Embedding Analogies for Evaluating Emotion in LLM-Generated Utterances

Sadegh Jafari, Els Lefever, Véronique Hoste
LT3, Language and Translation Technology Team
Ghent University, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{sadegh.jafari, els.lefever, veronique.hoste}@ugent.be

Abstract. Emotion plays a vital role in human communication, shaping not only language but also vocal tone, facial expression, and body posture. In the context of emotionally expressive text generation, the lack of reliable evaluation metrics still remains a key challenge. This paper introduces a two-step evaluation framework using embedding analogy-based metrics to assess the emotional expressiveness of large language models (LLMs). In the first step, we evaluate the model's ability to neutralize emotional content from a given text while preserving its semantic meaning. In the second step, we test the model's capacity to reinject the intended emotion back into the neutralized text. Our experiments demonstrate that GPT-4.1 outperforms other models in both semantic retention and emotional reconstruction, while llama-3.3-70b-instruct performs best among open-source models. This work lays the foundation for future research on cross-modal affective computing, aiming to build emotionally intelligent agents capable of nuanced and empathetic communication across text, speech, and video.

1 Introduction

Understanding and responding to human emotions is critical for AI systems operating in professional settings, particularly in education, where teachers and students engage in complex emotional interactions. In second language (L2) learning environments, emotionally supportive conversational agents can help teachers foster a safe and motivating atmosphere, alleviating workload and enhancing the student learning experience. Such systems require robust emotional understanding and generation capabilities, which are still underdeveloped due to fundamental challenges in emotion evaluation.

To function effectively in such roles, these systems must be capable of detecting and generating emotional content in real-life, unscripted scenarios. This ability is especially important in high-stakes domains such as healthcare, education, and crisis management. In such contexts, the ability to recognize and respond to genuine human emotions, rather than acted or exaggerated affect, is crucial for building trust, ensuring user well-being, and improving decision-making [30]. Recent efforts to build empathically aware AI systems rely heavily on the generation and interpretation of affective content [38, 26, 36]. However, evaluating the emotional quality of text generated by LLMs remains a fundamental challenge. Current evaluation methods for emotionally expressive text are either expensive, when relying on human annotations, or inadequate in quality and generalization when using existing automatic metrics [9, 17, 44]. This limits their usefulness for scalable and robust assessment of emotion generation models.

In this paper, we address the gap in effective and efficient evaluation of emotional text generation. We propose an embedding-based evaluation pipeline that measures emotional alignment in LLM-generated text without requiring human labels. Our method builds on analogical reasoning in emotion embedding spaces, incorporating steps of emotion neutralization and re-injection to isolate and assess the emotional expressiveness of different LLMs. We apply our evaluation framework to a range of state-of-the-art LLMs and find that GPT-4.1 [29] consistently produces the most emotionally aligned outputs. Among open-source models, LLaMA-3.3-70B-Instruct [13] performs best. Our results demonstrate that embedding-based emotion evaluation is a practical and scalable alternative to existing methods, providing a reliable benchmark for future emotion generation tasks.

2 Related Works

Recent research has explored emotional text generation using LLMs, with a growing interest in evaluating their ability to generate affectively aligned content. In this section, we review state-of-the-art models and evaluation strategies for emotional control in LLMs. Dong et al. [9] introduced continuous emotion vectors to steer LLM outputs toward target affective states. For evaluation, they generated two synthetic datasets using GPT-4o-mini [28] and assessed performance using perplexity, topic adherence (via prompt engineering), emotion probability score (using the zero-shot classifier `facebook/bart-large-mnli` [11]), and an emotion absolute score derived from prompt-based heuristics. However, the prompt-based scores were not evaluated or validated, as it simply relied on the LLM's own response to a scoring prompt. Ishikawa and Yoshino [17] explored emotional expression in LLMs using the circumplex model of affect. They fine-tuned a model on the GoEmotions dataset [8], but the resulting classifier, `sentimentmodel-sample-27go-emotion` [20], achieved 58.9% accuracy, which was deemed insufficient for further use in evaluation. To circumvent the limitations of discrete emotion classification, they instead projected the generated outputs into the arousal–valence space. This alternative approach was implemented to simplify the evaluation task, though it did not aim primar-

ily at improving reliability.

To improve emotional appropriateness in generation, Li et al. [23] proposed emotional chain-of-thought prompting, grounded in Goleman's emotional intelligence framework [12]. They argued that current emotion recognizers are inadequate for evaluation and introduced the Emotional Generation Score (EGS), a prompt-based metric evaluated via GPT-3.5 [27], supplemented by a small-scale human study with three annotators. Wang et al. [44] incorporated commonsense reasoning to enhance empathetic dialogue generation in LLMs. Using the EmpatheticDialogues [35] and Emotional Support Conversation datasets [25], they employed traditional metrics, BLEU [31], ROUGE-L [24], METEOR [3], Distinct-n [22], and CIDEr [43], along with cosine similarity and human evaluation. Human evaluation is valuable but costly and lacks repeatability. A disadvantage of existing automatic metrics is that they often fall short, as lexical overlap between gold-standard and generated emotional expressions remains high regardless of the actual emotional effectiveness. Janssens et al. [18] show that even advanced models struggle to detect miscommunications from facial expressions in natural human-robot dialogue, performing no better than chance. Their findings reveal that users often do not express confusion in visibly detectable ways, highlighting the limitations of current affect recognition tools, which are predominantly trained or fine-tuned on corpora of acted, non-naturalistic emotions and reinforcing the need for more robust, context-aware emotion evaluation strategies.

While these studies propose creative methods for controlling and evaluating emotional content, their reliance on unstable, non-repeatable, or costly approaches leaves the quality assessment of generated emotions an open challenge. Popular metrics like BLEU and ROUGE-L are often inadequate, as lexical overlap between gold-standard and generated emotion expressions remains high regardless of emotional success, rendering these metrics non-discriminative. Prompt-based LLM evaluation (e.g., using GPT-4 to judge GPT-3) also suffers from bias and circularity, especially when assessing commercial or closed-source systems. Lastly, human evaluation, while insightful, is costly and non-repeatable.

Our study addresses these gaps by highlighting the urgent need for robust, repeatable, and model-agnostic emotion evaluation strategies that can generalize across diverse generation setups. Unlike prior works, we initiate a neutralization–re-injection process: first stripping emotions from the original dataset, then prompting models to regenerate emotional variants. This setup enables us to evaluate models based on their capacity to reintroduce appropriate emotions while preserving semantic content.

3 Dataset

In recent years, a growing number of multimodal emotion recognition datasets have been introduced to support research in affective computing and emotionally intelligent systems. Notable among these is the **MELD** dataset [34], which comprises multi-party conversations extracted from the Friends TV show. Although MELD provides valuable dialogic emotion labels, it is based on acted and scripted television content, which may not generalize well to spontaneous emotional behaviors. Similarly, the **IEMOCAP** dataset [5] features dyadic interactions between professional actors performing scripted and semi-scripted scenarios, offering rich annotations across modalities, but again lacks true spontaneity. Similar corpora for Chinese include **EmotionTalk** [40] and **M3ED** [45] introducing large-scale, multimodal emotion data from Chinese TV dramas and controlled dialogues. To address the lack of spontaneous emotion data, the **K-EmoCon** dataset [32] captured natural interactions during real-time debates and provided multi-perspective annotations, including physiological signals, but is limited in scale and does not cover monologue settings. While these datasets advance the field significantly, they still reflect contextual and cultural biases, often rely on acted emotions, and typically do not isolate modalities during annotation, which limits their utility for fine-grained unimodal vs. multimodal analysis.

These limitations, namely, the lack of spontaneous, non-acted emotional expressions, limited diversity of monologue data, and insufficient attention to isolated modality annotations, motivates the use of new datasets designed to better reflect natural emotional communication. The **UniC** [10] dataset is a multimodal emotion dataset comprising 965 video clips sourced from YouTube, selected to capture natural, spontaneous emotional expressions rather than acted performances. The videos primarily include monologues such as book and movie reviews, where a single visible speaker expresses emotions clearly in both speech and facial expressions. The dataset was constructed through a multi-step filtering process using keyword searches, sentiment-based subtitle filtering, and manual validation. Each clip, approximately 10 seconds long, was annotated independently across four modalities: text, audio, silent video, and all modalities combined. Emotion annotations use both categorical and dimensional frameworks. Initially based on 26 categorical emotion labels from Shaver et al. [39], these were reduced to seven emotion clusters (joy, contentment, surprise, confusion, neutral, disappointment, and disgust) via clustering analysis, alongside valence and arousal scores. Figure 1 shows a sample from the UniC dataset.

but like, in a good way, not shockingly bad, shockingly absurd. I experienced a really visceral and physical response to it. Like, it was making my whole body tense and cringe by how wild it is, and also quite disgusting at times.

Figure 1. A video clip example from the UniC dataset

For our experiments, we focused on the text modality as a stepping stone to multimodal emotion expression generation in follow-up research. Noteworthy to mention is that for this text modality, the inter-annotator agreement (IAA) was highest, reaching a Fleiss' kappa of 0.47 after annotator training and emotion clustering. Among the different labeled emotions, emotions such as *confusion* and *surprise* were less reliably detected from text alone, highlighting the added value of multimodal signals. We evaluated the text modality of the UniC dataset using several baseline models, for which we used 100% of the dataset for testing. Due to the limited size of the dataset, we employed 5-fold cross-validation for training and evaluating our custom model.

As shown in Table 1, our model does not achieve the highest performance across any metric. Among the evaluated models, *michelle-jieli* and *j-hartmann* are fine-tuned emotion classifiers based on the DistilRoBERTa-base [37] architecture. The *bart-large-mnli* model, a zero-shot classifier built on the BART-large [21] transformer, is used without fine-tuning. The *gpt-4o-mini* model, on the other hand, is an LLM that predicts emotions through prompt-based reasoning. No-

Table 1. Evaluation results of various models for emotion recognition on the UniC dataset. Emotion recognition on real conversational data is inherently challenging; for instance, the best model here (gpt-4o-mini) achieves an F1-score of only 35.79%, substantially lower than the 60.25% commonly seen on acted datasets like MELD.

Method	Accuracy	Precision	Recall	F1
gpt-4o-mini [28]	0.4046	0.3765	**0.4496**	**0.3579**
michellejieli [14]	**0.4492**	**0.4205**	0.3095	0.3301
j-hartmann [15]	0.3880	0.4098	0.3724	0.3268
bart-large-mnli [11]	0.1639	0.3084	0.3470	0.1585
Our model	0.3885	0.3204	0.3068	0.3134

tably, *michellejieli* achieves the highest accuracy (0.4492) and precision (0.4205), while *gpt-4o-mini* performs best in recall (0.4496) and F1 score (0.3579). Our approach, which combines `BAAI-bge-m3` embeddings [7] with a tuned Random Forest classifier [4], yields moderate but consistent results across all metrics because it just trained on UniC dataset(772 training samples). The classifier's hyperparameters are shown in Table 2. It is important to highlight that these relatively low performance scores are primarily due to the nature of the dataset, which consists of natural, non-acted emotional expressions.

Table 2. Hyperparameters used for the Random Forest classifier with BAAI-bge-m3 embeddings.

Hyperparameter	Value
n_estimators	316
max_depth	488
min_samples_split	50
criterion	gini
class_weight	balanced_subsample

4 Methodology

Emotional text generation and its evaluation have been less explored through analogical methods, despite their proven utility in measuring structured semantic relations. Chen et al. [6] systematically analyzed vector-based analogies, confirming their reliability in capturing such relations, and Zhu and De Melo [46] extended analogical reasoning to contextualized sentence embeddings, showing that some models preserve analogical structures at the sentence level. To our knowledge, no prior work has applied analogy-based evaluation specifically to the assessment of emotional expressiveness in generated text.

Building on these insights, our methodology employs analogy-based evaluation to quantify the emotional expressiveness of LLMs. To rigorously isolate the model's generative capabilities, we begin by neutralizing the emotional content of each ground-truth (GS) text in our dataset using an LLM. Following neutralization, the model is prompted to regenerate the emotional version of each text. The neutralization step is crucial: by comparing the regenerated emotional outputs with the original GS emotions, we ensure that any observed affective content arises from the model's learned patterns rather than residual cues in the input. Finally, we compute embedding-based similarity and analogy metrics between the GS and regenerated texts, enabling quantification of both semantic fidelity and emotional alignment.

4.1 Embedding Evaluation Metric

Before focusing on the embedding evaluation metric, we should mention that all embeddings were calculated using the BGE-M3 [7] language model, and the 2D space was generated using the t-SNE [42] method applied to the BGE-M3 embedding space.

In our embedding evaluation metric, we draw inspiration from the well-known linguistic analogy: "**king - man + woman ≈ queen**". This example illustrates how word embeddings can capture semantic relationships through vector arithmetic [2]. By representing words as vectors in a high-dimensional space, operations such as subtraction and addition can reveal underlying relationships, such as gender or emotional tone. This property enables the assessment of emotional quality in generated text by analyzing geometric relationships between word vectors, offering a quantitative measure of emotional expressiveness in language models. Figure 2 visually demonstrates this concept, showing how vector operations can encode semantic relationships in the embedding space.

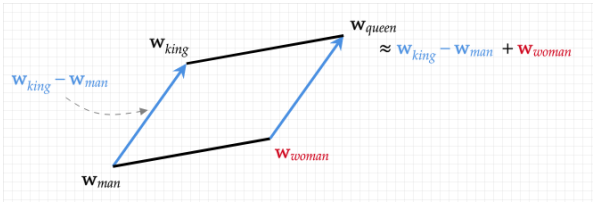

Figure 2. A hand-drawn illustrative example of the 'King and Queen' analogy in an ideal embedding space.

In Figure 2, the length and direction of the vectors $W(\text{king}) - W(\text{queen})$ and $W(\text{man}) - W(\text{woman})$ appear to be the same. However, this does not reflect reality. In a realistic scenario, we would expect the vector $W(\text{king}) - W(\text{man}) + W(\text{woman})$ to be close to $W(\text{queen})$. Using BGE-M3, we calculated the embeddings for *queen*, *king*, *man*, and *woman*. As shown in Figure 3, the expression $W(\text{king}) - W(\text{man}) + W(\text{woman})$ is not exactly equal to $W(\text{queen})$, but it is close.

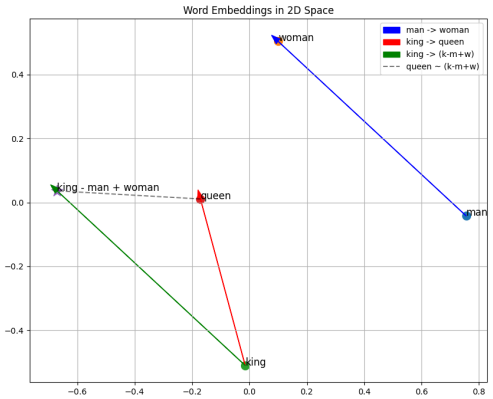

Figure 3. 2D visualization of word embeddings using t-SNE. The vector $W(\text{king}) - W(\text{man}) + W(\text{woman})$ lies near $W(\text{queen})$, illustrating a plausible semantic relationship in the reduced space.

4.1.1 Cosine Similarity vs. Manhattan Distance

A common method for measuring similarity between two vectors is the *cosine similarity* metric. However, in analogy tasks, this method has a major limitation: the results can vary based on the operation

order. Consider the analogy: *king is to queen as man is to woman*. The similarity and distance scores for various formulations are summarized in Table 3.

Table 3. Comparison of cosine similarity and Manhattan distance for different analogy vector operations.

Pair	Cosine Similarity	Manhattan Distance
(king, queen)	0.7119	19.2969
(man, woman)	0.6343	21.7031
(man, woman − queen + king)	0.7188	19.5469
(woman, man − king + queen)	0.7461	19.5469
(king, queen − woman + man)	0.7368	19.5469
(queen, king − man + woman)	0.7759	19.5469

As shown in Table 3, different operation orders produce varying cosine similarity scores, revealing inconsistency in the evaluation of the cosine-based analogy. In contrast, the *Manhattan distance* produces stable results across all permutations, indicating its robustness for analogy reasoning tasks. Due to its consistent behavior, we further used the Manhattan distance for the analogy evaluation in our experiments.

4.1.2 *Real Emotional Example*

To better understand the role of emotional analogy in our framework, we illustrate a representative example from our experiments. The goal is to analyze how vector arithmetic in the embedding space can capture shifts in emotional expression between sentences. Figure 4 visualizes this example. The corresponding text for each variable in the figure is as follows:

- **joy** = "joy"
- **neutral** = "neutral"
- **neutral_sent** = "It's my first day as a student"
- **joy_sent** = "I'm so happy, it's my first day as a student!"

In the Figure 4, we observe that the distance between the `neutral` and `joy` emotion embeddings is relatively large. This discrepancy poses a challenge for emotional analogy, as the semantic distance between the two sentence embeddings (`neutral_sent` and `joy_sent`) is significantly smaller than the distance between their corresponding emotion labels. To mitigate this, we construct an analogy vector using the following equation:

$$\text{analogy\_vector} = \text{neutral} - \text{neutral\_sent} + \text{joy\_sent} \quad (1)$$

This vector is then compared with the `joy` embedding. As shown in the Figure 4, the analogy vector lies closer to `joy` than `neutral`, indicating that the analogy operation effectively captures the intended emotional shift.

Recognizing emotions in real user utterances is particularly challenging due to their subtle and nuanced nature. As shown in Table 1, the best model achieves an F1-score of only 35.79%, significantly lower than the 60.25% observed on acted datasets like MELD [34]. To further investigate this phenomenon, we visualized the semantic structure of emotion representations using the BGE-M3 embedding model. Figure 5 shows a 2D projection of both the emotion label embeddings and the average embeddings of real user utterances associated with each emotion. In this plot, each circle represents an emotion label (e.g., *joy*, *disgust*, *neutral*), and each square denotes the average embedding of utterances tagged with that emotion. Two sets of relationships are highlighted:

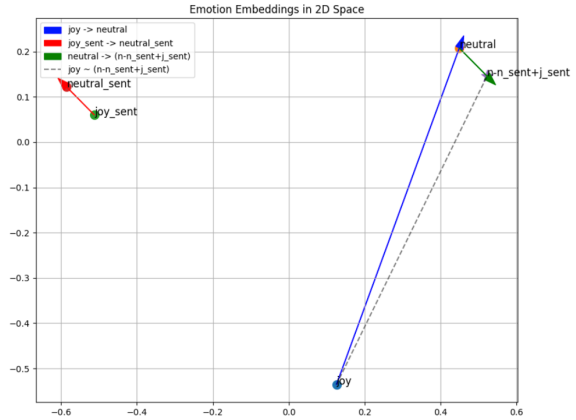

Figure 4. Visualization of emotional analogy in the embedding space. The plot shows the positions of the `neutral`, `joy`, and sentence embeddings (`neutral_sent` and `joy_sent`) in the embedding space.

- **Red lines** connect the embedding of the label *neutral* to other emotion labels.
- **Green lines** connect the average embedding of utterances labeled as *neutral* to the average embeddings of utterances for other emotions.

The figure reveals that while the emotion labels are well-separated in the embedding space, indicating clear semantic distinctions, the average embeddings of real user expressions are clustered more closely together, especially around the *neutral* region. This supports the idea that emotional language in real interactions is often more subtle, making automatic emotion detection more challenging in natural contexts.

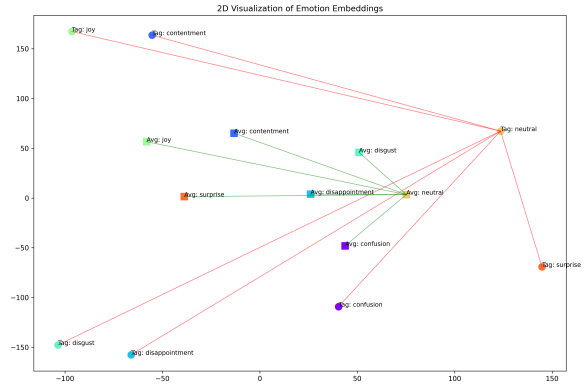

Figure 5. 2D projection of emotion embeddings using t-SNE. Circles represent the embeddings of emotion labels (e.g., *joy*, *neutral*), while squares represent the average embeddings of real user utterances associated with each emotion. Red lines show distances from the *neutral* label to other labels, and green lines show distances from the average *neutral* embedding to other emotion averages.

To better understand how emotional meaning is encoded in sentence embeddings, we explore the relationship between *labeled* and *unlabeled* emotional expressions. Specifically, we aim to approximate the embedding of an emotionally tagged utterance using its neutral version and the emotional shift encoded in a semantically aligned sentence. Here, *labeled emotion* refers to utterances that include direct emotion labels from the gold-standard data in UniC dataset (e.g., "I'm so happy, it's my first day as a student! (joy emotion)"), while

unlabeled emotion refers to emotionally expressive content without such tags but still conveying affect (e.g., "I'm so happy, it's my first day as a student!"). *Neutral* versions are affectively flat and omit emotional cues.

Our approach applies an analogy-style vector transformation of the form: `neutral − neutral_sent + joy_sent`, where `neutral_sent` and `joy_sent` are the neutral and emotionally expressive versions of the same utterance. This transformation enriches the affective content of the neutral-tagged embedding by injecting the emotional variation from the unlabeled expression, while preserving the shared semantic structure. The goal is to reduce the distance between the synthesized embedding and its explicitly emotional counterpart, effectively revealing how emotional meaning can be reconstructed through compositional operations. Figure 6 visualizes this transformation. The green arrow illustrates the analogy vector described above, and the dashed lines indicate the proximity between the predicted and actual emotion embeddings. The text associated with each vector in the figure is as follows:

- **joy** = "I'm so happy, it's my first day as a student! (joy emotion)"
- **neutral** = "It's my first day as a student (neutral emotion)"
- **neutral_sent** = "It's my first day as a student"
- **joy_sent** = "I'm so happy, it's my first day as a student!"

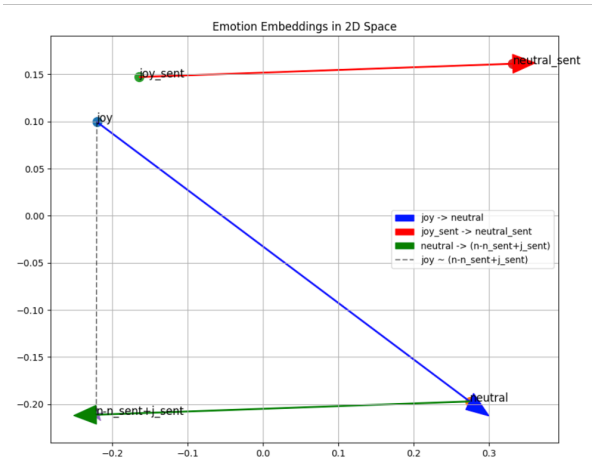

Figure 6. Visualization of emotion-related embedding transformations. The green arrow represents the analogy vector `neutral - neutral_sent + joy_sent`, used to approximate the emotional embedding. Dashed lines indicate proximity between the original and approximated vectors.

4.2 *Emotion Embedding Extraction Using Prompted Text Templates*

As discussed in Section 4.1.1, we use Manhattan distance as our similarity metric due to its sensitivity to subtle semantic variations in the embedding space. This metric is essential for evaluating how emotional content can be manipulated while preserving the original meaning. Our goal is to identify the most effective prompt template for extracting emotion embeddings from textual descriptions. These embeddings, denoted as E_o, E_n, and E_t, represent the original, neutral, and target emotional states, respectively. By inserting emotion-related phrases into structured prompt templates, we derive these embeddings for use in analogy-based transformations. The transformation involves two steps: neutralization and emotionalization. Let S_o, S_n, and S_t be the sentence embeddings for the original, neutral, and target emotional versions of the same sentence. Let $\mathrm{MD}(A, B)$ denote the Manhattan distance between embeddings A and B. The neutralization step tests whether removing the original emotion and inserting the neutral emotion embedding moves it closer to S_n:

$$\mathrm{MD}(S_o, S_n) \geq \mathrm{MD}(S_o - E_o + E_n, S_n) \tag{2}$$

The emotionalization step checks whether inserting the target emotion into the neutral embedding moves it closer to S_t:

$$\mathrm{MD}(S_n, S_t) \geq \mathrm{MD}(S_n - E_n + E_t, S_t) \tag{3}$$

These conditions validate whether modifying sentence embeddings via emotional vectors steers them toward the intended emotional states. A transformation is deemed successful when both inequalities are satisfied.

System Prompt 1: Text Neutralization

Your task is to neutralize the text by removing emotional expressions.
The text is a transcription of a video.
The text may contain emotional expressions.
The text should be neutral and not contain any emotional expressions.
The text should be in the same language, format, style, tone, and context as the input text.
Please try to change the text as little as possible.
Please neutralize the following text: {text}
The original emotion of the text is: {emotion}
Please make sure to remove all emotional expressions from the text.

System Prompt 2: Emotional Text Generation

Your task is to make the text more emotional by adding emotional expressions.
The text is a transcription of a video.
The text should be in the same language, format, style, tone, and context as the input text.
Please try to change the text as little as possible.
Don't mention the emotion in the text directly.
Please add emotional expressions to the following text: {text}
The current emotion of the text is: neutral.
The target emotion of the text should be: {emotion}.

To identify the most effective prompt template for extracting emotion embeddings, we evaluated five candidate prompt formulations across several LLMs. These templates vary in how they contextualize emotion labels with respect to the text, ranging from labeled structures (e.g., "joy emotion: {text}") to minimal expressions (e.g., just "joy").

Our evaluation follows a two-step analogy-based framework. In the neutralization step, we generated neutral versions of emotional sentences using each LLM with a fixed system instruction based on **System Prompt 1**. To extract the emotion embeddings E_o and E_n used in Equation 2, we tested the five emotion prompt templates by plugging them into an embedding encoder. In the emotionalization step, we used **System Prompt 2** to generate emotionalized sentences from neutral ones and evaluated how well each emotion prompt template performed using Equation 3 with the target emotion embedding E_t. The followings are the details about the emotion embedding prompts:

- **Prompt 1**: {emotion} emotion: {text}
- **Prompt 2**: This content has {emotion} emotion: {text}
- **Prompt 3**: {text} ({emotion} emotion)
- **Prompt 4**: {emotion}
- **Prompt 5**: This content has {emotion} emotion

As shown in Table 4, we identify the best-performing emotion prompt template for each step of the evaluation. Using the entire

Table 4. Performance of different LLMs and prompt templates in analogy-based emotion embedding evaluation. Each cell shows the percentage of samples that satisfied the analogy inequality in the neutralization (left) and emotionalization (right) steps.

Model	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5
gemma-3-1b-it [41]	98.96 / 100.00	98.76 / 100.00	99.79 / 96.00	33.82 / 28.00	33.82 / 24.00
llama-3.1-8b-instruct [13]	98.96 / 100.00	97.51 / 99.90	100.00 / 100.00	37.86 / 53.22	36.62 / 43.15
mistral-nemo-12b-instruct [1]	98.96 / 99.90	95.85 / 99.90	99.69 / 100.00	38.49 / 58.30	37.55 / 45.12
llama-3.3-70b-instruct [13]	99.38 / 99.48	98.96 / 99.38	100.00 / 100.00	37.14 / 57.47	36.20 / 45.64
gpt-4.1 [29]	99.07 / 98.96	96.58 / 95.95	99.59 / 99.48	**48.55 / 78.42**	**41.80 / 59.44**
gpt-4o-mini [28]	**99.48 / 99.90**	**98.34 / 99.90**	**99.90 / 100.00**	34.75 / 76.04	34.65 / 58.20
Average	99.13 / **99.71**	97.67 / 99.17	**99.83** / 99.25	38.44 / 58.57	36.77 / 45.93

text-only UniC dataset for evaluation, we conduct experiments on two tasks: neutralization and emotionalization. For neutralization, Prompt 3 achieves the highest analogy satisfaction rates across most models. For emotionalization, Prompt 1 performs best, indicating its effectiveness in reintroducing emotional content through embedding manipulation. These findings suggest that different prompt styles may be optimal for extracting emotion embeddings depending on the specific transformation goal.

5 Analysis and Results

Having decided on the **Manhattan Distance** to compare the embedding vectors (Section 4.1.1) and on using distinct prompt templates for extracting emotion embeddings depending on the transformation stage (Section 4.2), we set up an experiment in which our goal was to evaluate the impact of emotion generation by comparing the original emotional data with the emotionally re-generated text. Specifically, we used **Prompt 3** for the neutralization stage and **Prompt 1** for the emotionalization stage, as each achieved the highest analogy satisfaction rates for their respective tasks across most models. To enable a broad comparison, we evaluated a range of LLMs, including open-source models such as **Gemma [41]**, **LLaMA-3 [13]**, and **Mistral-NeMo [1]**, as well as commercial models like **GPT-4.1 [29]** and **GPT-4o-Mini [28]** from OpenAI. This mix allowed us to assess the effectiveness of emotion embedding manipulation across both accessible, community-driven models and state-of-the-art proprietary systems. All evaluations are on the UniC dataset's text modality. The process consisted of the following two main steps:

5.1 Neutralization

We used an LLM to neutralize the emotional content of the original text samples. This step aimed to remove any labeled or unlabeled emotional signals, resulting in emotionally flat, semantically preserved text. In this experiment we used **System Prompt 1**. The following formulas were used in the tables to evaluate the performance of different models. In these equations, A denotes the analogy vector.

$$A = E_n - S_n + S_o$$
$$R1_c = \cos(E_o, E_n), \quad R2_c = \cos(E_o, A)$$
$$R1_m = \|E_o - E_n\|_1, \quad R2_m = \|E_o - A\|_1$$

As shown in Table 5, we evaluate each LLM's ability to perform emotional neutralization based on how well the transformed sentence embedding aligns with the original emotional context vector. The evaluation uses both cosine similarity and Manhattan distance to capture different aspects of embedding relationships. In both cosine similarity and Manhattan distance metrics, **GPT-4.1** demonstrates the most controlled and semantically faithful emotion neutralization among all evaluated models. While **llama-3.3-70b-instruct** achieves the highest post-neutralization cosine similarity ($R2_{\text{c}} = 0.9746$) and lowest Manhattan distance ($R2_{\text{m}} = 5.33$), **GPT-4.1** yields the smallest changes in both cosine ($\Delta R_{\text{c}} = 0.0410$) and Manhattan metrics ($\Delta R_{\text{m}} = -3.297$), indicating minimal semantic distortion during transformation. This suggests that **GPT-4.1** preserves original sentence meaning more effectively while removing emotional content. Overall, while high-capacity open-source models like **mistral-nemo-12b-instruct** are increasingly competitive, commercial models such as **GPT-4.1** still lead in performance when performing nuanced tasks like emotion neutralization.

Table 5. Evaluation results of LLMs for emotion neutralization. R1/R2: before/after transformation. Subscripts $_{\text{m}}$/$_{\text{c}}$: Manhattan/Cosine.

Model	$R1_{\text{m}}$	$R2_{\text{m}}$	ΔR_{m}	$R1_{\text{c}}$	$R2_{\text{c}}$	ΔR_{c}
gemma-3-1b-it [41]	16.160	7.484	-8.670	0.7734	0.9536	0.1802
llama-3.1-8b-instruct [13]	10.180	5.965	-4.215	0.9106	0.9683	0.0576
mistral-nemo-12b-instruct [1]	9.640	5.785	-3.855	0.9185	0.9690	0.0508
llama-3.3-70b-instruct [13]	13.414	**5.330**	-8.086	0.8370	**0.9746**	0.1377
gpt-4.1 [29]	**8.920**	5.625	**-3.297**	**0.9290**	0.9700	**0.0410**
gpt-4o-mini [28]	11.990	6.450	-5.543	0.8790	0.9644	0.0855

Table 6 shows the emotion-wise results for **GPT-4.1** neutralization. The neutral category exhibits the smallest changes in both Manhattan distance and cosine similarity, reflecting that converting originally neutral utterances to neutral is inherently easier. In contrast, other emotions require more substantial transformations to remove affective content while preserving meaning, resulting in larger embedding changes.

Table 6. Emotion-wise evaluation results of GPT-4.1 [29] for emotion neutralization. R1/R2: before/after transformation. Subscripts $_{\text{m}}$/$_{\text{c}}$: Manhattan/Cosine.

Emotion	$R1_{\text{m}}$	$R2_{\text{m}}$	ΔR_{m}	$R1_{\text{c}}$	$R2_{\text{c}}$	ΔR_{c}
confusion	9.650	7.254	-2.395	0.9224	0.9565	0.0342
joy	10.010	6.453	-3.555	0.9175	0.9663	0.0488
neutral	6.266	2.521	-3.744	0.9630	0.9946	0.0317
disgust	11.760	8.164	-3.594	0.8857	0.9460	0.0601
contentment	9.870	7.016	-2.852	0.9190	0.9590	0.0400
disappointment	10.410	7.414	-2.992	0.9097	0.9560	0.0464
surprise	10.600	7.227	-3.375	0.9062	0.9580	0.0518

To evaluate the consistency between different similarity metrics, we computed the Pearson correlation [33] between the Manhattan distance and Cosine similarity values for both R2 and R1 scores. As shown in Figure 7, there is a very strong negative correlation between the two measures for both R2 ($r = -0.99365$) and R1 ($r = -0.99501$). These results indicate that as the Manhattan distance increases, the Cosine similarity decreases almost linearly, suggesting that both metrics are capturing highly similar trends in eval-

uating the transcripts, albeit in opposite directions due to their different mathematical formulations.

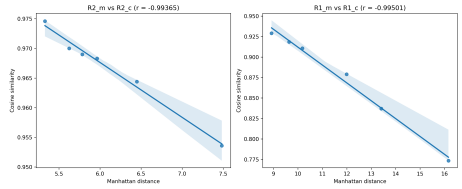

Figure 7. Scatter plots showing the relationship between Manhattan distance and Cosine similarity for R2 (left) and R1 (right) metrics. Each point represents the result of a different LLM on the UniC dataset. A regression line is included in each subplot to visualize the correlation.

5.2 Emotion Injection

In the emotion injection phase, we used an LLM to reintroduce a target emotion (e.g., joy) into the neutralized text. To guide this process, we prompted the model using system-level instructions and emotion-specific cues. The goal was to generate emotionally expressive text that closely resembles the original emotional content while preserving the core semantics of the neutralized version. In this experiment we used **System Prompt 2**. To perform the re-injection, we compute the analogy vector A based on the following relationship:

$$A = E_t - S_t + S_n$$
$$R1_c = \cos(E_n, E_t), \qquad R1_m = \|E_n - E_t\|_1$$
$$R2_c = \cos(E_n, A), \qquad R2_m = \|E_n - A\|_1$$
$$R3_c = \cos(S_o, S_t), \qquad R3_m = \|S_o - S_t\|_1$$

Table 7 shows the performance of various LLMs in emotion injection. GPT-4.1 achieves the best results overall, with the lowest distances and highest cosine similarities (e.g., $R3_m$ = 9.99, $R3_c$ = 0.9130), indicating strong emotional alignment and reinjection ability. In contrast, gemma-3-1b-it performs the weakest, especially in re-injection quality ($R3_c$ = 0.6990). While commercial models like GPT-4.1 and GPT-4o-mini outperform others due to superior training and architecture, larger open-source models such as LLaMA-3.3-70B and Mistral-Nemo-12B show competitive performance, suggesting that open models can still be effective in emotion-aware tasks.

Table 7. Evaluation results of LLMs for emotion injection. R1: neutral–target emotion, R2: analogy vector, R3: original–re-injected sentence. Subscripts $_m$/$_c$: Manhattan/Cosine.

Model	$R1_m$	$R2_m$	ΔR_m	$R3_m$	$R1_c$	$R2_c$	ΔR_c	$R3_c$
gemma-3-1b-it [41]	15.560	9.190	-6.375	18.890	0.7930	0.9287	0.1357	0.6990
llama-3.1-8b-instruct [13]	12.125	6.824	-5.301	13.830	0.8740	0.9590	0.0850	0.8374
mistral-nemo-12b-instruct [1]	12.164	6.625	-5.539	13.390	0.8740	0.9610	0.0869	0.8486
llama-3.3-70b-instruct [13]	12.950	5.824	-7.126	13.290	0.8486	**0.9700**	0.1216	0.8496
gpt-4.1 [29]	**8.086**	**5.470**	**-2.617**	**9.990**	**0.9310**	0.9680	**0.0366**	**0.9130**
gpt-4o-mini [28]	10.390	6.016	-4.375	13.414	0.9077	0.9670	0.0591	0.8500

Table 8 reports emotion-wise performance of GPT-4.1 on the emotion injection task. The model performs consistently across all emotions, with strong alignment scores (e.g., $R3_c > 0.88$) and small Manhattan distances. Notably, the *neutral* class achieves the best results ($R3_c$ = 0.9280, $R1_m$ = 2.30), which is expected since the model is converting a neutralized utterance back to a neutral form, making the reinjection task considerably easier in this case.

To further validate the metric alignment, we conducted a correlation analysis between Manhattan distance and cosine similarity across the three relations (R1, R2, R3). All pairs exhibit strong negative correlations below -0.9819, confirming the inverse relationship between the two metrics (see Figure 8). These results confirm that increased directional similarity corresponds closely with reduced embedding distance, validating the use of both metrics to quantify emotional fidelity in the reinjection process.

Table 8. Emotion-wise evaluation results for GPT-4.1 in emotion injection. R1: neutral–target emotion, R2: analogy vector, R3: original–re-injected sentence. Subscripts $_m$/$_c$: Manhattan/Cosine.

Emotion	$R1_m$	$R2_m$	ΔR_m	$R3_m$	$R1_c$	$R2_c$	ΔR_c	$R3_c$
surprise	11.766	7.688	-4.078	11.410	0.8870	0.9517	0.0645	0.8920
disgust	12.200	8.480	-3.720	11.610	0.8780	0.9414	0.0635	0.8860
confusion	10.734	7.977	-2.757	10.160	0.9060	0.9480	0.0425	0.9140
contentment	10.490	7.598	-2.892	9.970	0.9097	0.9517	0.0420	0.9160
neutral	2.303	1.225	-1.078	8.920	0.9920	0.9980	0.0059	0.9280
joy	11.000	7.242	-3.758	10.164	0.9010	0.9575	0.0566	0.9126
disappointment	11.230	7.477	-3.753	10.990	0.8975	0.9546	0.0571	0.8975

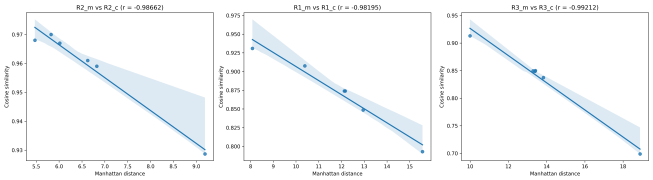

Figure 8. Scatter plots showing correlation between Manhattan distance and Cosine similarity for R1, R2, and R3 embedding relations. Each point represents the result of a different LLM on the UniC dataset. A regression line is included in each subplot to visualize the correlation.

6 Conclusion and Future Works

In this study, we explored the capability of LLMs to manipulate and generate emotionally expressive text through a two-step process: emotional neutralization followed by targeted emotion injection. Using embedding-based similarity metrics such as Manhattan distance and cosine similarity, we quantitatively evaluated the extent to which LLMs can remove and reintroduce specific emotions while preserving the semantic core of the original text. Our findings indicate that GPT-4.1, a commercial model, consistently outperforms other models in maintaining semantic fidelity and accurately reconstructing emotional nuances. Among open-source models, LLaMA-3.3-70B-Instruct demonstrates the best performance in our experiments, making it a strong candidate for accessible, open research in emotion-aware language generation. These results underscore the effectiveness of large-scale LLMs for emotion control and expression in text and provide a foundation for broader affective computing applications. Although our current focus is on the text modality, the proposed framework is explicitly designed to extend to speech and visual channels by leveraging shared embedding spaces. In particular, recent work by Jha et al. [19], which builds upon the *Platonic Representation Hypothesis* introduced by Huh et al. [16], demonstrates that as neural networks scale, internal representations across modalities converge toward a shared statistical model of reality. This convergence enables cross-modal affective analysis without requiring paired training data, providing a strong theoretical and practical foundation for our future work.

In addition to aligning emotional content across text, speech, and visual modalities within unified embedding spaces, our future efforts will also involve improved prompt engineering and the development of more expressive embedding models to enhance emotional transformation capabilities. As a concrete application, we aim to develop a multimodal empathetic conversational agent for second language

(L2) learning. By engaging students in emotionally supportive interactions, such agents can foster psychologically safe and motivating learning environments while assisting teachers in managing affective dynamics in the classroom.

7 Acknowledgments

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (FAIR) (174K02325).

References

[1] M. AI. Mistral nemo. https://mistral.ai/news/mistral-nemo/, 2024. Accessed: September 23, 2024.

[2] C. Allen and T. Hospedales. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR, 2019.

[3] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[4] L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[6] D. Chen, J. C. Peterson, and T. L. Griffiths. Evaluating vector-space models of analogy. *arXiv preprint arXiv:1705.04416*, 2017.

[7] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[8] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[9] Y. Dong, L. Jin, Y. Yang, B. Lu, J. Yang, and Z. Liu. Controllable emotion generation with emotion vectors. *arXiv preprint arXiv:2502.04075*, 2025.

[10] Q. Du, S. Labat, T. Demeester, and V. Hoste. Unic: A dataset for emotion analysis of videos with multimodal and unimodal labels. *Language Resources and Evaluation*, pages 1–36, 2025.

[11] Facebook-AI. facebook/bart-large-mnli. https://huggingface.co/facebook/bart-large-mnli, 2020. Accessed: 2025-06-12.

[12] D. Goleman. *Emotional intelligence: Why it can matter more than IQ*. Bantam, 2005.

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[14] J. Hartmann. Fine-tuned DistilRoBERTa-base for Emotion Classification. https://huggingface.co/michellejieli/emotion_text_classifier/, 2022. Accessed: 2025-05-07.

[15] J. Hartmann. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[16] M. Huh, B. Cheung, T. Wang, and P. Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[17] S.-n. Ishikawa and A. Yoshino. Ai with emotions: Exploring emotional expressions in large language models. *arXiv preprint arXiv:2504.14706*, 2025.

[18] R. Janssens, J. De Bock, S. Labat, E. Verhelst, V. Hoste, and T. Belpaeme. Why robots are bad at detecting their mistakes: Limitations of miscommunication detection in human-robot dialogue. In *IEEE RO-MAN 2025 conference*, 2025.

[19] R. Jha, C. Zhang, V. Shmatikov, and J. X. Morris. Harnessing the universal geometry of embeddings. *arXiv preprint arXiv:2505.12540*, 2025.

[20] J. Khan. sentiment-model-sample-27goemotion. https://huggingface.co/jkhan447/sentiment-model-sample-27go-emotion, 2022. Accessed: February 11, 2024.

[21] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019. URL http://arxiv.org/abs/1910.13461.

[22] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

[23] Z. Li, G. Chen, R. Shao, Y. Xie, D. Jiang, and L. Nie. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*, 2024.

[24] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[25] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*, 2021.

[26] A. Mari, A. Mandelli, and R. Algesheimer. Empathic voice assistants: Enhancing consumer responses in voice commerce. *Journal of Business Research*, 175:114566, 2024.

[27] OpenAI. Introducing gpt-3.5. https://platform.openai.com/docs/models/gpt-3.5-turbo, 2025. Accessed: 2025-06-04.

[28] OpenAI. Introducing gpt-4o-mini. https://platform.openai.com/docs/models/gpt-4o-mini, 2025. Accessed: 2025-06-04.

[29] OpenAI. Introducing gpt-4.1. https://platform.openai.com/docs/models/gpt-4.1, 2025. Accessed: 2025-06-04.

[30] E. Ortega-Ochoa, M. Arguedas, and T. Daradoumis. Empathic pedagogical conversational agents: a systematic literature review. *British Journal of Educational Technology*, 55(3):886–909, 2024.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[32] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020.

[33] K. Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

[34] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[35] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018.

[36] M. Rubin, H. Arnon, J. D. Huppert, A. Perry, et al. Considering the role of human empathy in ai-driven therapy. *JMIR Mental Health*, 11(1): e56529, 2024.

[37] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[38] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, and D. Meyer. Empathic conversational agent platform designs and their evaluation in the context of mental health: Systematic review. *JMIR Mental Health*, 11:e58974, 2024.

[39] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

[40] H. Sun, X. Wang, J. Zhao, S. Zhao, J. Zhou, H. Wang, J. He, A. Kong, X. Yang, Y. Wang, et al. Emotiontalk: An interactive chinese multimodal emotion dataset with rich annotations. *arXiv preprint arXiv:2505.23018*, 2025.

[41] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[42] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[44] L. Wang, J. Li, C. Yang, Z. Lin, H. Tang, H. Liu, Y. Cao, J. Wang, and W. Wang. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 123–140, 2025.

[45] J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, and H. Li. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*, 2022.

[46] X. Zhu and G. De Melo. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th international conference on computational linguistics*, pages 3389–3400, 2020.