Disentangling Emotion Understanding and Generation in Large Language Models

Sadegh Jafari, Els Lefever and 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

Large language models (LLMs) have demonstrated strong performance on emotion understanding tasks, yet their ability to faithfully generate emotionally aligned text remains less well understood. We propose a semantic evaluation framework that jointly assesses emotion understanding, emotion generation, and internal consistency, using a VAE-based emotion cost matrix that captures graded semantic similarity between emotion categories. Our framework introduces four complementary metrics that disentangle baseline understanding, human-perceived emotion in generated text, generation quality, and model consistency. Experimental results show that while understanding and consistency scores are highly correlated, emotion generation exhibits substantially weaker correlations with these metrics. These findings motivate the development of specialized evaluation protocols that independently measure emotional understanding and generation, enabling more reliable assessments of LLM emotional intelligence.

1 Introduction

The advancement of LLMs has brought their emotional intelligence to the forefront of AI research, as it is a critical component for effective and meaningful human-computer interaction (Wang et al., 2023; Li et al., 2023; Hu et al., 2025). The ability of these models to comprehend and express emotions is essential for their integration into real-world applications that require social awareness and empathy (Sabour et al., 2024; Ishikawa and Yoshino, 2025). Within this domain, two fundamental yet distinct capabilities are often discussed: **emotion understanding**, the capacity to accurately identify emotions from text, and **emotion generation**, the ability to produce text that conveys a specific, appropriate emotion (Li et al., 2024; Liu et al., 2025). While both are integral to emotional intelligence, a key challenge lies in evaluating them, as current benchmarks often focus more on recognition than other essential emotional skills (Sabour et al., 2024; Zhang et al., 2025). This may lead to an incomplete assessment of an LLM's true emotional capabilities.

This paper argues that evaluating the emotion generation capability of LLMs is fundamentally different from evaluating their emotion understanding abilities. To investigate this distinction, we conduct experiments with five distinct LLMs (see Table 1). Our methodology centers on the UniC dataset (Du et al., 2025), which consists of expert-labeled conversational texts annotated with emotions. We adopt a two-phase process involving emotion neutralization followed by emotion re-injection, prompting the models to generate text samples that align with the original emotion labels provided in UniC. To establish a reliable ground truth for the generated texts, all generated samples were annotated by human evaluators.

By analyzing the performance of the five LLMs across emotion understanding (OERS, GERS), emotion generation (EGS), and internal consistency (ECS) metrics (see Section 4), this study aims to demonstrate the divergence between their ability to understand emotion and their ability to generate emotionally aligned text. Our central hypothesis is that high performance on understanding-based metrics does not necessarily translate to high performance on generation-based metrics. Our findings reveal a stark contrast in the relationships between these scores: the average correlation among the understanding (OERS, GERS) and consistency (ECS) metrics is 95%, indicating that they measure similar constructs. However, the average correlation between the Generation Score (EGS) and these other three metrics (OERS, GERS, and ECS) is only 73%. This discrepancy strongly suggests that the other metrics are not good representatives for evaluating an LLM's emotion generation capability. These findings underscore the need for specialized

March 29, 2026 ©2026 Association for Computational Linguistics

evaluation protocols that can independently assess both understanding and generation, thereby providing a more accurate and nuanced picture of an LLM's emotional intelligence and guiding the development of more genuinely empathetic artificial agents (Li et al., 2024; Liu et al., 2025).

2 Related Works

Emotional Intelligence (EI) is broadly defined as the ability to manage one's own emotions and to understand the emotions of others (Sabour et al., 2024). As LLMs become more integrated into human-centric applications, imbuing them with EI has become a critical area of research (Wang et al., 2023; Raj, 2024). In the context of LLMs, this dual capability translates into two distinct research strands: emotion understanding (perceiving emotion in text) and emotion generation (expressing emotion through text).

A significant body of research has focused on benchmarking an LLM's emotion understanding, which corresponds to the ability to "recognize emotions" (Samad, 2014). This is most commonly evaluated as a text classification task using established datasets. The GoEmotions dataset, for instance, is frequently used to test model performance (Demszky et al., 2020). Recent work has used GoEmotions to compare various LLMs, analyzing their accuracy in identifying human emotions from text (Lecourt et al., 2025). However, some researchers argue that such datasets, while useful, are insufficient for a comprehensive evaluation and have proposed new benchmarks like EmoBench, which includes hand-crafted questions designed to require deeper emotional reasoning beyond simple recognition (Sabour et al., 2024). These studies treat classification accuracy on such benchmarks as a primary indicator of an LLM's emotion understanding capability.

In parallel, research on emotion generation focuses on the other aspect of EI: the ability to "invoke and reason with emotions" (Samad, 2014). For an LLM, this translates into generating text that is both emotionally appropriate and coherent. This has motivated the development of techniques aimed at enhancing the emotional expressiveness of model outputs. For instance, the Emotional Chain-of-Thought (ECoT) method prompts LLMs to reason about emotions before generating a response, thereby improving their performance on generation tasks (Li et al., 2024). Evaluating the quality of such emotionally generated text presents its own challenges, which have led to the proposal of new metrics. The Emotional Generation Score (EGS) evaluates generated outputs based on psychological theories (Li et al., 2024). This metric is different from the EGS defined in the Equation 3, although the two share a similar name. In contrast, Jafari et al. (2025) introduced an embedding-based automatic evaluation metric for emotional text generation.

While these two components of EI are often studied in isolation, some prior work has acknowledged the need to connect them for a more holistic view of emotional intelligence (Zhao et al., 2024). However, our work is motivated by the hypothesis that strong performance in emotion understanding does not necessarily translate to high performance in emotion generation. We directly investigate this potential divergence by quantitatively comparing these two capabilities across multiple LLMs to empirically demonstrate the degree to which they differ and motivate the need for distinct evaluation metrics for understanding versus generation.

3 Dataset

We use two different datasets: one for evaluating emotional text understanding and another for evaluating emotional text generation. The UniC dataset (Du et al., 2025) is used for the understanding evaluation, while the generation evaluation relies on texts produced by various LLMs that are annotated with emotion labels. The details of each component are provided below.

3.1 Emotion Understanding Dataset

A considerable body of research has investigated the emotion understanding capabilities of LLMs by evaluating them on established emotion classification datasets. Many of these datasets originate from social media platforms or news headlines. For example, the widely used GoEmotions dataset contains 58,000 Reddit comments annotated with 27 fine-grained emotion categories (Demszky et al., 2020; Fitriana and Setiawan, 2025). Similarly, datasets constructed from microblogging platforms have been employed to study emotion classification in short and informal text (Wen and Wan, 2014). Other studies have utilized news headlines, such as the dataset from the SemEval 2007 workshop (Strapparava and Mihalcea, 2007), to explore relationships between lexical semantics and emotion

Model	Developer	Parameters	Context Window(tokens)	License / Source
GPT-4.1 (OpenAI, 2025a)	OpenAI	Proprietary	1M	Closed / Commercial API
GPT-4o-mini (OpenAI, 2025b)	OpenAI	Proprietary	128K	Closed / Commercial API
Llama-3.3-70B-Instruct (Grattafiori et al., 2024)	Meta AI	70B	128K	Open (Meta Llama License)
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Meta AI	8B	128K	Open (Meta Llama License)
Mistral-NeMo-12B-Instruct (AI, 2024)	Mistral AI	12B	128K	Apache-2.0 (Open)

Table 1: Technical overview of the LLMs used in this study, including developer, parameter count, context window size, and licensing/source information.

(Danisman and Alpkocak, 2008). Although these datasets are valuable for training and benchmarking models on explicit emotional expressions, they are limited in their ability to capture more nuanced forms of emotional intelligence. Social media posts and headlines tend to be brief, highly contextual, and often include performative or fact-based language rather than implicitly conveyed, naturally expressed personal emotions (Yang et al., 2023). These limitations motivated our adoption of the UniC dataset (Du et al., 2025), which was curated from YouTube monologues (such as book and film reviews). UniC captures multimodal, non-acted, implicit, and naturally occurring emotional expressions, making it well-suited for our study. The dataset comprises 964 video clips derived from a set of source videos and curated through a multi-step pipeline involving keyword search, subtitle filtering, and manual validation. The clips are short (approximately 10 seconds each) and were independently annotated across four modalities: text, audio, silent video, and all modalities combined. Annotations include categorical labels (26 initial categories later clustered into seven emotions) as well as dimensional valence-arousal scores. The seven emotions are disappointment, disgust, confusion, neutral, contentment, joy, and surprise. In this study, we rely solely on the text modality (i.e., transcripts).

3.2 Emotion Generation Dataset

To develop a dataset suitable for evaluating emotion-aware text generation systems, we construct parallel textual variants that differ only in emotional style while preserving semantic meaning. The UniC dataset consist of 964 transcripts drawn from 18 videos, each containing spontaneous emotional cues.

3.2.1 Generate Neutralized and Emotionalized Samples

Each transcript is processed through a two-step prompting pipeline. In the first step, a neutralization prompt (Box A.1) removes explicit emotional expressions while maintaining the original semantics and sentence structure. This ensures that emotional cues present in the source text do not influence the LLM during emotion generation, enabling a controlled comparison with the original transcript. Second, an emotion reinjection prompt (Box A.2) introduces a target emotion with minimal stylistic deviation from the neutralized version. This *neutralize* $\rightarrow$ *reinject* procedure produces three aligned versions per sample: the original transcript, a neutralized form, and an emotionalized version. Five LLMs (see Table 1) are used for generation, producing a total of 964 samples $\times$ 5 models $=$ 4,820 generated texts. If the original text is labeled as neutral, both the neutralization and neutral emotion re-injection stages should ideally leave the text unchanged, resulting in identical original, neutralized, and re-injected outputs.

3.2.2 Sample Selection for Human Annotation

Direct human annotation of all 4,820 generated texts would be prohibitively costly. Therefore, we selected a representative subset using a two-step sampling strategy. First, for each emotion category, transcripts were sampled from both the beginning and the end of each of the 18 videos to ensure content diversity. Second, we prioritized samples with high-quality ground-truth labels. Specifically, from a smaller manually annotated portion of the UniC dataset (61 samples labeled by three annotators), we retained only samples whose majority score exceeded 0.6, corresponding to agreement in at least 2 of the 3 annotations, resulting in 55 reliable samples. Applying both criteria resulted in a final subset of 277 high-quality samples. Across all models, this corresponds to: 277 samples $\times$ 5 models $=$ 1,385 generated texts. Thus, for emotional text generation, we annotated a total of 1,385 samples.

The neutralization process is not the main component of the pipeline, so we annotated a representative subset of the 277 samples. Specifically, we selected 12 samples per emotion category from the 277 samples. Given 7 emotion categories, this

Model	Krippendorff's α	Fleiss' κ
GPT-4.1	0.2918	0.2903
LLaMA-3.3-70B	0.3023	0.3008
LLaMA-3.1-8B	0.2851	0.2835
Mistral-Nemo-12B	**0.3924**	**0.3911**
GPT-4o-mini	0.2268	0.2252

Table 2: Inter-annotator agreement (IAA) scores for emotion neutralization evaluation across five different LLMs.

resulted in: $12 \times 7 = 84$ samples. Across the 5 models, this yields a total of: $84 \times 5 = 420$ samples for neutralization annotation.

3.2.3 Human Annotation of Neutralized Texts

Human evaluation of neutralization was conducted on an 84-sample subset per LLM to assess the effectiveness of emotion neutralization. Annotators were informed that each text was an automatically neutralized version of an originally emotional sentence and were asked to identify the perceived emotion. Selecting *neutral* indicates successful neutralization, whereas choosing any other emotion suggests a failure in the neutralization process. Five annotators independently annotated each sample. To assess the reliability and consistency of the annotations, we computed two robust inter-annotator agreement (IAA) metrics: Krippendorff's α (Krippendorff, 2018) and Fleiss' κ (Fleiss, 1971). Unlike simple percentage agreement, both metrics correct for chance agreement and therefore provide a more conservative estimate of annotation reliability. The IAA results for all evaluated models are shown in Table 2. Overall, the agreement scores indicate *fair agreement* across models. Higher values of α and κ reflect greater consistency among annotators in identifying neutralized text, whereas lower values suggest residual emotional signals or ambiguity in the generated outputs.

3.2.4 Human Annotation for Generated Emotional Texts

Human evaluation was performed on the 277-sample subset to assess the perceived emotional accuracy and consistency of the generated outputs. Annotators are asked to select the most appropriate emotion label from the predefined set of seven categories. Five annotators independently annotated each sample. To further validate the consistency of the annotations, we again computed the two inter-annotator agreement metrics Krippendorff's α and Fleiss' κ. As shown in Table 3, α and κ values fall between 0.38 and 0.45, which indicates a moderate level of agreement among annotators. These values are typical in emotion annotation tasks (Du et al., 2025), which inherently involve subjective interpretation. Notably, GPT-4.1 yields the highest agreement scores ($\alpha = 0.4497$, $\kappa = 0.4469$), suggesting that annotators slightly more consistently perceived the emotional cues in its outputs compared to the other models.

Model	Krippendorff's α	Fleiss' κ
GPT-4.1	**0.4497**	**0.4469**
LLaMA-3.3-70B	0.4261	0.4232
LLaMA-3.1-8B	0.4394	0.4364
Mistral-Nemo-12B	0.4451	0.4422
GPT-4o-mini	0.3841	0.3807

Table 3: Inter-annotator agreement scores for emotion re-injection across models using Krippendorff's α and Fleiss' κ. Higher values indicate stronger annotator consensus.

4 Metrics

To evaluate the performance of LLMs in emotion understanding (OERS, GERS), emotion generation (EGS), and internal consistency (ECS), we define four metrics based on a VAE-based emotion cost matrix. Let N denote the total number of samples, o_i the target (ground-truth) emotion for the i-th sample, $\hat{o}_i$ the emotion predicted by the model, and $C(\cdot,\cdot)$ an emotion cost function derived from a VAE-based emotion embedding space. Unlike standard classification metrics such as F1, the VAE-based cost matrix (VCM) explicitly captures semantic similarity between emotion categories (for a more detailed discussion, see Section 4.2).

- **Original Emotion Recognition Score (OERS)**: This metric evaluates a model's ability to recognize emotions in the original data by comparing its predictions with the reference emotion labels.

$$\text{OERS} = \frac{1}{N}\sum_{i=1}^{N} C\left(o_i^{\text{original}}, \hat{o}_i^{\text{original}}\right) \quad (1)$$

- **Generated Emotion Recognition Score (GERS)**: This metric measures how well a model recognizes emotions in generated text as perceived by human annotators, by comparing model predictions with human-labeled

Model	OERS	GERS	EGS	ECS
GPT-4.1	**0.4943**	**0.7389**	**0.7248**	**0.7669**
LLaMA-3.3-70B	0.4559	0.7406	0.7129	0.7515
LLaMA-3.1-8B	0.2925	0.5452	0.6470	0.5010
Mistral-Nemo-12B	0.3306	0.4964	0.6033	0.4970
GPT-4o-mini	0.3436	0.5818	0.5988	0.5499

Table 4: Macro-averaged F1 scores across four evaluation dimensions: UniC and Prolific Understanding Scores (OERS, GERS), Generation Score (EGS), and Consistency Score (ECS).

emotion annotations.

$$\text{GERS} = \frac{1}{N}\sum_{i=1}^{N} C\left(o_i^{\text{generated}}, \hat{o}_i^{\text{generated}}\right) \tag{2}$$

- **Emotion Generation Score (EGS)**: This metric evaluates emotion generation quality by measuring the alignment between the reference emotion label of the source data and the emotion perceived by human annotators in the generated output.

$$\text{EGS} = \frac{1}{N}\sum_{i=1}^{N} C\left(o_i^{\text{original}}, o_i^{\text{generated}}\right) \tag{3}$$

- **Emotion Consistency Score (ECS)**: This metric assesses the internal emotional consistency of the model by measuring whether the emotion explicitly specified during text generation is correctly identified by the same model when it is subsequently asked to infer the dominant emotion of its own generated text.

$$\text{ECS} = \frac{1}{N}\sum_{i=1}^{N} C\left(o_i^{\text{original}}, \hat{o}_i^{\text{generated}}\right) \tag{4}$$

We begin by evaluating model performance using the macro-averaged F1 score, computed uniformly across all emotion classes. This metric treats all misclassifications as equally severe and is commonly used for multi-class emotion recognition. Table 4 reports the macro F1 results for different LLMs across four evaluation dimensions.

Under this uniform-penalty evaluation, GPT-4.1 achieves the strongest overall performance. To better understand the source of these scores, Table 5 provides a per-emotion breakdown for GPT-4.1 across all metrics. The reason the results for the *confusion* label appear excessively high is that, during the neutralization stage, 9 of the GPT-4.1 errors were *confusion* label. As a result, the neutralized texts generated by GPT-4.1 still contain confusion cues in 9 samples(for more details, see Table 9).

Emotion	# Samples	OERS	GERS	EGS	ECS
Disappointment	60	0.5833	0.8750	0.8833	0.9000
Disgust	22	0.4091	0.8500	0.6818	0.8636
Confusion	17	0.9412	0.9375	0.8235	1.0000
Neutral	66	0.3939	0.6438	0.8485	0.6667
Contentment	68	0.4118	0.7895	0.4706	0.6765
Joy	32	0.3750	0.6818	0.7813	0.7500
Surprise	12	0.7500	0.5909	0.8333	0.8333
Macro Average	277	0.4943	0.7379	0.7248	0.7669

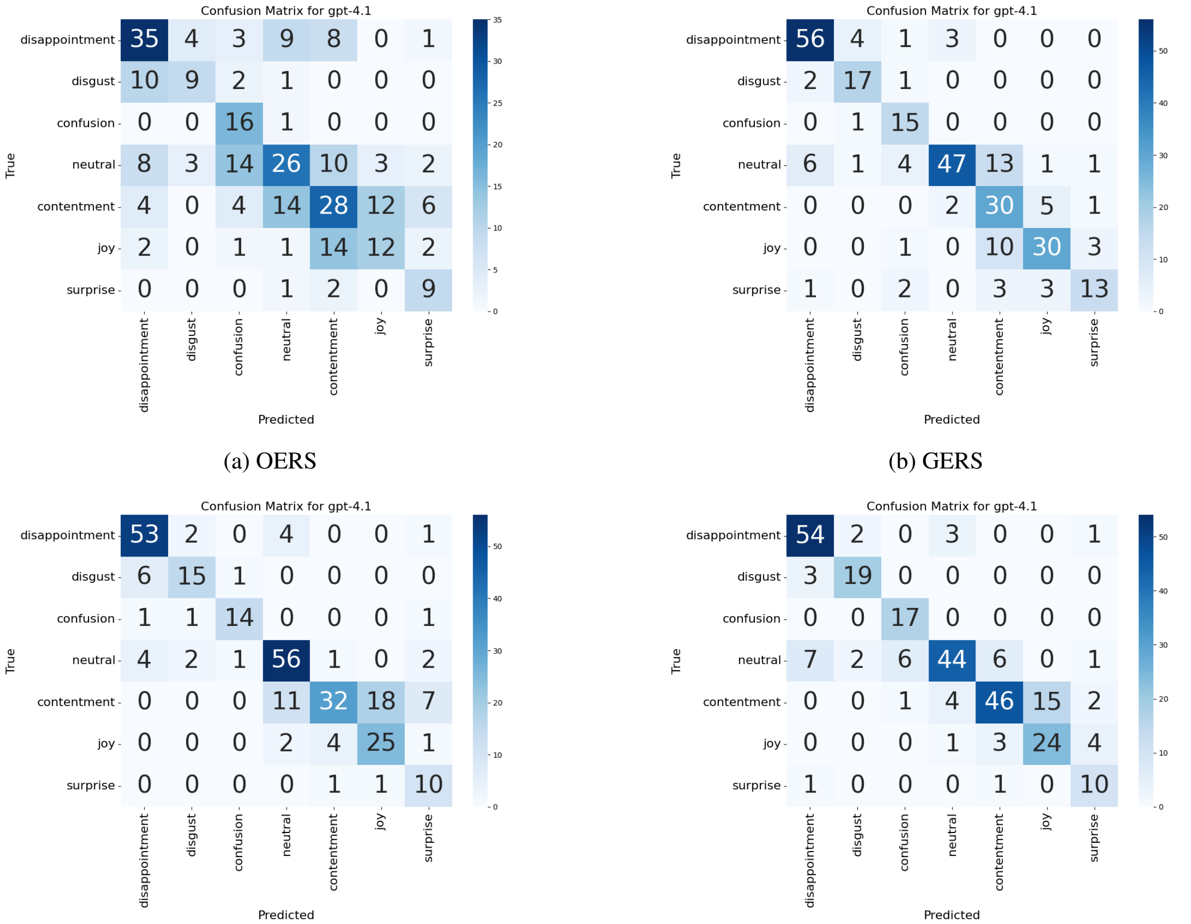
Table 5: Per-emotion performance of GPT-4.1 across evaluation metrics.

Figure 1 presents the confusion matrices corresponding to the four evaluation metrics for GPT-4.1. A consistent pattern emerges across metrics: a substantial portion of errors arises from confusion between *joy* and *contentment*. While one might suggest merging these classes, a similar confusion pattern is also observed between *contentment* and *neutral*. Merging classes in response to such overlaps would therefore lead to an undesirable collapse of distinct emotional states (e.g., *joy*, *contentment*, and *neutral*), which are conceptually and functionally different. Given this continuum-like relationship among emotions, penalizing confusions between semantically adjacent classes in the same way as confusions between semantically distant emotions (e.g., *joy* and *disgust*) is arguably inappropriate. This observation motivates the need for an evaluation framework that explicitly accounts for semantic proximity among emotion classes, rather than relying on coarse class merging.

4.1 Manually Defined Cost Matrix

To address the limitations of uniform-penalty metrics, we first introduce a manually defined semantic cost matrix (De Bruyne, 2022), shown in Figure 2. This matrix assigns lower costs to misclassifications between semantically similar emotions and higher costs to confusions between semantically distant ones.

Table 4 reflects performance under the standard macro F1 setting, whereas Table 6 reports results obtained using the manual cost matrix. Unlike macro F1, this evaluation incorporates polarity-aware misclassification costs between emotion categories. While the overall ranking of models remains broadly consistent, the cost-sensitive evaluation amplifies performance differences. Stronger



(a) OERS

(b) GERS

(c) EGS

(d) ECS

Figure 1: Confusion matrices for GPT-4.1 across the four evaluation metrics.

Model	OERS	GERS	EGS	ECS
GPT-4.1	**0.7004**	**0.8700**	**0.8761**	**0.8809**
LLaMA-3.3-70B	0.6871	0.8664	0.8556	0.8676
LLaMA-3.1-8B	0.5872	0.7774	0.8111	0.7341
Mistral-Nemo-12B	0.6258	0.7906	0.8002	0.7401
GPT-4o-mini	0.6474	0.8279	0.7858	0.7714

Table 6: Model performance under the manually defined semantic cost matrix.

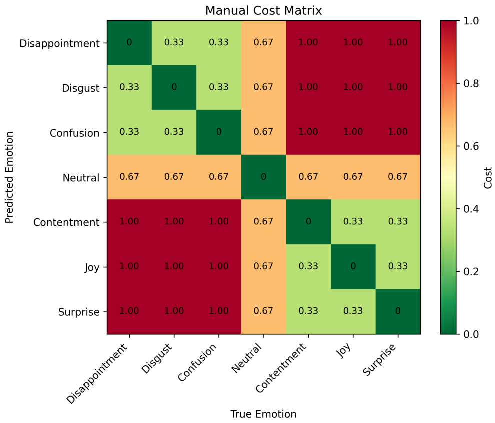
models benefit from making errors within the same polarity, which are penalized less, whereas weaker models incur higher penalties due to confusion across opposite polarities. This suggests that macro F1 may obscure qualitative differences in how models reason about emotional polarity.

4.2 VAE-Based Cost Matrix

Although effective, manually designing a cost matrix is inherently subjective. To overcome this limitation, we propose an automatic, data-driven approach based on latent affective representations learned by a variational autoencoder (VAE). We employ three-dimensional sentiment embeddings from the SentiVAE model (Hoyle et al., 2019), which encode affective semantics in a continuous latent space. Each emotion label is mapped to its corresponding latent vector, and semantic similarity between emotions is computed using cosine similarity. To obtain a misclassification cost, similarity is transformed into dissimilarity by subtracting the cosine similarity from one, followed by normalization to the $[0, 1]$ range (see Algorithm 1). The resulting VAE-based cost matrix (VCM) is visualized in Figure 3, while Table 7 reports evaluation results using this VAE-based cost matrix.

Model	OERS	GERS	EGS	ECS
GPT-4.1	**0.8549**	**0.9535**	**0.9640**	**0.9581**
LLaMA-3.3-70B	0.8494	0.9524	0.9531	0.9533
LLaMA-3.1-8B	0.7871	0.9120	0.9396	0.8998
Mistral-Nemo-12B	0.8103	0.9155	0.9305	0.9196
GPT-4o-mini	0.8265	0.9410	0.9310	0.9235

Table 7: Evaluation results using the VAE-based cost matrix.



	Disappointment	Disgust	Confusion	Neutral	Contentment	Joy	Surprise
Disappointment	0	0.33	0.33	0.67	1.00	1.00	1.00
Disgust	0.33	0	0.33	0.67	1.00	1.00	1.00
Confusion	0.33	0.33	0	0.67	1.00	1.00	1.00
Neutral	0.67	0.67	0.67	0	0.67	0.67	0.67
Contentment	1.00	1.00	1.00	0.67	0	0.33	0.33
Joy	1.00	1.00	1.00	0.67	0.33	0	0.33
Surprise	1.00	1.00	1.00	0.67	0.33	0.33	0

Figure 2: Manually defined semantic cost matrix for emotion classification. Darker colors indicate higher misclassification costs, while lighter colors correspond to lower misclassification cost for emotion pairs.

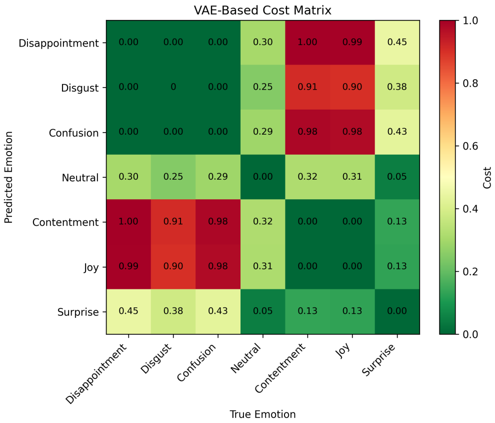
Algorithm 1 VAE-Based Cost Matrix Construction

```
Require: Emotion set $\mathcal{E} = \{e_1, \ldots, e_N\}$; VAE sentiment dictionary $\mathcal{D}$
Ensure: Normalized cost matrix $\mathbf{C} \in [0, 1]^{N \times N}$
for $i = 1$ to $N$ do
    Retrieve latent vector $\mathbf{v}_i \in \mathbb{R}^3$ for emotion $e_i$
end for
for $i = 1$ to $N$ do
    for $j = 1$ to $N$ do
        $s_{ij} = \cos(\mathbf{v}_i, \mathbf{v}_j)$
        $C_{ij} = 1 - s_{ij}$
    end for
end for
Normalize $\mathbf{C}$ to $[0, 1]$
return $\mathbf{C}$
```

Finally, Table 8 reports the Pearson correlation (Pearson, 1895) between the VAE-based and manually defined cost matrices. The correlations are consistently high across all metrics, indicating that the VAE-based approach closely approximates the manually designed costs. Given this near-perfect alignment, we adopt the VAE-based cost matrix for all subsequent experiments, as it provides a principled, automated, and semantically grounded alternative to manual cost specification.

Metric	OERS	GERS	EGS	ECS
Correlation	0.9985	0.9873	0.9850	0.9656
p-value ($\times 10^{-5}$)	6.90	171.55	220.14	763.27

Table 8: Pearson correlation and corresponding p-values (all reported in units of 10^{-5}) between VAE-based and manually defined cost matrices.



	Disappointment	Disgust	Confusion	Neutral	Contentment	Joy	Surprise
Disappointment	0.00	0.00	0.00	0.30	1.00	0.99	0.45
Disgust	0.00	0	0.00	0.25	0.91	0.90	0.38
Confusion	0.00	0.00	0.00	0.29	0.98	0.98	0.43
Neutral	0.30	0.25	0.29	0.00	0.32	0.31	0.05
Contentment	1.00	0.91	0.98	0.32	0.00	0.00	0.13
Joy	0.99	0.90	0.98	0.31	0.00	0.00	0.13
Surprise	0.45	0.38	0.43	0.05	0.13	0.13	0.00

Figure 3: VAE-based cost matrix derived from latent sentiment embeddings.

5 Results and Discussion

We first present the results of the neutralization stage, which constitutes the initial component of the proposed pipeline. This is followed by the results of the emotion re-injection stage. Finally, we provide a qualitative analysis based on two representative samples.

5.1 Neutralization Results

Table 9 summarizes the performance of different models on the emotion neutralization task over 84 annotated samples. In the ideal (best-case) scenario, all outputs should be classified as *neutral*, since the models were explicitly instructed to neutralize the emotional content of the input texts. However, after conducting human evaluation, we found that a subset of the generated outputs still conveyed residual emotions and were therefore annotated as non-neutral. For F1-score computation, the ground-truth labels consist of 84 instances of the *neutral* class (i.e., the true values assume all samples should be neutral). The predicted labels correspond to the *majority emotion* assigned by human annotators to each generated neutralized text. Each sample was annotated by five annotators, and the dominant emotion among them was used as the final prediction.

Among the evaluated models, *LLaMA-3.3-70B* achieves the highest F1-score (0.95), indicating a strong ability to remove emotional cues and produce genuinely neutral outputs, with very few residual emotional mistakes. *LLaMA-3.1-8B* follows with an F1-score of 0.83, demonstrating robust performance despite its smaller model size. *GPT-4.1* and *Mistral-Nemo-12B* achieve comparable F1-

scores (0.78), suggesting similar neutralization behavior across proprietary and open-source models. Finally, *GPT-4o-mini* records the lowest F1-score (0.77), reflecting comparatively weaker neutralization performance and a higher number of non-neutral outputs. Overall, these results demonstrate that instruction-tuned open-source models, particularly larger variants, can achieve strong and competitive performance in the neutralization stage of the proposed pipeline.

5.2 Emotion Re-injection Results

Table 7 reports the evaluation results for all models using the VAE-based cost matrix on 277 samples. According to the metrics defined in Section 4, o_i^{original} denotes the original UniC emotion labels, while $\hat{o}_i^{\text{original}}$ represents the model predictions on the UniC dataset. Furthermore, $o_i^{\text{generated}}$ corresponds to the human annotations of the generated texts, and $\hat{o}_i^{\text{generated}}$ denotes the model predictions on these generated samples. Overall, larger models consistently outperform smaller ones across all four metrics, with GPT-4.1 achieving the highest scores in understanding, generation, and consistency.

To further analyze the relationship between the proposed metrics, Figure 4 presents the Pearson correlation matrix computed over average VAE-based cost matrix scores. The results reveal a strong correlation between OERS and GERS ($r = 0.96$), indicating that models that better understand emotions in the original dataset also tend to align more closely with human emotion perception in generated text. The ECS exhibits the highest correlation with OERS ($r = 0.98$), suggesting that strong emotion understanding is closely tied to internally consistent emotion generation. In contrast, the EGS exhibits comparatively lower correlations with understanding-based metrics ($r \approx 0.70$). This discrepancy strongly suggests that emotion understanding and internal consistency metrics are not sufficient proxies for evaluating an LLM's emotion generation capability. While models may accurately recognize or internally align with emotional intent, this does not necessarily translate into a good emotional text generation model. These findings underscore the need for specialized evaluation protocols that independently assess emotion understanding and emotion generation, thereby providing a more accurate and nuanced characterization of an LLM's emotional intelligence.

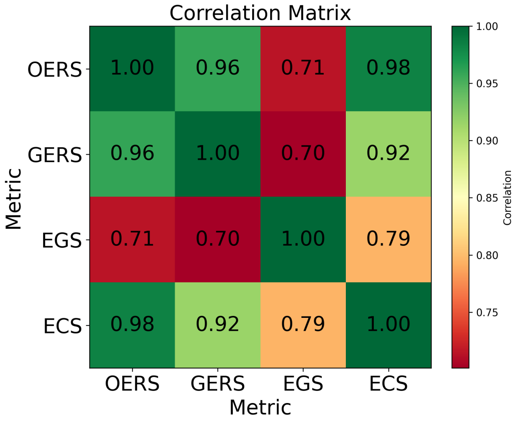


Figure 4: Pearson correlation matrix between emotion understanding, generation, and consistency metrics using average VCM scores.

5.3 Qualitative Analysis

Table 10 presents two representative examples illustrating how emotional cues introduced during enrichment influence both human annotations and GPT-4.1 emotion predictions across different annotation sources. Importantly, *UniC Pred.* refers to GPT-4.1 predictions on samples annotated in the UniC dataset, while *Prolific Pred.* denotes GPT-4.1 predictions on emotionally enriched texts whose labels originate from Prolific crowd annotations.

In the first example, the original text describes a disturbing narrative involving cannibalism and violence, yet includes a subtle positive evaluative cue ("trust me this is a good book"). Despite the dark subject matter, Prolific annotations predominantly label the emotion as *contentment*, suggesting that annotators attend to the narrator's engaged and appreciative stance rather than the literal events described. The neutralized version removes this evaluative framing, yielding a more descriptive and emotionally flattened account. In the emotionally enriched version, affective signals are explicitly amplified through phrases such as "fascinating book," "eager curiosity," and "strangely satisfying." While GPT-4.1 predictions on the generated emotional text aligned with Prolific annotations indicate *contentment*, GPT-4.1 predictions on the original text (UniC-annotated version) shift to *surprise*. The second example exhibits a clearer alignment between emotional framing and semantic content. The original text expresses a preference for "silly, light-hearted stories," which is further reinforced in the emotionally enriched version through explicit affective language such as "gentle smile" and

Model	F1	Disappointment	Disgust	Confusion	Neutral	Contentment	Joy	Surprise
GPT-4.1	0.7826	3	5	9	54	3	3	7
LLaMA-3.3-70B	**0.9500**	2	2	1	76	2	0	1
LLaMA-3.1-8B	0.8333	3	2	3	60	3	7	6
Mistral-Nemo-12B	0.7826	8	4	5	54	3	3	7
GPT-4o-mini	0.7737	11	0	6	53	6	3	5

Table 9: Emotion neutralization performance of different models measured using F1-score. The per-emotion columns report the number of samples annotated with each dominant emotion by human evaluators. Since all inputs were expected to be neutral after processing, non-neutral categories represent emotional mistakes made by the models.

UniC Emotion	Prolific Emotions	Original Text	Neutral Text	Emotional Text	UniC Pred.	Prolific Pred.
contentment	{*joy*: 1, *contentment*: 2, *surprise*: 1, *neutral*: 1}	about cannibalism trust me this is a good book we follow our female main character who is a literal psychopath and we follow her as she goes after her ex-lovers kills them and then eats them	about cannibalism this is a book we follow our female main character who is a literal psychopath and we follow her as she goes after her ex-lovers kills them and then eats them	about cannibalism this is a fascinating book we follow our female main character who is a literal psychopath and we follow her with a sense of eager curiosity as she confidently goes after her ex-lovers, kills them with chilling precision, and then eats them. Each page unfolds her unnerving world in a strangely satisfying way.	surprise	contentment
contentment	{*neutral*: 5}	noticed that I personally have a preference for like the more like silly light-hearted stories as opposed to beach read which falls more into like the sweeping emotional side of things	noticed that I personally have a preference for the more silly, light-hearted stories as opposed to beach read which falls more into the sweeping emotional side of things	I have a preference for stories that are more light-hearted, the kind that bring a gentle smile or a comforting warmth, as opposed to 'Beach Read,' which falls more into the sweeping emotional side of things.	contentment	contentment

Table 10: An example annotation from the dataset illustrating the original, neutralized, and emotionally enriched texts, along with their corresponding emotion labels and model predictions. The yellow highlights indicate emotionally salient cues and are shown solely for clarification purposes in these two samples. For text generation and emotion prediction, we used GPT-4.1.

"comforting warmth". In this case, GPT-4.1 predictions for both UniC and Prolific annotated samples consistently yield *contentment*. This consistency suggests that when emotional cues are explicit and thematically congruent, model predictions remain stable.

6 Conclusion

In this work, we proposed a semantic, VAE-based evaluation framework for assessing LLMs' emotional intelligence across emotion understanding, generation, and internal consistency. Through a comprehensive analysis of the UniC dataset and a model-generated dataset annotated by human evaluators, we showed that strong emotion understanding does not necessarily imply accurate emotion generation. Our results, supported by a correlation analysis, highlight emotion generation as a distinct and more challenging capability, motivating the need for dedicated evaluation protocols beyond traditional understanding-based metrics.

7 Future Work

In future work, we plan to include a sensitivity analysis and a discussion on how slight variations in neutralization and emotion reinjection prompts could impact the final scores. Moreover, the current framework depends on human annotators to compute GERS and EGS, which may limit its scalability to larger datasets. To address this, future research could explore automated evaluators that reduce annotation costs while maintaining reliable assessment quality. Additionally, other embedding methods, such as transformers or word vectors, should be tested for constructing the cost matrix, as our current experiments only used VAE-based embedding vectors for each emotion label.

Acknowledgments

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

Limitations

Our evaluation is limited to a fixed emotion taxonomy and a single cost matrix, which may not fully capture culture-specific or context-dependent emotional nuances. Additionally, human annotations are subject to individual perception variability, which can introduce noise into the annotation process.

References

Mistral AI. 2024. Mistral nemo. https://mistral.ai/news/mistral-nemo/.

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 convention communication, interaction and social intelligence*, pages 53–59. Aberdeen, Scotland.

Luna De Bruyne. 2022. *Feeling EmotioNL : automatically detecting emotions in Dutch texts*. Ph.D. thesis, Ghent University.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025. Unic: a dataset for emotion analysis of videos with multimodal and unimodal labels: Q. du et al. *Language Resources and Evaluation*, pages 1–36.

Frizka Fitriana and Hendrik Setiawan. 2025. Performance analysis of svm in emotion classification: A comparative study of tf-idf and countvectorizer. *Journal of Embedded Systems, Security and Intelligent Systems*, pages 133–145.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Ryan Cotterell, and Isabelle Augenstein. 2019. Combining sentiment lexica with a multi-view variational autoencoder. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 635–640.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.

Shin-nosuke Ishikawa and Atsushi Yoshino. 2025. Ai with emotions: Exploring emotional expressions in large language models. *arXiv preprint arXiv:2504.14706*.

Sadegh Jafari, Els Lefever, and Véronique Hoste. 2025. Embedding analogies for evaluating emotion in llm-generated utterances. In *28th European Conference on Artificial Intelligence (ECAI 2025)-BEHAIV workshop*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2025. 'only chatgpt gets me': An empirical analysis of gpt versus other large language models for emotion detection in text. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2603–2611.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Chenyang Zhao, Zhongwei Wan, Chaofan Tao, Wendong Xu, and 1 others. 2025. Longemotion: Measuring emotional intelligence of large language models in long-context interaction. *arXiv preprint arXiv:2509.07403*.

OpenAI. 2025a. Introducing gpt-4.1. https://platform.openai.com/docs/models/gpt-4.1. Accessed: 2025-12-12.

OpenAI. 2025b. Introducing gpt-4o-mini. https://platform.openai.com/docs/models/gpt-4o-mini. Accessed: 2025-12-12.

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

Pinaki Raj. 2024. A literature review on emotional intelligence of large language models (llms). *International Journal of Advanced Research in Computer Science*, 15(4).

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.

Haybat Abdul Samad. 2014. Emotional intelligence the theory and measurement of eq. *European Scientific Journal*.

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.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Shiyang Wen and Xiaojun Wan. 2014. Emotion classification in microblog texts using class sequential rules. In *Proceedings of the AAAI conference on artificial intelligence*.

Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, and 1 others. 2025. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. *arXiv preprint arXiv:2402.10073*.

A Prompts

Box A.1: Emotion Neutralization Prompt

Objective. This prompt removes emotional expressions from a text while preserving semantic content, structure, and linguistic properties.

Prompt Template.
Your task is to neutralize the text by removing emotional expressions.
The text is a transcription of a video and may contain emotional cues.

The output text must:
- be emotionally neutral,
- remain in the same language,
- preserve the original format, style, tone, and context,
- and differ from the input as little as possible.

Please neutralize the following text:
{text}

The original emotion of the text is: {emotion}.
Ensure that all emotional expressions are removed.

Return the result in the following JSON format:
{"neutral_text": "The neutralized text"}

Box A.2: Emotion Reinjection Prompt

Objective. This prompt reintroduces emotional expressions into a neutral text, targeting a specific emotion while maintaining semantic fidelity.

Prompt Template.
Your task is to make the text more emotional by adding emotional expressions.
The text is a transcription of a video.

The output text must:
- remain in the same language,
- preserve the original format, tone, and context,
- and differ from the input as little as possible.

Do not explicitly mention the target emotion in the text.

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}.

Return the result in the following JSON format:
{"emotional_text": "The emotionalized text"}