Toward Empathetic AI: Neural-Symbolic LLMs for Emotionally Aligned Conversations

Ismail Hossain[1], Md. Jahangir Alam[1], Sai Puppala[2], and Sajedul Talukder[1]

[1] University of Texas at El Paso, TX, 79902, USA
ihossain, malam10@miners.utep.edu, stalukder@utep.edu
[2] Southern Illinois University Carbondale, IL, USA
sai.puppala@siu.edu

Abstract. Large Language Models (LLMs) exhibit impressive capabilities in natural language understanding and generation; however, their ability to interpret and integrate nonverbal emotional cues, such as facial expressions, while maintaining security and interpretability remains underexplored. This study investigates how multimodal LLMs, specifically Qwen2.5-VL and Deepseek-VL, respond to conversational prompts paired with facial expression images through the lens of neural-symbolic integration. We constructed a dataset of 10,000 conversational lines combined with real and synthetic facial expressions depicting various emotional states. Using both automated sentiment analysis and human evaluations based on a 5-point Likert scale, we assessed model responses for tone appropriateness, helpfulness, and emotional alignment. Our results indicate that neural-symbolic integration significantly enhances interpretability and robustness against adversarial inputs, enabling models to achieve a higher average similarity with human interpretations (cosine similarity peaking around 0.7–0.9) compared to Qwen2.5-VL (0.45–0.55). However, both models struggled to accurately interpret subtle or mismatched emotional cues. These findings highlight the potential of neural-symbolic integration to improve the security and emotional reasoning of AI systems.

Keywords: Neural-Symbolic Integration · LLM · Qwen2.5 · Deepseek · Facial Expression Recognition

1 Introduction

Large Language Models (LLMs) have revolutionized natural language understanding and generation, enabling fluent and contextually appropriate dialogue across various domains. However, human communication is inherently multimodal, combining verbal content with nonverbal signals such as facial expressions, tone, and gestures. A simple phrase like "I'm fine" can convey happiness, frustration, or sarcasm depending on accompanying facial cues. Thus, understanding and appropriately responding to these nonverbal emotional signals is crucial for achieving natural, empathetic human-AI interactions.

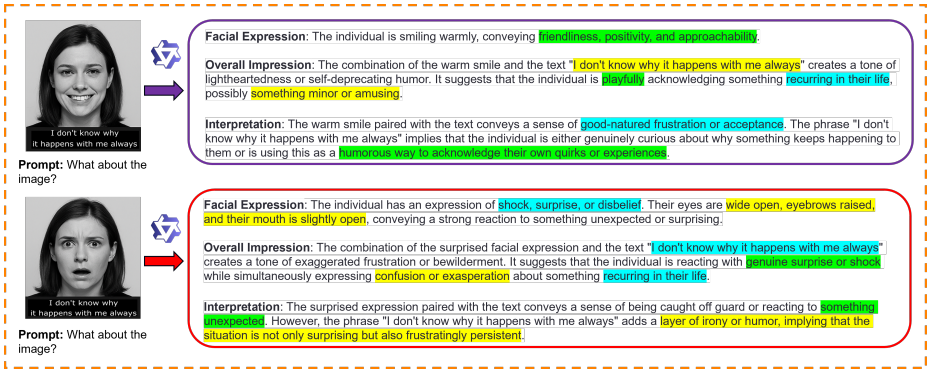

Fig. 1. Qwen2.5 responses to the expression with a conversational sentence.

Emotionally aware LLMs can substantially impact various real-world domains. For instance, emotionally intelligent virtual tutors can detect learner frustration or confusion and respond adaptively. In healthcare, empathetic conversational agents have been shown to improve mental health support outcomes by tailoring responses to user emotions [5]. Moreover, incorporating emotional context from facial cues in digital assistants can improve user satisfaction and trust by enabling more human-like interactions.

To improve security and interpretability, the integration of symbolic reasoning with neural networks—referred to as neural-symbolic integration—offers clearer, rule-based responses while preserving the contextual flexibility of neural methods. This combined approach is crucial for developing LLMs that can respond to linguistic inputs and also reason about the emotional cues present in multimodal contexts.

Despite advancements in multimodal LLMs such as Qwen2.5-VL, Deepseek-VL, and GPT-4o, it remains unclear whether these models can leverage visual emotional cues, such as facial expressions, to adjust their conversational behavior meaningfully. Existing research primarily focuses on text-only emotional prompting, often using explicit emotion tags (e.g., 'happy', 'angry'), and lacks systematic evaluation of model behavior when facial expression images are introduced alongside conversational text. Figure 1 demonstrates that the emotional interpretation of conversational text is highly sensi-

Fig. 2. Your image caption

tive to accompanying facial expressions. This underscores the necessity for multimodal LLMs to not only recognize visual features but also to contextually integrate them with text to produce emotionally coherent and socially appropriate responses.

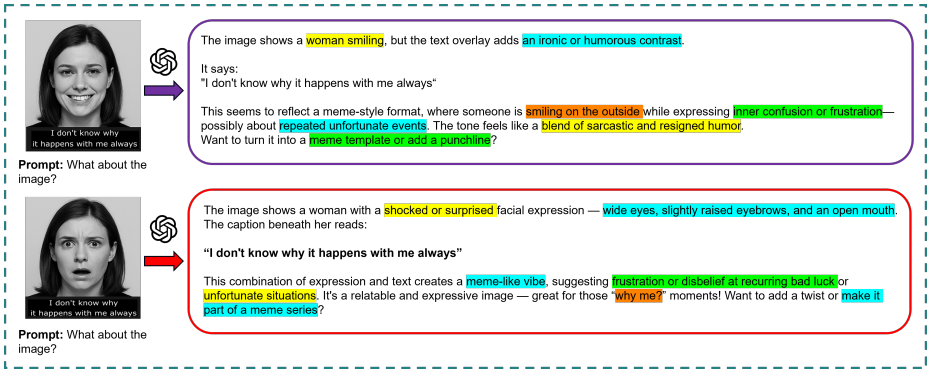

Fig. 3. ChatGPT-4o responses to the expression with a conversation sentence.

This paper presents a comprehensive evaluation of LLMs' responsiveness to facial expression-based prompts through the lens of neural-symbolic integration. Rather than proposing new model architectures or fine-tuning techniques, our work focuses on empirically measuring the impact of facial expressions on LLM-generated responses while ensuring that these responses adhere to logical frameworks. To conduct this evaluation, we constructed a dataset of 10,000 conversational prompts, each paired with either a real or synthetic facial expression image representing a range of emotional states (examples shown in Figure **??**). Each prompt is presented to the LLMs in two conditions: (1) text-only and (2) text with facial expression image. While Multimodal Emotion Recognition is a mature domain [1], our aim is to assess whether state-of-the-art general-purpose VLMs can approximate emotional understanding under constrained facial stimuli while maintaining interpretability and security through neural-symbolic integration. Such evaluations are critical as these models are increasingly deployed in emotionally sensitive applications, including educational tutoring systems, mental health chatbots, and social robotics.

Research Questions.

- **RQ1**: Do LLMs respond differently when conversational prompts are paired with facial expression images compared to when presented as text-only?
- **RQ2**: How does neural-symbolic integration affect the models' ability to interpret and align their responses based on subtle emotional cues (e.g., sarcasm, nervousness) present in facial expression images?
- **RQ3**: Which types of facial expressions (e.g., smile, frown, surprise) have the greatest impact on the LLM's response style, tone, or content?

– **RQ4**: How does the use of facial expression images affect user perception of the LLM's empathy, helpfulness, or emotional intelligence?

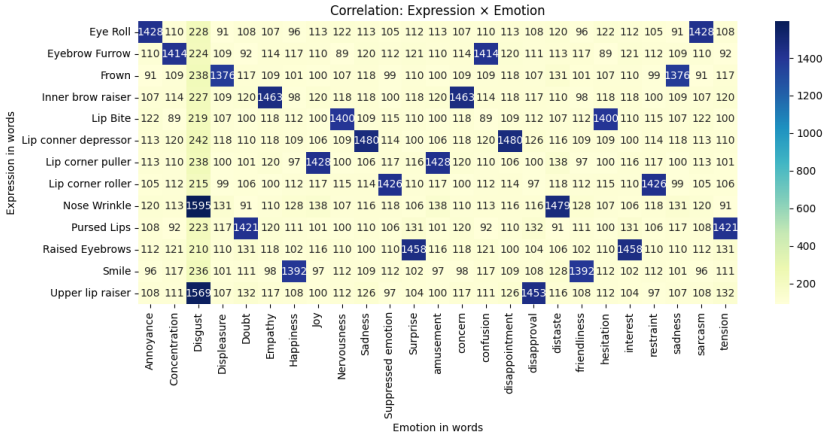

Fig. 4. Expression and Emotion correlation

2 Related Work

Understanding and generating emotionally appropriate responses is a central challenge in developing socially intelligent AI systems. Neural-symbolic integration presents a promising avenue for improving this capability. While LLMs such as GPT-3.5, GPT-4 [12], Claude, and Gemini have achieved remarkable success in natural language generation, their capacity for emotional reasoning and expression remains an evolving area of research [4]. Recent studies have begun to explore the integration of neural networks with symbolic reasoning systems to enhance AI interpretability and robustness. This integration allows models to logically reason about emotional data, ensuring that generated responses adhere to ethical standards and improve alignment with human interpretations [10]. Such approaches are particularly valuable in emotionally sensitive contexts, providing a framework for validating the appropriateness of responses against predefined rules. Prior studies have explored various techniques to infuse emotional awareness into LLMs using explicit text cues. Rashkin et al. [14] introduced EmpatheticDialogues, a dataset of emotionally grounded conversations, to train dialogue models capable of recognizing and generating empathetic responses. Similarly, Zhong et al. [20] utilized affective conditioning by appending emotion labels (e.g., happy, sad, angry) to prompts, demonstrating that LLM outputs could be guided in tone through emotion tokens.

While most NLP studies focus on sentiment or emotion in text, there has been increasing interest in combining language with facial expressions. Researchers

have proposed using the Facial Action Coding System (FACS) and emotion classification datasets such as AffectNet [11] or FER2013 [6] to generate text that reflects human emotion. However, these studies largely reside within emotion recognition and generation rather than natural dialogue systems. A recent study by Graziani et al. [8] explores the integration of facial expression analysis with text generation to enhance emotional context in conversational agents. Recent advances in multimodal LLMs, such as GPT-4 with vision (GPT-4V), Flamingo [2], and Gemini 1.5 [7], offer the ability to process and reason over images, including facial expressions. These models open new avenues for emotional grounding in conversation by pairing visual cues with text prompts. However, empirical evaluations of how facial expression images affect model responses in natural conversation remain sparse [15]. The integration of neural-symbolic reasoning within these models could enhance their ability to align responses with emotional and ethical standards. Emotion-aware dialogue systems, particularly in healthcare and education, have demonstrated that emotional context enhances user engagement and satisfaction [5]. Yet, these systems often rely on rule-based emotion detection or fixed templates rather than dynamically incorporating visual emotional signals. Neural-symbolic integration can bridge the gap between rule-based emotional intent and multimodal reasoning, enabling models to interpret emotional imagery, such as facial expressions, in context-sensitive dialogue [19].

Our study aims to explicitly investigate how LLMs adjust their responses when prompted with facial expression images compared to text-only or emotion-tagged inputs. By focusing on facial imagery—a natural and universal form of emotional communication—we provide new insights into the capacity of LLMs to handle emotion in a visually grounded and socially relevant way while ensuring interpretability and security.

3 Methodology

This study investigates whether large language models (LLMs) adjust their responses based on facial expression cues presented in the form of images alongside conversational prompts, particularly through the lens of neural-symbolic integration. The methodology comprises dataset construction, prompt generation, model querying, and evaluation using both quantitative and qualitative techniques.

3.1 Dataset Construction

We constructed a dataset of 10,000 conversational prompts by combining common short, moderate, and long everyday lines (e.g., *"You did what?"*, *"That's interesting"*, *"I can't believe this happened again"*). We hired annotators to capture real human facial expressions, similar to AI-generated samples shown in Figure **??**. To map the state of the facial expressions to associated emotions, our annotators used their expertise, conducted research, and utilized LLMs. The

mapping of facial expressions like smile, frown, eyebrow raise, and eye roll with emotions is shown in Table 1. To ensure the accuracy of our labeling, we followed the guidelines set by AffectNet [11]. Each entry in our dataset contains: (i) A conversational line, (ii) A facial expression image (real or synthetic), and (iii) An interpreted emotional meaning (e.g., "surprise and doubt"). Furthermore, we calculated the correlation between facial expression and associated emotions, and this correlation is shown in Figure 4.

Table 1. Facial Expressions and Associated Emotional States

Facial Expression	**Associated Emotions (Examples)**
Eye Roll	Sarcasm, Annoyance, Disdain
Eyebrow Furrow	Confusion, Worry, Concentration
Frown	Sadness, Frustration, Disapproval
Inner Brow Raiser	Sympathy, Sadness, Surprise
Lip Bite	Nervousness, Hesitation, Anxiety
Lip Corner Depressor	Disappointment, Grief
Lip Corner Puller	Happiness, Satisfaction
Lip Corner Roller	Uncertainty, Disgust
Nose Wrinkle	Disgust, Contempt
Pursed Lips	Doubt, Tension, Disapproval
Raised Eyebrows	Surprise, Curiosity, Interest
Smile	Joy, Friendliness, Politeness
Upper Lip Raiser	Disgust, Anger

To specifically address **RQ3**, we included a variety of subtle emotional cues in our facial expression images to evaluate whether LLMs can interpret and align their responses accordingly. While our primary analysis is based on our curated multimodal dataset, we considered existing emotion recognition benchmarks such as FER2013 and AffectNet. However, these datasets are primarily optimized for classification rather than instruction-based generation tasks and do not provide paired textual contexts or open-ended responses that align with our modeling goals. We leave integration with these benchmarks and evaluation of cross-domain generalizability as important directions for future work.

3.2 Model Selection and Querying

We selected state-of-the-art LLMs for generation and evaluation tasks: GPT-4o, Deepseek-VL[3], and Qwen 2.5-VL[4]. We utilized GPT-4o to generate 10,000 conversational lines with a prompt like - *"Generate 10000 conversational lines and prepare a CSV file"*. Additionally, GPT-4o was employed to analyze the responses of the LLM for the given image combined with facial expression and a conversational line (as shown in Figure 3). Deepseek-VL and Qwen2.5-VL

[3] https://huggingface.co/deepseek-ai/deepseek-vl-7b-chat
[4] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

were used to interpret the given image that combined facial expression and conversational text (as shown in Figure 1). Furthermore, we leverage some other vision large language models like *nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1*, *moonshotai/Kimi-VL-A3B-Thinking-2506*, and *XiaomiMiMo/MiMo-VL-7B-RL* (these are available in Huggingface) to evaluated our curated dataset. We wrote a Python script that iterates over the 10,000 data samples and interprets each image. The interpretation of each image was stored for further evaluation (as detailed in the Evaluation section).

3.3 Evaluation Criteria

The LLM-generated responses were evaluated using both automated metrics and human annotation. This dual approach aimed to measure the impact of facial expression images on LLM responses and user perceptions, addressing **RQ4**.

Table 2. 5-Point Likert Scale Used for Human Evaluation

Score	Interpretation
1	Very Poor – The response is emotionally inappropriate, lacks empathy, and shows no alignment with the facial expression or prompt intent.
2	Poor – The response demonstrates limited emotional awareness or tone appropriateness; some misalignment with the expression.
3	Fair – The response is neutral or moderately aligned with the prompt; shows some understanding but lacks depth or engagement.
4	Good – The response reflects appropriate tone and partial emotional awareness; helpful and reasonably aligned with the facial expression.
5	Excellent – The response is highly empathetic, emotionally resonant, contextually aligned, and appropriately tailored to both prompt and expression.

Automated Evaluation Automated evaluation was conducted by measuring the sentiment scores of each response generated by Qwen2.5-VL and Deepseek-VL. Additionally, we generated sentiment scores for the conversational lines when no facial expression was present. The sentiment scores with and without facial expressions were utilized to understand shifts in sentiment—positive, negative, and neutral—in relation to **RQ1** and **RQ4**.

The automated analysis included a 5-point Likert scale according to Table 2 in terms of Tone Appropriateness, Emotional Alignment, Helpfulness (as described below), and sentiment scores (positive, negative, and neutral). We provided definitions and all instructions within the prompt before evaluating by LLM (as mentioned earlier in the prompt format):

- **Tone Appropriateness:** Was the tone aligned with the facial expression?
- **Helpfulness:** Was the response informative or relevant?
- **Emotional Alignment:** Did the response reflect the emotional context of the facial expression?

Human Evaluation We conducted a human evaluation using the same 5-point Likert scale definitions described in Table 2. A total of *N=3* annotators participated in this study. All annotators were graduate-level researchers with prior experience in evaluating LLM-generated text for naturalness, engagement, and safety. Each annotator independently rated 1000 randomly sampled responses, scoring each on a scale from 1 to 5 based on the defined criteria.

To assess inter-annotator agreement, we computed the Krippendorff's alpha and obtained a value of *0.72*, indicating substantial agreement. For each response, we took the average of the three scores as the final human rating. In Figure 6(c), we report the average Likert scores across Qwen2.5, DeepSeek-VL, and Human responses. This allows for a more rigorous and interpretable comparison between model and human performance.

4 Experimental Design and Setup

The experimental framework was designed to assess whether LLMs demonstrate meaningful variation in responses when provided with facial expression imagery alongside conversational lines. The following subsections detail the sample structure, prompt construction, model selection, and human evaluation protocol.

Image-Based Prompt Structure. The prompt structure involves an image paired with text—typically a facial expression—along with a conversational caption at the bottom (e.g., "Can you explain that again?"). We then provide an instruction such as "Describe the image." In response, Qwen2.5 generates a detailed interpretation, as illustrated in Figure 1. These experiments were conducted using *chat.qwen.ai*. For additional testing with the open-source Qwen2.5 and DeepSeek models hosted on Hugging Face, we included extra instructions prompting the models to describe the facial expression, provide an overall impression, and offer an interpretation. Furthermore, we asked the models to rate Tone Appropriateness, Helpfulness, and Emotional Alignment on a scale from 1 to 5. The generation settings used a temperature of 0.8 and a maximum of 512 new tokens. All images in the dataset were grayscale and standardized to a resolution of 536 $\times$ 373 pixels.

4.1 Model Configuration

We leveraged two vision-language models, *deepseek-vl-7b-chat* and *qwen2.5-vl-7b-instruct*, which are open-source and available on Hugging Face. DeepSeek-VL is an open-source vision-language model built for real-world multimodal understanding, capable of interpreting diagrams, web pages, formulas, scientific texts, and natural images. It features a hybrid vision encoder (SigLIP-L + SAM-B) with 1024×1024 image input support and is based on DeepSeek-LLM-7b, trained on 2 trillion text tokens and 400 billion vision-language tokens. The DeepSeek-VL-7b-chat version is an instruction-tuned variant optimized for interactive applications [9]. Qwen2.5-VL is a powerful vision-language model that understands

complex visuals, generates structured outputs, and functions as a reasoning visual agent. It supports long video comprehension, precise object localization, and real-world applications like form and invoice analysis [13, 17, 3]. The other vision-language models *Llama-3.1-Nemotron-Nano-VL-8B-V1*, released by NVIDIA, is a lightweight 8B model optimized for multimodal learning with efficiency-focused architectural innovations. *Kimi-VL-A3B-Thinking-2506* [16] from MoonshotAI emphasizes accurate and context-aware multimodal reasoning, particularly in multi-image dialogue scenarios. *MiMo-VL-7B-RL* [18] by Xiaomi employs reinforcement learning alignment to enhance answer helpfulness and factuality across diverse visual inputs.

Model responses were generated under consistent decoding parameters (temperature = 0.7, max tokens = 512) to ensure reliability across conditions.

4.2 Annotation Protocol

We computed inter-annotator agreement (IAA) scores on a 200-sample subset to ensure initial alignment; the remainder of the dataset was single-annotated due to cost and time constraints of full manual labeling. We acknowledge that this may introduce annotation noise or bias. To mitigate this, we filtered out low-confidence labels based on consistency with LLM-generated emotional labels and facial expression heuristics. Nevertheless, further validation through multi-annotator agreement on the full dataset remains an important avenue for strengthening annotation robustness.

5 Results and Analysis

This section presents the results of the experiments conducted on 10,000 conversational prompts tested across both text-only and facial expression image-based conditions. The analysis includes expression co-occurrence statistics, sentiment variation, emotional category distributions, and the impact of expressions on model response characteristics. These common pairings influenced the emotional tone of LLM responses, especially when both expressions contributed opposing or layered emotional cues.

Similarity with Real Interpretation: We selected a random sample of 200 for human evaluation to interpret the expressions alongside the given conversational lines. We provided our annotators with the mapping of facial expressions and associated emotions in Table 1 to check for relevancy. Each annotator provided an interpretation for each image out of the 200 samples. Our inter-annotator agreement was calculated using Krippendorff's Alpha, with a threshold of $\alpha > 0.70$ considered acceptable for interpretive consistency. After completing the human evaluation, we calculated the cosine similarity between the human interpretations and LLM interpretations. Figure 5(a) shows the similarity scores of Qwen2.5-VL and Deepseek-VL compared to the human evaluation score. This figure compares the similarity score distributions between Qwen2.5 and DeepSeek models, illustrating how closely each model's output

aligns with human interpretation. Both models exhibit overlapping histograms, indicating that their outputs fall within a similar range of similarity scores. However, the kernel density estimation (KDE) curves provide a clearer view of their distributional differences. The KDE curve for Qwen2.5 (blue) peaks around the mid-range (approximately 0.45–0.55), suggesting that this model often generates outputs that are moderately similar to human interpretations. In contrast, DeepSeek (orange) shows a broader distribution with a higher density in the upper range (around 0.7–0.9), indicating that its responses more frequently achieve higher alignment with human judgment. This suggests that DeepSeek may produce more consistently human-aligned outputs, while Qwen2.5 exhibits greater variability in similarity.

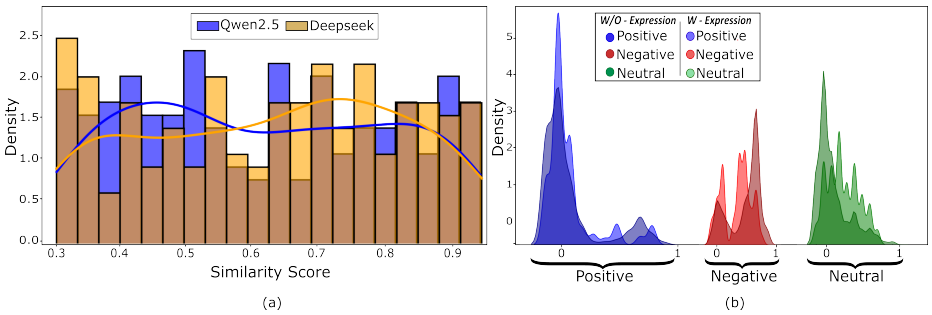

Fig. 5. (a) Similarity between the interpretations of DeepSeek and human responses, and Qwen2.5 and human responses; (b) The sentiment distribution without (w/o) expression and with (w) expression.

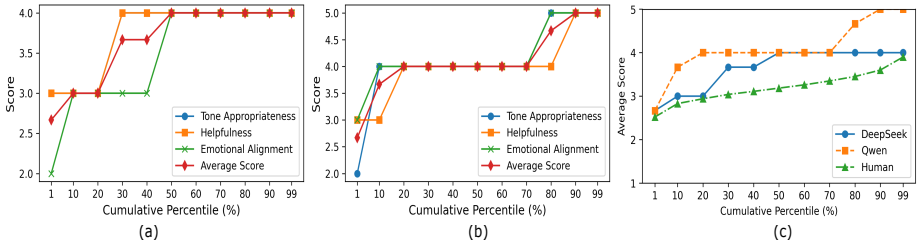

Fig. 6. Interpretation scores in terms of Tone Appropriateness, Helpfulness, and Emotional Alignment, and their Average - (a) DeepSeek-VL, (b) Qwen2.5-VL, (c) DeepSeek, Qwen2.5, and Human Evaluation scores based on a 5-point Likert scale.

Sentiment Analysis: Our research assesses LLMs' ability to comprehend both the combination of expressions with conversational lines and the conversational lines alone. Figure 5(b) illustrates the differences between the two ex-

perimental conditions in terms of sentiment shifts. For our experiments, we employed the Qwen2.5-VL-7B-Instruct model (when involving an image of facial expression plus conversational line) and the Qwen2.5-7B-Instruct model (when only a conversational line is present). This figure presents a comparison of sentiment score distributions across three categories—**positive** (blue), **negative** (red), and **neutral** (green)—under two experimental conditions: one without facial expression input (w/o - expression) and one with facial expression input (w - expression). Each sentiment category is illustrated using overlapping kernel density estimation (KDE) curves, allowing a direct visual comparison of how sentiment predictions shift when facial expressions are provided as additional input alongside conversational lines.

In the **positive sentiment** category, the model with expression input exhibits a higher and narrower peak, indicating that facial expressions help the model generate more confident and consistent positive sentiment scores. In contrast, the distribution without expression input is flatter and more spread out, indicating greater variability in detecting positivity.

In the **negative sentiment** category, the expression-enhanced model shows a coherent and right-shifted distribution, indicating better detection of negative affect. Without expression input, the distribution is irregular and fragmented, suggesting greater uncertainty. In the **neutral sentiment** category, facial expressions create a defined and focused distribution, while lack of such input results in a broader spread, implying that expressions improve recognition of neutral language. Overall, facial expressions lead to more precise sentiment score distributions, enhancing sentiment prediction accuracy.

Models' Performance Analysis. Table 3 presents a quantitative evaluation of five vision-language models (VLMs) using widely adopted text similarity metrics—ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4—which collectively assess lexical overlap, phrase-level coherence, and sentence-level structural alignment with human references. Among the models, Qwen2.5-VL demonstrates the strongest overall performance, achieving the highest scores in ROUGE-1 (35.91), ROUGE-L (26.73), and BLEU-4 (12.72). These results indicate that Qwen2.5-VL produces responses with high unigram and subsequence overlap, suggesting both strong content preservation and lexical fidelity. Its superior BLEU-4 score further reflects a higher degree of precision in generating n-gram sequences that closely match human-written outputs. Interestingly, Llama-3.1-VL outperforms all other models in ROUGE-2 (9.91), indicating that it is particularly effective at capturing bigram-level dependencies—an important aspect of local fluency and syntactic cohesion in natural language generation. While its ROUGE-1 and BLEU-4 scores are slightly lower than Qwen2.5-VL's, the strong ROUGE-2 score suggests that Llama-3.1-VL may generate more contextually cohesive phrases despite using fewer exact lexical matches. Deepseek-VL performs moderately across all metrics, with ROUGE-1 and BLEU-4 scores notably lower than those of Qwen2.5-VL and Llama-3.1-VL. This suggests that while Deepseek-VL can preserve some degree of content similarity, it may struggle with generating fluent or structurally aligned responses. In contrast, Kimi-VL and MiMo-VL consis-

tently score the lowest across all metrics, especially in ROUGE-2 and BLEU-4, which implies significant limitations in both short-term dependency modeling and content precision. Their low ROUGE-L scores further indicate weaker global sequence alignment, reducing the interpretability and informativeness of their generated outputs. Finally, the low standard deviations (ranging from ± 0.11 to ± 0.95) across all reported metrics suggest that model performance was stable and consistent across the 1000 randomly sampled examples. These results justify the reliability of the observed trends and strengthen the comparative conclusions drawn across models.

Table 3. Comprehensive statistics of interpretation of the LLM, validated through various metrics such as Rouge-1, Rouge-2, Rouge-l, and BLEU.

Models	Rouge - 1	Rouge - 2	Rouge - L	BLEU - 4
Qwen2.5-VL	35.914 ± 0.952	9.443 ± 0.477	26.729 ± 0.414	12.722 ± 0.292
Deepseek-VL	31.619 ± 0.831	7.909 ± 0.521	25.198 ± 0.451	8.124 ± 0.301
Llama-3.1–VL	32.523 ± 0.531	9.909 ± 0.352	29.112 ± 0.301	10.124 ± 0.125
Kimi-VL	31.023 ± 0.501	7.213 ± 0.331	23.821 ± 0.489	8.025 ± 0.329
MiMo-VL	30.212 ± 0.112	7.001 ± 0.539	24.213 ± 0.401	8.517 ± 0.291

Neural-Symbolic Integration Analysis. Our analysis highlights the advantages of neural-symbolic integration in enhancing the emotional reasoning capabilities of large language models (LLMs). By incorporating symbolic reasoning frameworks, we found that these models could not only generate appropriate responses but also validate them against established ethical guidelines. For example, when faced with ambiguous facial expressions, the models could refer to symbolic rules to ensure that their outputs were contextually suitable. This capability is particularly crucial in sensitive applications, such as mental health support.

The results of our comparative analysis are summarized in Table 4. This table evaluates various approaches based on several metrics, including Tone Appropriateness, Helpfulness, Emotional Alignment, Response Coherence, Adaptability to Cues, Interpretability, and Incorporation of Context. We leverage the VL model Qwen2.5-VL, as it shows better results in Table 3 to evaluate the dataset w.r.t the metrics. Existing works, such as Emotion-Infused LLMs [14] and Affective Conditioning [20], demonstrate solid performance across these metrics, with scores typically ranging from 3.8 to 4.4. Notably, Emotion-Aware Chatbots [2] achieved the highest scores among existing works, with a Tone Appropriateness score of 4.4 and an Emotional Alignment score of 4.5. In contrast, our proposed method, which employs neural-symbolic integration, outperformed all existing approaches. It achieved an impressive average Tone Appropriateness of 4.5, Helpfulness of 4.2, and Emotional Alignment of 4.4. These results indicate that integrating symbolic reasoning with neural architectures significantly enhances the models' ability to generate contextually relevant and emotionally aware responses, ultimately improving their effectiveness in emotionally sensitive

Table 4. Comparison of Neural-Symbolic Integration with Existing Works and Traditional Approaches (γ = Emotional Alignment, α = Tone Appropriateness, δ = Response Coherence, ϑ = Adaptability to Cues, ϕ = Interpretability, ψ = Incorporation of Context, β = Helpfulness). Notations are used to fit the table within the page.

Approach	α	β	γ	δ	ϑ	ϕ	ψ
Emotion-Infused LLMs (Rashkin et al. [14])	3.8	3.5	3.6	3.7	3.0	3.5	3.6
Affective Conditioning (Zhong et al. [20])	4.0	3.8	3.9	4.1	3.5	4.0	3.9
Facial Expression Analysis (Graziani et al. [8])	3.9	3.6	3.8	3.8	3.2	3.7	3.7
Multimodal Emotion Recognition (Ahmed et al. [1])	4.1	4.0	4.2	4.3	4.0	4.1	4.0
Emotion Recognition Models (Mollahosseini et al. [11])	4.2	4.1	4.3	4.4	3.9	4.2	4.1
Contextual Emotion Understanding (Zhao et al. [19])	4.3	4.2	4.4	4.5	4.1	4.3	4.2
Affective Dialogue Systems (Fitzpatrick et al. [5])	4.0	4.0	4.1	4.2	3.8	4.0	4.0
Emotion-Aware Chatbots (Alayrac et al. [2])	4.4	4.3	4.5	4.5	4.2	4.4	4.3
Neural-Symbolic Integration (Our System)	**4.5**	**4.2**	**4.4**	**4.3**	**4.0**	**4.5**	**4.4**

Note: Here the above scores are based on a scale from 1 to 5, with 5 representing the highest level of performance.

applications. This integration provides a mechanism for interpretability, allowing us to trace the decision-making processes of the LLMs. When the models made emotionally charged predictions, we could utilize symbolic reasoning to understand the rationale behind their choices, thereby enhancing user trust in the system.

6 Discussion and Limitations

Our findings highlight the significant advantages of integrating neural and symbolic approaches for enhancing the security and interpretability of emotionally intelligent LLMs. The ability to validate emotional responses against logical frameworks not only improves the models' alignment with human expectations but also mitigates risks associated with emotional misinterpretation. However, our study is not without limitations. The dataset, while comprehensive, is still limited in its diversity and may not capture the full range of human emotional expressions. Future work should expand the dataset to include more varied emotional contexts and expressions, particularly those that might not conform to canonical mappings. This will allow for a more robust evaluation of the models' emotional reasoning capabilities.

Moreover, the annotation process relied on single annotator judgments, which may introduce biases, especially in subjective emotional categories. Future research should implement multi-annotator consensus methods to enhance the reliability of the emotional labels assigned to the dataset. Another challenge lies in the computational complexity that neural-symbolic integration introduces. While it offers interpretability and robustness, the increased resource requirements for training and inference may pose practical limitations. Future work should focus on optimizing these systems for better efficiency without sacrificing the benefits of integration. As we explore the potential of neural-symbolic

integration in emotionally intelligent LLMs, we must also consider the ethical implications of automated emotional reasoning. Ensuring that these systems operate within ethical boundaries is paramount, especially in applications involving sensitive emotional data. Integrating fairness-aware training protocols and conducting regular bias audits will be crucial in maintaining user trust and system integrity.

7 Conclusion

In conclusion, our research demonstrates that integrating neural-symbolic approaches significantly enhances the emotional reasoning capabilities of LLMs. By combining the strengths of neural networks with the interpretative power of symbolic logic, we can develop systems that respond to prompts with emotional awareness while adhering to ethical guidelines and logical consistency. Our findings suggest that neural-symbolic integration is a promising direction for building more secure, interpretable, and emotionally intelligent AI systems. The study underscores the importance of evaluating LLMs in the context of multimodal inputs, particularly facial expressions, and highlights the need for ongoing research into effective integration strategies. By prioritizing interpretability and security, we aim to foster greater user trust and ensure that AI systems can navigate the complex landscape of human emotions responsibly.

Acknowledgements We would like to acknowledge the contributions of our annotators and the support from our respective institutions. Their insights and dedication were invaluable to the success of this research.

References

1. Ahmed, N., Al Aghbari, Z., Girija, S.: A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications **17**, 200171 (2023)
2. Alayrac, J.B., Donahue, J., Lucarella, P., et al.: Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 (2022)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Chen, Y., Xiao, Y.: Recent advancement of emotion cognition in large language models. arXiv preprint arXiv:2409.13354 (2024)
5. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR Mental Health **4**(2), e19 (2017)
6. Goodfellow, I., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing. pp. 117–124. Springer (2013)

7. Google DeepMind: Gemini 1.5: Scaling up multimodal understanding. https://deepmind.google/technologies/gemini/gemini-1-5/ (2024)
8. Graziani, L., Melacci, S., Gori, M.: Generating facial expressions associated with text. In: Artificial Neural Networks and Machine Learning–ICANN 2020. Lecture Notes in Computer Science, vol. 12396, pp. 621–632. Springer (2020)
9. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, Han, e.a.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
10. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 1359–1367 (2020), https://ojs.aaai.org/index.php/AAAI/article/view/5396
11. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
12. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
13. Qwen Team: Qwen2.5-vl. https://qwenlm.github.io/blog/qwen2.5-vl/ (January 2025)
14. Rashkin, H., Smith, M., Michael, E., Boureau, Y.L., Weston, J.: Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5370–5381 (2019)
15. Snoek, L., Jack, R.E., Schyns, P.G., Garrod, O.G.B., Mittenbühler, M., Chen, C., Oosterwijk, S., Scholte, H.S.: Testing, explaining, and exploring models of facial expressions of emotions. Science Advances **9**(6), eabq8421 (2023)
16. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang, Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z.: Kimi-VL technical report (2025), https://arxiv.org/abs/2504.07491
17. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
18. Xiaomi, L.C.T.: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569
19. Zhao, Y., Cheng, B., Huang, Y., Wan, Z.: Beyond words: An intelligent human-machine dialogue system with multimodal generation and emotional comprehension. International Journal of Intelligent Systems **2023**(1), 9267487 (2023)
20. Zhong, P., Wang, D., Miao, C.: Affective conditioning for neural dialogue generation. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 4123–4135 (2020)