Results & Analysis

This section includes details about the effect of α and Sensitivity of prompting strategies.

E.1 Per-Emotion Performance

Per-emotion precision and recall heatmaps for all seven open-source models are shown in Figure 6. Three patterns are consistent across the model panel. **Joy** achieves high precision and recall across every model; lexical cues for joy (smile, laughter, warmth) are reliable and frequent in EmoScene. **Anticipation** shows moderate precision and recall: it is over-predicted by smaller models (Gemma-2-2B), consistent with the false-positive pattern P2 in §6.1. **Disgust** and **Surprise** suffer critically low recall (near-zero in Llama-3.2-3B and Mistral-Nemo), indicating models predict these emotions only when evidence is (abundant). This is the over-prediction–vs–conservatism axis that BI exploits: the marginal prior θ_i pulls baselines that over-predict rare emotions (Qwen2.5-7B, Gemma-2-2B) toward the empirical zero-rate, raising precision without losing the high-recall Joy and Anticipation cells.

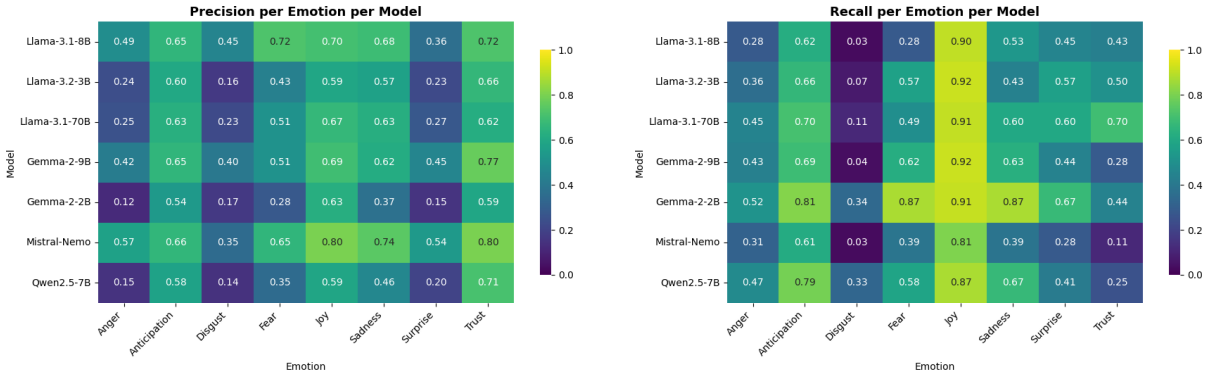
E.2 Significance of α

The hyperparameter α in our MAP objective controls how strongly the entanglement prior pulls the model's marginal predictions toward the empirical co-occurrence structure of EmoScene. We sweep α over $\{0, 0.1, 0.25, 0.5, 0.75, 1, 2, 5\}$ for each model and report the per-model F1-optimal value (the same value is used for Table 1). Figure 7 plots Macro F1 as a function of α for each model, alongside a linear-only ablation in which the pairwise term θ_{ij} is held at zero.

Three observations follow. **(1) Smooth curves.** The F1-vs-α curve is smooth and well-behaved for every model, with no narrow peak. BI's gains are not the artefact of fine-tuning a single value of α; small mis-calibrations in either direction leave the result essentially unchanged. **(2) Llama-3.2-3B's flat curve explains its zero improvement.** For Llama-3.2-3B, Macro F1 stays at the baseline value across all α, so the F1-optimal $\alpha = 0$ is a benign symptom of an inert response rather than a sign of α-shopping or BI failing to engage. This model's marginal predictions are already well-calibrated to EmoScene's emotion distribution, so the prior θ_{ij} has nothing to add. **(3) The pairwise term contributes meaningfully where BI helps.** For the two over-predicting baselines (Qwen2.5-7B, Gemma-2-2B), the dashed linear-only curve sits visibly below the solid full-prior curve at the operating α; an ablation (Appendix 6.2) quantifies that removing θ_{ij} at the operating α erases 48–52% of BI's F1 gain. The entanglement term is therefore not decorative – on the models where BI is effective, roughly half of the gain comes specifically from modelling pairwise emotion co-occurrence.

E.3 Effect of Prompting Strategy

The results in Table 1 contains results corresponding to two prompt variants. First, chain-of-thought that asks for step-by-step reasoning along with the answer and other is the zero-shot that only asks for



(a) Precision per Emotion per Model

(b) Recall per Emotion per Model

Figure 6: 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.

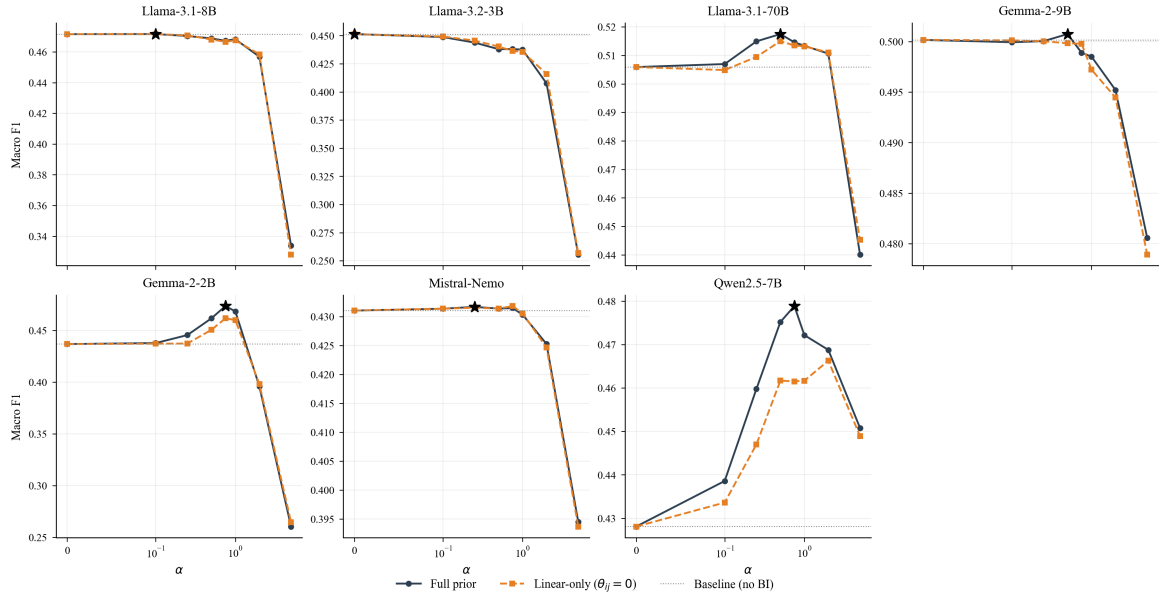


Figure 7: Macro F1 as a function of α for each open-source model under two prior configurations: the full entanglement-aware prior (solid) and the linear-only ablation with $\theta_{ij} = 0$ (dashed). Black stars mark the F1-optimal α used in Table 1; the dotted horizontal line marks the baseline ($\alpha = 0$).

answer without any intermediate reasoning.

Two observations follow. First, BI provides positive aggregate gains under *both* prompting strategies, indicating that the benefit of modeling emotion entanglement is not specific to any particular reasoning style. Second, several models attain a higher baseline F1 under the no-CoT prompt (e.g., Llama-3.1-8B: 0.505 vs. 0.472; Gemma-2-9b: 0.506 vs. 0.500), suggesting that step-by-step reasoning is not uniformly helpful for this task and can amplify systematic biases in the model's emotion priors.

E.4 Confidence and Calibration

Under our CoT prompt each model produces two confidence signals: a self-reported integer in $\{1, 2, 3, 4, 5\}$ inside the ⟨confidence⟩ tags, and a logit-derived value $\max(P_{\text{yes}}, P_{\text{no}})$ computed from the softmax at the answer token. We measure how well each signal tracks actual accuracy on EmoScene using the Expected Calibration Error (ECE). Per-model ECE values are reported in Appendix F.

At the token level, all six open-source models are over-confident. Mean logit-derived confidence ranges from 0.92 to 0.998 across models, while their actual accuracy is only 0.71 to 0.84. Token-level ECE ranges from 0.149 (Llama-3.1-

Pattern	Llama-3.1-8B	Llama-3.2-3B	Llama-3.1-70B	Gemma-2-9B	Gemma-2-2B	Mistral-Nemo	Qwen2.5-7B
P1: Trust under-prediction (FN)	1307	1133	694	1629	1270	2020	1699
P2: Anticipation over-prediction (FP)	677	904	842	738	1424	650	1190
P3: Anticipation + Trust co-occurrence miss	557	500	428	644	596	647	734
P4: Joy + Trust co-occurrence miss	392	269	263	529	415	640	586

Table 5: Incidence of the four dominant error patterns across six baseline LLMs on EmoScene. Each pattern is defined by a deterministic Boolean rule on the prediction array; counts are over all 4,731 scenarios. **P1–P2** are the top-2 single-emotion error modes (ranked across 8 emotions $\times$ {FN, FP}); **P3–P4** are the top-2 emotion-pair co-occurrence violations (ranked across the 28 emotion pairs).

Model	Q1 (lowest)	Q2	Q3	Q4 (highest)
Qwen2.5-7B	+1.06	+0.60	+0.31	+0.03
Gemma-2-2B	+1.49	+0.64	+0.33	+0.05
Gemma-2-9B	+0.04	~ 0	~ 0	~ 0
Llama-3.1-8B	~ 0	~ 0	~ 0	~ 0

Table 6: Mean per-scenario BI gain (Δcorrect-emotion-cells, out of 8) within each quartile of the baseline's logit-derived confidence. For the two models where BI is most effective (top), the gain decreases monotonically with confidence. For uniformly high-confidence baselines (bottom), the gain is small in every quartile.

8B) to 0.232 (Gemma-2-2B). Self-reported confidence shows a different pattern: some models are under-confident (Llama-3.1-8B reports a mean of 0.47/1.0 but is accurate 0.84 of the time; Qwen2.5-7B reports 0.43 and is accurate 0.76), while Mistral-Nemo is nearly perfectly calibrated.

The most useful finding for BI is in Table 6. For the two models where BI is effective, the gain scales directly with baseline uncertainty. In the lowest-confidence quartile (Q1), BI corrects an additional 1.06 (Qwen2.5-7B) to 1.49 (Gemma-2-2B) emotion cells per scenario on average. In the highest-confidence quartile (Q4), the same gain shrinks to 0.03 and 0.05. The decrease is monotonic across all four quartiles. This trend is expected is consistent with the behaviour from an uncertainty-aware prior: **BI provides the largest improvements in regions where the baseline model exhibits lower confidence**. For models whose baseline is uniformly high-confidence (Gemma-2-9B, Llama-3.1-8B), the same per-quartile gains are essentially zero – there is no uncertainty for the prior to compensate, which matches their small overall BI gains in Table 1.

E.5 Co-occurrence Error Patterns

At the emotion-pair level, the top two failure modes both involve Trust. Anticipation+Trust (P3) and Joy+Trust (P4) are the two pairs most often broken by the baseline, where the model predicts one of the pair but not the other. Together, P3 and P4 covers 661 to 1,320 scenarios per model, which is comparable in scale to the single-emotion patterns. These are exactly the failure modes that our entanglement term θ_{ij} is designed to fix. As we show in the Appendix, removing θ_{ij} removes 48–52% of BI's F1 gain on the models where BI is most effective. The boxes P3 and P4 below show one example of each.

P3: Anticipation + Trust co-occurrence miss

True: Anticipation, Trust. *Baseline:* Anticipation.

Edna stood on the edge of the field, part of a group of fellow football players, her new gear snug and ready for **the upcoming match**. Her heart was pounding steadily, and a light sheen of sweat glistened on her brow. She took a deep breath, muscles taut, and began her stretching routine, glancing occasionally towards her teammates. Edna's jaw was clenched as **she listened to their captain strategize, absorbing every word, her eyes steady and sure**.

The model correctly picks up Anticipation ("her physical symptoms and her focused attention on the upcoming match and her captain's strategy all point towards anticipation"), but it misses Trust. Its own reasoning explains why: "these actions could indicate respect or determination, [but] there's no direct evidence suggesting trust towards her teammates or the captain." The captain-team listening dynamic – absorbing every word, a steady gaze – is exactly the kind of relational Trust the entanglement prior θ_{ij} is designed to recover. In the learned prior, Anticipation and Trust are positively coupled ($\theta_{ij} = +0.12$): whenever Anticipation is correctly identified in a team setting, the coupling raises the posterior on Trust as well.

Model	Self-reported ECE	Logit-derived ECE
Llama-3.1-8B	0.400	0.149
Llama-3.2-3B	0.158	0.175
Llama-3.1-70B	0.075	0.145
Mistral-Nemo	0.090	0.164
Qwen2.5-7B	0.370	0.163
Gemma-2-2B	0.118	0.232
Gemma-2-9B	0.306	0.162

Table 7: Expected Calibration Error for each model under the CoT prompt.

P4: Joy + Trust co-occurrence miss

True: Joy, Trust. *Baseline:* Joy.

John stood with Anna in the bustling café, their eyes meeting for the first time in weeks. **Anna smiled warmly and remarked on John's striking new jacket**, noting how well it suited him. As she spoke, **John's face felt warmer, a smile spreading effortlessly across his lips, his posture easing**. He nodded in thanks, deciding he would keep experimenting with different styles, maybe even following a few fashion influencers for new ideas.

The model correctly identifies Joy from the warm smile and easing posture, but misses Trust with the reasoning that "while warmth and positive feelings are present, there's no explicit indication of trust being established or felt." Yet John's posture easing in Anna's presence – and his open willingness to take her suggestion – is precisely the kind of physical comfort that signals Trust in an intimate exchange. Joy and Trust together form Plutchik's "love" dyad, and the entanglement prior θ_{ij} is positively coupled here ($\theta_{ij} = +0.15$): whenever Joy is identified in a warm interpersonal exchange, the prior pulls the posterior on Trust upward, recovering exactly this kind of miss.

F Confidence Analysis

We measure calibration of each model's confidence signals against actual accuracy on EmoScene using the Expected Calibration Error (ECE). Two confidence signals are available under the CoT prompt: a self-reported integer in $\{1, 2, 3, 4, 5\}$ from the ⟨confidence⟩ tags, and a logit-derived value $\max(P_{\text{yes}}, P_{\text{no}})$ from the softmax at the answer token. Per-model ECE values are listed in Table 7.

At the token level (logit-derived), every open-source model is *over-confident*: mean confidence ranges from 0.92 to 0.998, but actual accuracy is only 0.71 to 0.84. The token-level ECEs sit between 0.149 (Llama-3.1-8B) and 0.232 (Gemma-2-2B), with the gap concentrated in the highest-confidence bins – the model says it is sure roughly 97% of the time but is wrong 15–30% of those cases. This is consistent across all six models at the logit level, current open-source LLMs are uniformly overconfident on subjective emotion judgements on EmoScene.

Self-reported confidence (the ⟨confidence⟩ tag) shows the opposite pattern for several models. Llama-3.1-8B reports a mean confidence of 0.47/1.0 while being accurate 0.84 of the time, and Qwen2.5-7B reports 0.43 while being accurate 0.76 – both are noticeably *under*-confident in their verbal reports relative to their actual accuracy (ECE = 0.40 and 0.37 respectively). Mistral-Nemo is nearly perfectly calibrated by self-report (ECE 0.090 and 0.078). The gap between the two signals – token-level over-confidence and report-level under-confidence on the same model – suggests that the verbal report is not a faithful reflection of the underlying token-level distribution. Our BI implementation uses the logit-level signal directly via the softmaxed Yes/No probabilities, sidestepping this disagreement.

G Scenario Generation Prompt

The exact system and user prompts used to instruct the language model for synthetic scenario generation are detailed in Figure 8. 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 9.

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 8: 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 9: EmoScene Evaluation Prompt.

Zero-Shot 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.

You must answer strictly by wrapping your Yes or No prediction inside `<answer></answer>` tags.

Do not provide any reasoning, explanation, or extra text. Output ONLY the tags and your answer.

User Message

Scenario:
text
Subject:
subject
Question:
Does subject feel emotion?
Note:
Based on the above-provided information, answer the question strictly with Yes or No wrapped inside `<answer></answer>` tags. Do not provide any explanation, reasoning, or additional text.

Figure 10: Zero-Shot Evaluation Prompt.