EmoLASP: Emotion Recognition with Language Models and Answer Set Programming

Thao Le and Michael Thielscher
School of Computer Science and Engineering, University of New South Wales, Australia
{thao.le1, mit}@unsw.edu.au

Abstract

Emotion recognition in conversations is increasingly tackled with language models, but these models can be unstable and expensive to fine-tune or to prompt with long dialogue histories. We propose **EmoLASP**, a framework that combines a language model with declarative reasoning via Answer Set Programming (ASP) to predict VAD scores (Valence-Arousal-Dominance) in conversations. Experiments on a widely used benchmark dataset (IEMOCAP) across six open-source LLMs (3B–120B) and two PLMs (BERT, RoBERTa) show that EmoLASP improves prediction performance compared to using the language model alone, even when the LLMs/PLMs are given no dialogue history in their prompts or input vectors. The gains are largest for prompt-only LLMs, which EmoLASP uses without any fine-tuning. However, for fine-tuned PLMs, the reasoner adds little once dialogue history is available. EmoLASP's LLM pipeline demonstrates the potential advantages of using a reasoning approach to ensure emotion prediction consistency and to reduce both the cost of fine-tuning and the cost of prompting with long dialogue histories.

1 Introduction

Emotion recognition in conversations (ERC) is the task of identifying the emotional state expressed in each utterance of a dialogue, and it supports various applications including empathetic dialogue systems and mental health support (Chu et al., 2025). Common approaches to ERC include using pre-trained language models (PLMs) and fine-tuning them on labelled emotion datasets to predict emotions (Park et al., 2021; Mendes and Martins, 2023). In recent years, large language models (LLMs) have been used to recognise emotions in text (Feng et al., 2024; Liu et al., 2024). However, there are several limitations of previous work that need to be addressed. First, most work on ERC focuses on categorical emotion classification, which is less expressive than dimensional representations such as Valence-Arousal-Dominance (VAD) (Russell and Mehrabian, 1977; Russell, 1980). Additionally, while LLMs/PLMs have shown promising results in emotion recognition, their instability and bias (Wake et al., 2023) mean more research is needed to ensure consistent emotional reasoning in conversations. LLMs are essentially black boxes, which might provide an answer without following logical reasoning such as continuity of an emotional state. Moreover, relying solely on language models can be costly, whether due to fine-tuning on domain-specific datasets or the large number of input tokens required when feeding the entire dialogue history in prompt-based approaches.

In this paper, we present a novel approach combining language models with declarative reasoning for emotion recognition. Declarative reasoning has proven effective in other areas such as logical anomaly detection in Vision-Language Models (Jin et al., 2025). We adopt Answer Set Programming (ASP) as our reasoning framework with the following advantages. First, ASP allows us to express *non-deterministic choice rules* (Eiter et al., 2022), which consider not only the language model's top prediction but also other possible hypotheses with lower confidence. This is particularly useful when outputs are highly uncertain. Second, and on top of this, *weak constraints* in ASP allow us to encode preferences over VAD hypotheses to select the best VAD hypothesis based on the dialogue context. Since *Emotional Contagion Theory* suggests that a person's emotions can be influenced by the emotions of others in a conversation (Hatfield et al., 1993), these weak constraints favour hypotheses consistent with prior turns. Overall, the crucial contribution of the ASP layer is its *declarative* and *extensible* nature, which allows us to define preferences for choosing between hypotheses by stating explicit rules. Additional preferences, including those derived from other psychological theories,

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 40322–40334
October 24-29, 2026 ©2026 Association for Computational Linguistics

can be incorporated as new rules rather than requiring modifications to imperative code or retraining of the language model. Our experiments show that the ASP reasoner can handle historical emotion dynamics and select context-consistent emotional hypotheses even when dialogue history is omitted from the language model's input. This is where our cost saving comes from: when EmoLASP is built on prompt-only LLMs, no fine-tuning is performed at all, and omitting the history further reduces the number of input tokens per turn.

To our knowledge, this is the first work combining language models and declarative reasoning for ERC. Our contributions are: (1) we propose the **EmoLASP** framework, a novel approach integrating a language model with a reasoner for emotion recognition in conversations; (2) we evaluate it against different open-source language models, including LLMs and PLMs, on a benchmark dataset; (3) we perform qualitative analysis to identify utterances for which LLMs find it most challenging to predict accurate VAD values, even with the help of an ASP reasoner and the presence of dialogue context. Our code is publicly available.[1]

2 Method

Our proposed method has two layers: (1) an LLM/-PLM and (2) an answer set program (ASP) implemented in Clingo (Gebser et al., 2019). The overall pipeline is as follows: for each utterance, we first use a language model to provide VAD hypotheses. Then, we feed those hypotheses into an ASP reasoner, which performs *aggregation* and *optimisation* to select the best VAD hypothesis for that turn by reasoning about VAD scores of previous turns. Figure 1 illustrates the overall framework of our method, which we call **EmoLASP** (**Emo**tion recognition with **L**anguage models and **A**nswer **S**et **P**rogramming).

2.1 Preliminaries

Given a conversation C with a sequence of n utterances, $C = \{(u_1, p_1), (u_2, p_2), \ldots, (u_n, p_n)\}$, where utterance u_i is spoken by party p_i, the task of *Emotion Recognition in Conversation* (ERC) is to predict emotional **VAD** scores for each utterance u_i. We write $y_i = (v_i, a_i, d_i)$ for the gold VAD label of u_i and $\hat{y}_i = (\hat{v}_i, \hat{a}_i, \hat{d}_i)$ for our system's prediction. The scores represent a 3-dimensional model of affect, where **valence** (V) represents the positivity or negativity of an emotion, **arousal** (A) the intensity or activation level, and **dominance** (D) the degree of control conveyed by the emotion. VAD scores can be represented as continuous or ordinal values within a specific range.

2.2 Providing VAD Hypotheses

Large language models (LLMs) or pre-trained language models (PLMs) provide VAD hypotheses for each utterance. Formally, given an utterance u_i, the model produces a set of m VAD hypotheses: $H_i = \{h_i^{(1)}, h_i^{(2)}, \ldots, h_i^{(m)}\}$, where each $h_i^{(j)} \in \mathbb{R}^3$ is a three-dimensional hypothesis with valence $v_i^{(j)}$, arousal $a_i^{(j)}$, and dominance $d_i^{(j)}$.

LLMs We prompt an LLM to output m ranked VAD hypotheses for u_i, ordered from most to least likely according to the model. Depending on the context condition, the prompt includes either only u_i (*nohist*), or u_i together with all previous turns $\{u_1, u_2, \ldots, u_{i-1}\}$ (*hist*). The *nohist* condition reduces the number of input tokens, which benefits smaller LLMs with limited context windows.

PLMs We fine-tune the encoder-based PLMs BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) on the IEMOCAP (Busso et al., 2008) training set with three independent classification heads, one per VAD dimension, each predicting a Likert score in {1, 2, 3, 4, 5}.

The *nohist* and *hist* conditions are realised by encoding u_i alone or by prepending the dialogue history with a speaker label before u_i. To produce a ranked set of m hypotheses, we apply a softmax to each dimension's logits. Each candidate triple (v, a, d) is scored by the product of its per-dimension probabilities, and the top m unique triples are returned as H_i, ranked from most to least likely.

2.3 Reasoning with Answer Set Programming

The second layer of our method takes the VAD hypotheses from the language model and applies ASP to optimise emotional dimensions. We first decide whether to invoke ASP using a *confidence threshold* τ; when invoked, ASP performs two steps: (1) *aggregation* and (2) *optimisation*.

Confidence Threshold We apply ASP only when the language model is not highly confident about its predictions. A per-turn confidence score is derived from the language model's own probability distribution (from digit-token logprobs for

[1] https://github.com/thaole25/emolasp

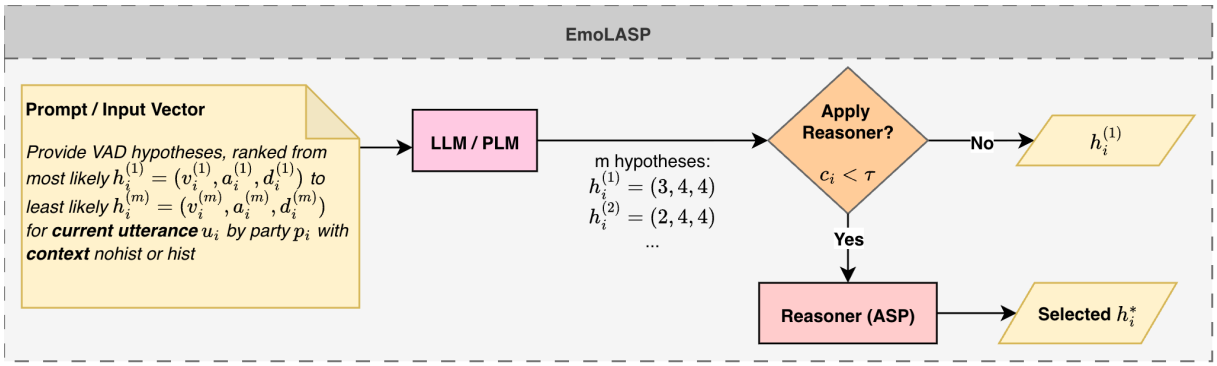


Figure 1: Overview of our EmoLASP framework.

LLMs, and from the per-dimension softmax product for PLMs). Let $p_i^{(j)}$ be the probability the model assigns to hypothesis $h_i^{(j)}$ for utterance u_i, normalised over the m hypotheses. We define the *confidence score* c_i based on Shannon entropy (Shannon, 1948) as:

$$c_i = 1 - \frac{E(p_i)}{\ln m}, \quad E(p_i) = -\sum_{j=1}^{m} p_i^{(j)} \ln p_i^{(j)}$$

Dividing the Shannon entropy $E(p_i)$ by its maximum value $\ln m$ ensures the score c_i lies in $[0, 1]$ regardless of m. We then apply the ASP reasoner when $c_i < \tau$, where threshold τ is predefined; otherwise, the top-ranked hypothesis is taken as the prediction, $\hat{y}_i = h_i^{(1)}$.

Aggregation We combine and average the *predicted* VAD scores from the most recent k turns; gold labels y_j are never used at inference time, so there is no label leakage. Formally, let the conversation history before turn i be $C_{i-1} = \{u_1, u_2, \ldots, u_{i-1}\}$ with corresponding predictions $\hat{Y}_{i-1} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{i-1}\}$. We then combine the k most recent predicted VAD scores as $A_i = \frac{1}{k} \sum_{j=i-k}^{i-1} \hat{y}_j$ so that A_i is the aggregated VAD score for turn i based on the most recent k turns.

Optimisation We use *weak constraints* in ASP to encode objectives for selecting the best VAD hypothesis based on the dialogue context:

```
:~ pick(N,H), hist_penalty(N,H,P). [P@2, N, H]
:~ pick(N,H). [H@1, N, H]
```

Intuitively, these constraints operate as follows: Priority `@2` (first preference) picks the hypothesis H for turn N that has the smallest distance P to the history aggregate A_i. Given a set of hypotheses H_i provided by the language model, let the best hypothesis be h_i^*; the distance is then measured using the L_1 norm: $h_i^* = \arg\min_{h_i^{(j)} \in H_i} \|h_i^{(j)} - A_i\|_1$ (the selected VAD hypothesis should be close to the aggregated VAD score, which captures the emotional context from previous turns). When several hypotheses share the minimal distance, priority `@1` (second preference) picks the hypothesis with the smallest index H, corresponding to the higher-ranked hypothesis from the language model.

3 Experiment

Dataset We use the IEMOCAP dataset (Busso et al., 2008), which is annotated with VAD scores that serve as labels for evaluation. This dataset contains a total of 115 dialogues across 5 dyadic conversational sessions. Each utterance in the dialogue is annotated with 5-point ordinal ratings (1 to 5) for VAD. The final VAD labels are obtained by averaging the ratings from multiple human annotators. Across 10,039 utterances, annotator counts range from 1 to 4: 8,666 have two annotations, 1,279 have three, 58 have four, and only 36 have just one.

Metrics There are two metrics used to evaluate the performance. The first is Mean Absolute Error (MAE), which quantifies the average absolute difference between model predictions and the average of human ratings (treated as continuous values). The second is Krippendorff's α, which measures how well the model agrees with the human annotators on the original ordinal ratings. To compute α, we add the model's prediction as one extra rater alongside the per-utterance human annotators. We also report Krippendorff's α among only the human annotators as a reference point for human-level agreement.

Baselines Open-source LLMs publicly available on Ollama[2] and two PLMs (BERT, RoBERTa) are

[2] https://ollama.com/

		MAE (lower is better)				Krippendorff's α (higher is better)			
		No Reasoner		(Ours) EmoLASP		No Reasoner		(Ours) EmoLASP	
Category	**Model**	nohist	hist	nohist	hist	nohist	hist	nohist	hist
Humans alone		—				$\bar{\alpha} = 0.39 \pm 0.03$			
< 10b	llama3.2:3b	0.88 ± 0.03	1.01 ± 0.02	$\mathbf{0.83 \pm 0.02}$	0.95 ± 0.01	$\mathbf{0.21 \pm 0.02}$	0.18 ± 0.02	$\mathbf{0.21 \pm 0.02}$	0.19 ± 0.02
	gemma3:4b	$0.76 \pm 0.03^{\dagger}$	0.84 ± 0.03	$\mathbf{0.74 \pm 0.03}^{\dagger}$	0.76 ± 0.04	$\mathbf{0.26 \pm 0.02}$	0.25 ± 0.02	0.23 ± 0.02	0.24 ± 0.02
20-30b	gpt-oss:20b	0.84 ± 0.02	0.82 ± 0.03	0.76 ± 0.02	$\mathbf{0.73 \pm 0.02}$	0.28 ± 0.02	$\mathbf{0.32 \pm 0.02}^{\dagger}$	0.28 ± 0.02	$\mathbf{0.32 \pm 0.02}^{\dagger}$
	gemma3:27b	0.77 ± 0.01	$\mathbf{0.75 \pm 0.02}$	0.77 ± 0.01	$\mathbf{0.75 \pm 0.01}$	0.28 ± 0.02	$\mathbf{0.30 \pm 0.02}$	0.22 ± 0.02	0.26 ± 0.03
> 60b	llama3.3:70b	0.96 ± 0.02	0.91 ± 0.04	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.81 \pm 0.03}$	$0.29 \pm 0.02^{\dagger}$	$\mathbf{0.32 \pm 0.02}^{\dagger}$	$0.29 \pm 0.02^{\dagger}$	$\mathbf{0.32 \pm 0.02}^{\dagger}$
	gpt-oss:120b	0.91 ± 0.02	0.85 ± 0.03	0.75 ± 0.01	$\mathbf{0.71 \pm 0.02}^{\dagger}$	0.27 ± 0.02	$\mathbf{0.31 \pm 0.02}$	0.27 ± 0.02	0.30 ± 0.02
PLM	bert	0.78 ± 0.03	0.74 ± 0.03	0.77 ± 0.03	$\mathbf{0.73 \pm 0.03}$	$0.29 \pm 0.03^{\dagger}$	$\mathbf{0.30 \pm 0.02}$	$0.29 \pm 0.03^{\dagger}$	$\mathbf{0.30 \pm 0.02}$
	roberta-large	0.78 ± 0.02	$0.72 \pm 0.04^{\dagger}$	0.77 ± 0.02	$\mathbf{0.71 \pm 0.05}^{\dagger}$	0.28 ± 0.02	$\mathbf{0.32 \pm 0.03}^{\dagger}$	$0.29 \pm 0.02^{\dagger}$	$\mathbf{0.32 \pm 0.03}^{\dagger}$

Table 1: MAE (left) and Krippendorff's α (right). *Humans alone* is α over only the human annotators. † = best in the same column; **bold** = best in the same row within its metric block; *gray highlight* = handicap comparison.

		MAE			Krippendorff's α		
Category	**Model**	nohist	hist	handicap	nohist	hist	handicap
< 10b	llama3.2:3b	$\mathbf{-0.049}^{***}$	$\mathbf{-0.062}^{***}$	$\mathbf{-0.174}^{***}$	-0.002	$+0.007$	$\mathbf{+0.024}^{**}$
	gemma3:4b	-0.014	$\mathbf{-0.075}^{**}$	$\mathbf{-0.093}^{***}$	-0.029^{**}	-0.010	-0.016
20-30b	gpt-oss:20b	$\mathbf{-0.078}^{***}$	$\mathbf{-0.091}^{***}$	$\mathbf{-0.062}^{**}$	$+0.002$	-0.002	-0.036^{***}
	gemma3:27b	$+0.002$	$+0.001$	$+0.022$	-0.061^{***}	-0.047^{***}	-0.079^{***}
> 60b	llama3.3:70b	$\mathbf{-0.153}^{***}$	$\mathbf{-0.094}^{***}$	$\mathbf{-0.097}^{***}$	$+0.000$	$+0.000$	-0.025^{**}
	gpt-oss:120b	$\mathbf{-0.158}^{***}$	$\mathbf{-0.133}^{***}$	$\mathbf{-0.096}^{***}$	-0.004	-0.004	-0.039^{***}
PLM	bert	$\mathbf{-0.015}^{***}$	-0.009	$+0.030$	$\mathbf{+0.006}^{*}$	-0.002	-0.012
	roberta-large	$\mathbf{-0.014}^{**}$	-0.005	$+0.051^{*}$	$\mathbf{+0.004}^{*}$	-0.003	-0.029^{*}

Table 2: Paired t-test of **No Reasoner** vs. **(Ours) EmoLASP** on the IEMOCAP dataset. Each cell reports $\Delta = \overline{\text{EmoLASP}} - \overline{\text{No Reasoner}}$, the mean score difference over the paired units. For MAE a negative Δ favours EmoLASP; for α a positive Δ favours EmoLASP. Significance is from a two-sided paired t-test: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. **Bold** marks a statistically significant improvement by EmoLASP.

used as baselines for comparison. For each language model, there are two context settings: (1) *nohist*: prompt/input vector has only the current utterance; (2) *hist*: prompt/input vector includes all previous conversation turns and the current utterance. Since there are five different sessions of the IEMOCAP dataset, for PLMs we test on each session and use the remaining four sessions for training. For LLMs, we evaluate on each session separately. Results are reported as the mean and standard deviation across five sessions. Prompts and hyperparameters are detailed in Appendix C.

4 Results

RQ1: Can EmoLASP improve ERC performance compared to the LM alone? EmoLASP achieves comparable-to-better MAE at substantially lower input-token cost, which is concentrated in the prompt-only LLM pipeline. Table 2 reports paired t-tests for every cell of Table 1. Over the 24 MAE comparisons, 16 show a significant improvement of EmoLASP over the No Reasoner baseline, 7 show no significant difference, and only 1 shows a significant decrease. Splitting these by pipeline (LLMs vs. PLMs) makes things clearer. For the six prompt-only LLMs (18 comparisons), 14 show significant improvements and *none* shows a significant decrease; the 4 non-significant cells are all three comparisons for `gemma3:27b`, plus `gemma3:4b` under *nohist*. For the two fine-tuned PLMs (6 comparisons), only the two *nohist* cells improve significantly, and the single significant decrease in the whole table is `roberta-large` under *handicap*.

The *handicap* comparison, highlighted in Table 1, is the key one for the cost claim: it pits the No Reasoner baseline with full dialogue history in the prompt (*hist*) against EmoLASP with no history at all (*nohist*), so any win cannot be attributed to dialogue context. For LLMs, EmoLASP (*nohist*) significantly beats the full-context baseline in 5 of 6 cases, with the remaining case (`gemma3:27b`) showing no significant difference. For example, MAE drops from 0.91 to 0.81 for `llama3.3:70b` ($p < 0.001$), even though the LLM sees only the current utterance. This matters because EmoLASP (*nohist*) uses fewer input tokens than the No Reasoner baseline (*hist*) and requires no fine-tuning.

In contrast, the PLMs behave differently: the reasoner brings no benefit under *handicap* for `bert` ($p > 0.05$) and significantly hurts `roberta-large` ($0.72 \rightarrow 0.77, p < 0.05$), indicating that fine-tuned PLMs already make effective use of dialogue history in their input. We therefore restrict the cost claim to the prompt-only LLM pipeline, where EmoLASP matches or beats a history-fed baseline without fine-tuning and with fewer input tokens.

The results for Krippendorff's α tell a slightly different story. The human inter-annotator agreement is 0.39, while the best model α is 0.32 (e.g., `gpt-oss:20b`, `roberta-large` with *hist*). This means that no model reaches human-level agreement yet, but larger LLMs and PLMs come closer than the smaller ones. Across the 24 α comparisons, EmoLASP leaves agreement statistically unchanged in 13, improves it in 3, and decreases it in 8. These decreases are not spread evenly: 3 come from `gemma3:27b` under all three comparisons, 4 more fall in the *handicap* column, where EmoLASP is denied the dialogue history that the baseline receives, and the last is `gemma3:4b` under *nohist*. Setting the two `gemma3` models aside, every matched-context comparison (*nohist* and *hist*) is therefore either unchanged or an improvement, so the MAE gains above do not come at the cost of agreement with human annotators.

RQ2: Which utterances have high VAD prediction errors across all LLMs? In this section, we identify some examples of utterances that are challenging to predict. A longer list of examples can be found in Appendix D.3. A clear pattern appearing across these cases is that while the ground-truth annotations reflect extreme emotions in the VAD space (1 or 5), every model predicts values close to the middle (around 2 to 4). These issues have two main causes. The first involves very short or surface-neutral utterances. One example utterance is *"Hmm"* with human-annotated scores (5, 4, 0.5), meaning the speaker feels very positive, yet every model predicts mid-range valence values from 2 to 3. In another example (*"Are you cold, huh? Do you want to go home?"*), the utterance looks like a caring question, but the gold label (4.50, 4.50, 1.00) reflects strong feeling with very low dominance, which all models miss.

The second cause is figurative or metaphorical language. In one case (*"...I'd like a family. I'd like some kids."*), the label is (1, 4.5, 5), expressing deep frustration, but models read the words "family" and "kids" as positive and predict high valence (around 4 to 5). Another example (*"God I feel as though my entire life is going to be spent standing on the beach...waiting for the fish to show up."*) uses the fish as a metaphor for life-long disappointment, with label (1, 5, 5). Some models (e.g., `gpt-oss:120b`, `gpt-oss:20b`) correctly pick up the low valence and high arousal, but they underestimate the dominance by predicting 1 instead of 5. ASP does not change this much because LLMs are responsible for generating the VAD hypotheses.

Central-tendency shrinkage A fair concern is that the MAE gains reflect central-tendency shrinkage, that is, predictions pulled toward the middle of the VAD scale, rather than better emotion recognition. However, three points argue against pure shrinkage. First, the reasoner can only pick among the m hypotheses proposed by the language model, so it cannot regress toward a midpoint that is not already a candidate. It is also invoked only when the model is not confident, leaving confident predictions untouched. Second, if the mechanism were pure smoothing, Krippendorff's α should degrade uniformly. Instead, α is unchanged in 13 of 24 comparisons and improves in 3, and the 8 significant decreases are concentrated in one model (`gemma3:27b`) and in the *handicap* comparison, where the reasoner is not given the dialogue history that the baseline receives in its prompt. Third, RQ2 shows that every tested model already predicts mid-range values (around 2 to 4) for utterances annotated at the extremes (1 or 5), and ASP cannot help in these cases because the extreme values never appear among the hypotheses it selects from.

5 Conclusions

We presented EmoLASP, which adds an ASP reasoner on top of a language model. Our system improves MAE for most prompt-only LLMs, while the fine-tuned PLMs gain little as they already use the dialogue history well. EmoLASP beats a history-fed baseline for 5 of 6 LLMs while seeing only the current utterance, so a small reasoner can replace a long history in the prompt at a lower input-token cost and with no fine-tuning. Agreement with human annotators mostly stays the same, so these gains do not come at the cost of agreement. However, qualitative analysis reveals a limitation of LLM-based approaches: they struggle to accurately interpret extreme emotions in brief utterances or those relying on metaphorical language.

Limitations

Our study has several limitations, which concern the data, the framework, and our evaluation. First, regarding the data, we evaluate EmoLASP on a single dataset (IEMOCAP). To the best of our knowledge, this is the only publicly available dataset with human VAD annotations at the dialogue level. While there are other datasets with VAD annotations, such as EmoBank (Buechel and Hahn, 2017) or NRC VAD (Mohammad, 2025), they are sentence-level or word-level datasets. In the IEMOCAP dataset, human annotators annotate the video data, which includes both audio and visual cues, but our models only have access to the text data. This mismatch may contribute to the disagreement between model predictions and human annotations, and it also means that our models are not fully leveraging the multimodal nature of the data. Also, the number of annotators per utterance varies across utterances, with the majority (8,666 out of 10,039) having only two annotators.

Second, the framework itself has limitations. The current ASP implementation is modest with a choice rule and two weak constraints for optimisation. In future work, we plan to explore more elaborate ASP implementations. The aggregation window k, the number of hypotheses m, and the confidence threshold τ are also fixed to $k = 2$, $m = 3$, and $\tau = 0.25$ in our main experiments. Our ablation studies in Appendix D.2 show that the results barely change across a range of these values, and that weighting the history by speaker identity does not help either. However, these sweeps vary one parameter at a time on a single dataset, so the best settings for other datasets or for a joint search may still differ. Furthermore, our analysis in Section 4 shows that the framework still struggles with very short utterances (e.g., *"Hmm"*) and figurative language. Addressing this issue likely requires changes to how hypotheses are generated, which depends on the language models, not just how they are selected by the ASP reasoner.

Third, our evaluation is limited in scope. We currently use fine-tuned PLMs as supervised VAD baselines that are representative of the encoder-based approach and serve as fair points of comparison for the LLM pipeline. Previous ERC models, such as DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019) and SKIER (Li et al., 2023), predict categorical emotion labels and are not directly comparable on VAD dimensions. Future work could explore the use of these models as baselines by mapping their categorical predictions to VAD scores. Moreover, our PLMs use three independent classification heads over a 5-point Likert scale, which assumes independence across V, A, and D. Finally, all evaluated LLMs are open-source models available through Ollama, and we do not test closed-source models such as GPT-5, so the conclusions may not generalise to all LLMs.

Ethics Considerations

We use the IEMOCAP dataset (Busso et al., 2008) for research purposes only, consistent with the terms of its license. Our experiments use only the text transcriptions of utterances and do not process the audio or video recordings. The transcriptions provided by the dataset have been de-identified. In terms of demographics, the corpus consists of acted dyadic dialogues in English, recorded in a single cultural and institutional setting (a US university drama department). The speakers are ten actors (five male, five female) whose demographic characteristics beyond gender are not documented in the original release.

Potential Risks Emotion recognition systems carry some risks. First, dimensional VAD predictors can be repurposed for affect-based advertising or surveillance without user consent, even though our motivating applications are empathetic dialogue and mental health support. Second, our qualitative analysis (Section 4) shows that most evaluated LLMs systematically compress predictions toward the middle of the VAD scale and miss extreme states conveyed through short or figurative utterances. In safety-critical settings such as mental health support, this could cause systems to underestimate distress precisely when accurate recognition matters most, a limitation that EmoLASP does not overcome. Third, IEMOCAP consists of acted English dyadic dialogues from a single cultural setting, which limits generalisation to other populations and languages.

Acknowledgements

This research was partially supported by the Australian Research Council (grant #DP250101822). The experiments were conducted on Katana (UNSW Sydney, 2010).

References

Mario Alviano, Lorenzo Grillo, Fabrizio Lo Scudo, and Luis Angel Rodriguez Reiners. 2025. Integrating Answer Set Programming and Large Language Models for Enhanced Structured Representation of Complex Knowledge in Natural Language. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 4330–4338.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Margaret M Bradley and Peter J Lang. 1999. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical report, Technical report C-1, the center for research in psychophysiology.

Andreas Brännström and Juan Carlos Nieves. 2022. Emotional Reasoning in an Action Language for Emotion-Aware Planning. In *International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 103–116. Springer.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards Multimodal Emotional Support Conversation Systems. *IEEE Transactions on Multimedia*, 27:8276–8287.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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 4171–4186.

Thomas Eiter, Tobias Geibinger, Nelson Higuera, and Johannes Oetsch. 2023. A Logic-based Approach to Contrastive Explainability for Neurosymbolic Visual Question Answering. In *IJCAI*, pages 3668–3676.

Thomas Eiter, Nelson Higuera, Johannes Oetsch, and Michael Pritz. 2022. A Neuro-Symbolic ASP Pipeline for Visual Question Answering. *Theory and Practice of Logic Programming*, 22(5):739–754.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. 2024. Affect Recognition in Conversations Using Large Language Models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 259–273.

Mauricio J Osorio Galindo, Luis A Montiel Moreno, David Rojas-Velázquez, and Juan Carlos Nieves. 2020. E-Friend: A Logical-Based AI Agent System Chat-Bot for Emotional Well-Being and Mental Health. In *International Workshop on Deceptive AI*, pages 87–104. Springer.

Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. 2019. Multi-shot ASP solving with clingo. *Theory and Practice of Logic Programming*, 19(1):27–82.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Er Jin, Qihui Feng, Yongli Mou, Gerhard Lakemeyer, Stefan Decker, Oliver Simons, and Johannes Stegmaier. 2025. LogicAD: Explainable Anomaly Detection via VLM-based Text Feature Extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4129–4137.

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. SKIER: A Symbolic Knowledge Integrated Model for Conversational Emotion Recognition. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 13121–13129.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers. In *European Conference on Information Retrieval*, pages 84–100. Springer.

Saif M Mohammad. 2025. NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English Terms. *arXiv preprint arXiv:2503.23547*.

Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro–Symbolic Reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204.

Andrew Ortony, Gerald L Clore, Allan Collins, and 1 others. 1988. *The Cognitive Structure of Emotions*, volume 18. Cambridge university press Cambridge, UK.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional Emotion Detection from Categorical Emotion. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 4367–4380.

PVC (Research Infrastructure), UNSW Sydney. 2010. Katana.

Abhiramon Rajasekharan, Yankai Zeng, Parth Padalkar, and Gopal Gupta. 2023. Reliable Natural Language Understanding with Large Language Models and Answer Set Programming. *ICLP*.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Claude Elwood Shannon. 1948. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):379–423.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Bias in Emotion Recognition with ChatGPT. *arXiv preprint arXiv:2310.11753*.

Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling Large Language Models with Logic Programming for Robust and General Reasoning from Text. In *Findings of the association for computational linguistics: ACL 2023*, pages 5186–5219.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2025. Beyond Verbal Cues: Emotional Contagion Graph Network for Causal Emotion Entailment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1755–1767.

Yankai Zeng, Abhiramon Rajasekharan, Kinjal Basu, Huaduo Wang, Joaquín Arias, and Gopal Gupta. 2024. A Reliable Common–Sense Reasoning Socialbot Built Using LLMs and Goal-Directed ASP. *Theory and Practice of Logic Programming*, 24(4):606–627.

A AI Assistants in Research or Writing

We use AI assistants, including Claude and ChatGPT, for coding assistance, grammar checks and stylistic improvements.

B Extended Background

In this section, we provide more details about other related work on applying answer set programming (ASP) to affective computing. Galindo et al. (2020) introduced an E-friend chatbot for mental well-being that maintains a model of the user's emotion and mental state using logic rules. In their system, psychological theories such as the Ortony, Clore, and Collins (OCC) model (Ortony et al., 1988) are represented as ASP rules. The ASP reasoning module helps the chatbot decide on strategies such as asking questions or suggesting activities to improve the user's emotional state. Another example is the work by Brännström and Nieves (2022) on emotion-aware agent planning, in which the authors formalise emotional dynamics as an action language and translate it to ASP for reasoning about how an agent's actions influence human emotions. Moreover, Eiter et al. (2023) also took advantage of weak constraints in ASP, similar to our approach, but in the context of generating counterfactual explanations in the visual question answering domain. This approach uses ASP to provide explanations given an image and a user query; in contrast, we use ASP to select among emotion hypotheses produced by a language model. They nevertheless highlight that ASP can be powerful for ensuring logical consistency and integrating human knowledge about emotions.

Moreover, LM-ASP pipelines have been explored in Natural Language Understanding (NLU) tasks, but to our knowledge, not yet for emotion recognition. For example, a common approach is to use LLMs to translate the problem description and user query into ASP syntax, then apply an ASP solver to perform reasoning tasks (Yang et al., 2023; Alviano et al., 2025). Another line of work uses LLMs to generate candidate continuations of a story, and ASP for filtering out illogical continuations (Nye et al., 2021). In handling dia-

logues, the LM-ASP combination has been used in the task-oriented chatbot domain (Rajasekharan et al., 2023; Zeng et al., 2024), where LLMs are used to extract knowledge from the dialogue, and ASP is executed to decide on the next action of the chatbot, such as recommending a restaurant or answering a question. However, those works are restricted to movie and restaurant recommendation domains and have not been applied to emotion understanding. Unlike pipelines that translate the full problem into ASP and let the solver do the heavy reasoning, EmoLASP uses ASP specifically for its choice rules and weak constraints with priorities, which let us encode preferences over LM-produced VAD hypotheses while still leveraging the language model for semantic interpretation.

In the ERC domain, several related works have used purely neural or neuro-symbolic approaches. Majumder et al. (2019) applied an RNN to track individual speaker states for emotion classification. Ghosal et al. (2019) presented a graph convolutional network (GCN) to model the inter-speaker dependency. Additionally, Yu et al. (2025) developed a graph neural network to simulate the emotional contagion process (Hatfield et al., 1993). Furthermore, to integrate knowledge beyond the conversation context, Li et al. (2023) utilised an external commonsense knowledge base to understand relations between two different concepts and hence infer the emotion class of an utterance from its context. Unlike these neural approaches that predict categorical emotion labels, EmoLASP targets VAD dimensions and uses ASP to enforce contextual consistency over LM-produced hypotheses, without requiring the extensive neural-network training that these methods rely on; such training is outside the scope of our paper.

C Experiment Details

C.1 Licenses

Our code is publicly released under the MIT License to allow reproducibility and future research[3]. The dataset used in this paper, IEMOCAP (Busso et al., 2008), is shared for research purposes only.

C.2 Hyperparameters

Our experiments were conducted on a shared computing infrastructure equipped with NVIDIA V100, A100, and H200 GPUs. We used a single GPU per job. A total of 8 (models) $\times$ 2 (contexts: nohist, hist) $\times$ 5 (dialogue sessions) = 80 jobs were run.

When training PLMs, we set the learning rate to 2e-5, batch size to 16, max context length to 512, and trained for 20 epochs. There are two PLMs: BERT has $\sim$ 110 million parameters, and RoBERTa-large has $\sim$ 355 million parameters. For each model, we ran 10 jobs (2 contexts $\times$ 5 dialogue sessions). We estimate the total running time to be approximately 10 hours for the two models.

For LLMs, we set the temperature to 0, and the number of VAD hypotheses m to 3. For ASP, we set the confidence threshold τ to 0.25 and the aggregation window size k to 2. We use six LLMs with sizes ranging from 3 billion to 120 billion parameters. For each model, we ran 10 jobs (2 contexts $\times$ 5 dialogue sessions). The total running time is estimated to be approximately 170 hours for all six LLMs.

Why $\tau = 0.25$? Specifically, with $m = 3$, let the highest probability for utterance u_i be $p_i^* = 0.7$ and assume the two other hypotheses have the same probability of 0.15 (highest uncertainty); the corresponding Shannon entropy is $E(p_i) = -0.7 \ln 0.7 - 0.15 \ln 0.15 - 0.15 \ln 0.15 \approx 0.82$. The confidence score is then $c_i = 1 - \frac{E(p_i)}{\ln m} = 1 - \frac{0.82}{\ln 3} \approx 0.25$. Therefore, setting $\tau = 0.25$ means that we apply ASP when the language model's best hypothesis has a probability lower than 0.7.

[3] https://github.com/thaole25/emolasp

C.3 Prompt Templates

We provide the full prompt templates for LLM hypothesis generation in Figures 2 and 3. The only difference between the two prompts is that the `hist` prompt includes the dialogue context, while the `nohist` prompt does not. The 5-point Likert scale for VAD is based on the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994, 1999).

```
Analyze the emotion expressed in the following
     text and provide 3 plausible Valence,
    Arousal, and Dominance (VAD) hypotheses,
    ranked from most likely (H1) to least
    likely (H3).

Each score must be an integer from 1 to 5 (
    inclusive). Use only whole numbers 1, 2,
    3, 4, or 5.

**Valence (V)** - happiness vs unhappiness:
- 1 = Complete unhappiness, annoyance,
    unsatisfaction, melancholia, or boredom.
- 2 = Mostly negative, low pleasure.
- 3 = Neutral; neither clear happiness nor
    sadness.
- 4 = Mostly positive, moderate pleasure.
- 5 = Complete happiness, pleasure,
    satisfaction, contentment, or hope.

**Arousal (A)** - calm vs excited:
- 1 = Complete relaxation, calmness,
    sluggishness, dullness, sleepiness, or
    lack of excitement.
- 2 = Mostly calm, low energy.
- 3 = Neutral; neither calm nor excited.
- 4 = Mostly stimulated, moderate energy.
- 5 = Complete stimulation, excitement, frenzy,
     nervousness, or high arousal.

**Dominance (D)** - controlled vs in-control:
- 1 = Complete feeling of being controlled,
    influenced, cared-for, awed, submissive,
    or guided.
- 2 = Mostly submissive, low control.
- 3 = Neutral; neither controlling nor
    controlled.
- 4 = Mostly in-control, moderate dominance.
- 5 = Complete feeling in-control, influential,
     important, dominant, autonomous, or
    controlling.

Text: "<UTTERANCE>"

Return exactly 3 ranked hypotheses, one per
    line, with H1 the most likely and H3 the
    least likely:
H1: V=X, A=X, D=X
H2: V=X, A=X, D=X
H3: V=X, A=X, D=X
(X must be one of 1, 2, 3, 4, or 5)
Your output:
```

Figure 2: LLM hypothesis-generation prompt without dialogue context (`nohist`).

```
Analyze the emotion expressed in the current
    text based on the dialogue context and
    provide 3 plausible Valence, Arousal, and
    Dominance (VAD) hypotheses,
ranked from most likely (H1) to least likely (
    H3).

Each score must be an integer from 1 to 5 (
    inclusive). Use only whole numbers 1, 2,
    3, 4, or 5.

**Valence (V)** - happiness vs unhappiness:
- 1 = Complete unhappiness, annoyance,
    unsatisfaction, melancholia, or boredom.
- 2 = Mostly negative, low pleasure.
- 3 = Neutral; neither clear happiness nor
    sadness.
- 4 = Mostly positive, moderate pleasure.
- 5 = Complete happiness, pleasure,
    satisfaction, contentment, or hope.

**Arousal (A)** - calm vs excited:
- 1 = Complete relaxation, calmness,
    sluggishness, dullness, sleepiness, or
    lack of excitement.
- 2 = Mostly calm, low energy.
- 3 = Neutral; neither calm nor excited.
- 4 = Mostly stimulated, moderate energy.
- 5 = Complete stimulation, excitement, frenzy,
     nervousness, or high arousal.

**Dominance (D)** - controlled vs in-control:
- 1 = Complete feeling of being controlled,
    influenced, cared-for, awed, submissive,
    or guided.
- 2 = Mostly submissive, low control.
- 3 = Neutral; neither controlling nor
    controlled.
- 4 = Mostly in-control, moderate dominance.
- 5 = Complete feeling in-control, influential,
     important, dominant, autonomous, or
    controlling.

Dialogue Context:
Person A: <PRIOR UTTERANCE 1>
Person B: <PRIOR UTTERANCE 2>
...

Current Text: "<SPEAKER LABEL>: <UTTERANCE>"

Return exactly 3 ranked hypotheses, one per
    line, with H1 the most likely and H3 the
    least likely:
H1: V=X, A=X, D=X
H2: V=X, A=X, D=X
H3: V=X, A=X, D=X
(X must be one of 1, 2, 3, 4, or 5)
Your output:
```

Figure 3: LLM hypothesis-generation prompt with dialogue context (`hist`). All prior same-dialogue turns are included, each prefixed with its `Person A` or `Person B` speaker label.

Ablation	Setting	**MAE** (lower is better)		**Krippendorff's** α (higher is better)	
		nohist	hist	nohist	hist
k (history window)	$k = 1$	0.79 ± 0.03	0.78 ± 0.07	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.05}$
	$\underline{k = 2}$	0.78 ± 0.03	0.77 ± 0.07	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.04}$
	$k = 3$	$\mathbf{0.77 \pm 0.03}$	0.76 ± 0.07	$\mathbf{0.26 \pm 0.04}$	0.27 ± 0.05
	$k = 5$	$\mathbf{0.77 \pm 0.03}$	$\mathbf{0.75 \pm 0.07}$	$\mathbf{0.26 \pm 0.04}$	$\mathbf{0.28 \pm 0.04}$
m (# hypotheses)	$\underline{m = 3}$	0.78 ± 0.03	0.77 ± 0.07	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.04}$
	$m = 5$	$\mathbf{0.76 \pm 0.06}$	0.76 ± 0.09	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.05}$
	$m = 7$	$\mathbf{0.76 \pm 0.08}$	0.76 ± 0.09	$\mathbf{0.26 \pm 0.04}$	0.27 ± 0.05
	$m = 9$	$\mathbf{0.76 \pm 0.08}$	$\mathbf{0.75 \pm 0.09}$	0.25 ± 0.04	0.27 ± 0.05
τ (confidence gate)	$\tau = 0.1$	0.79 ± 0.03	0.78 ± 0.08	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.05}$
	$\underline{\tau = 0.25}$	0.78 ± 0.03	0.77 ± 0.07	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.04}$
	$\tau = 0.4$	$\mathbf{0.77 \pm 0.03}$	0.77 ± 0.08	$\mathbf{0.26 \pm 0.04}$	$\mathbf{0.28 \pm 0.04}$
	$\tau = 0.6$	$\mathbf{0.77 \pm 0.04}$	$\mathbf{0.76 \pm 0.08}$	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.04}$

Table 3: Ablation of the EmoLASP reasoner hyperparameters. Each row holds two of $\{k, m, \tau\}$ at the anchor $(k = 2, m = 3, \tau = 0.25)$ (underlined) and sweeps the third. Cells aggregate across 6 LLMs + 2 PLMs by first averaging across the 5 cross-session splits per model, then reporting mean $\pm$ std across the 8 per-model means. **Bold** marks the best value within each parameter group and column.

w_{self}:w_{other}	**MAE** (lower is better)		**Krippendorff's** α (higher is better)	
	nohist	hist	nohist	hist
1:3	$\mathbf{0.78 \pm 0.03}$	0.78 ± 0.07	$\mathbf{0.26 \pm 0.03}$	0.27 ± 0.05
1:2	$\mathbf{0.78 \pm 0.03}$	0.78 ± 0.07	$\mathbf{0.26 \pm 0.04}$	0.27 ± 0.05
$\underline{1{:}1}$ *(neutral)*	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.77 \pm 0.07}$	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.04}$
2:1	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.77 \pm 0.07}$	$\mathbf{0.26 \pm 0.04}$	$\mathbf{0.28 \pm 0.05}$
3:1	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.77 \pm 0.07}$	$\mathbf{0.26 \pm 0.03}$	$\mathbf{0.28 \pm 0.04}$

Table 4: Ablation of the EmoLASP speaker-weighting ratio w_{self}:w_{other}, holding the other reasoner hyperparameters at the anchor $(k = 2, m = 3, \tau = 0.25)$. The 1:1 row (underlined) is the speaker-blind default. Cells aggregate across 6 LLMs + 2 PLMs by first averaging across the 5 cross-session splits per model, then reporting mean $\pm$ std across the 8 per-model means. **Bold** marks the best value per column.

D Results

D.1 Speaker-Weighted Aggregation

The ablation in Table 4 changes only the aggregation step of Section 2.3. In the main method, the aggregate A_i weights the k most recent turns equally. In the speaker-weighted variant, each history turn j is weighted by whether its speaker p_j is the speaker p_i of the target turn:

$$w_j = \begin{cases} w_{\text{self}} & \text{if } p_j = p_i \\ w_{\text{other}} & \text{if } p_j \neq p_i \end{cases}$$

and the aggregate becomes the weighted mean

$$A_i^{\text{sw}} = \frac{1}{W_i} \sum_{j=i-k}^{i-1} w_j \, \hat{y}_j, \quad W_i = \sum_{j=i-k}^{i-1} w_j.$$

Setting $w_{\text{self}} > w_{\text{other}}$ pulls the aggregate toward the speaker's own previous turns, and $w_{\text{other}} > w_{\text{self}}$ toward the interlocutor. The optimisation step is unchanged apart from using A_i^{sw} in place of A_i:

$$\hat{y}_i = \arg \min_{h_i^{(j)} \in H_i} \| h_i^{(j)} - A_i^{\text{sw}} \|_1,$$

with ties again broken by the higher-ranked hypothesis.

D.2 Ablation Studies

In this section, we present ablation studies to investigate the impact of the hyperparameters: the history window size k, the number of hypotheses m, and the confidence threshold τ. We also add speaker-weighted aggregation variants to explore whether weighting the history aggregate by speaker identity influences performance. The results are reported in Tables 3 and 4.

In Table 3, each sweep changes one parameter and keeps the other two at the default setting $(k = 2, m = 3, \tau = 0.25)$. The main finding is that EmoLASP is stable: across all twelve settings, MAE stays between 0.75 and 0.79 and Krippendorff's α between 0.25 and 0.28. These gaps are much smaller than the standard deviation across models (up to ± 0.09), so no setting is clearly better than another. Looking at a longer history ($k = 5$) or asking for more hypotheses ($m = 9$) lowers MAE by about 0.02, but α does not improve and even drops slightly in some cases. In other words, a small gain in average error does not mean better agreement with the human ratings. Changing τ has the smallest effect of all: raising it from 0.1 to 0.6 makes the reasoner run on many more turns, yet the scores barely move, which shows that applying ASP more often does not hurt performance.

Table 4 tests whether the history aggregate should treat the two speakers differently. Instead of averaging the last k turns equally, we give weight w_{self} to turns from the same speaker and w_{other} to turns from the interlocutor, as described in Appendix D.1. The results are almost identical for every ratio: MAE is 0.78 without history and 0.77–0.78 with history, and α is 0.26–0.28. The $w_{\text{self}} > w_{\text{other}}$ settings (2:1 and 3:1) match the speaker-blind default, while the $w_{\text{other}} > w_{\text{self}}$ settings (1:2 and 1:3) are worse on α with history (0.27 vs. 0.28). Since no ratio gives a real gain, we keep the simpler speaker-blind 1:1 aggregate.

Overall, these two ablations suggest that EmoLASP is not significantly affected by hyperparameter tuning, which is useful in practice because the reasoner has no trained parameters. We therefore keep the default $(k = 2, m = 3, \tau = 0.25)$ with equal speaker weights.

D.3 Qualitative Analysis

Table 5 shows the five highest-error cases by MAE where all LLMs fail to predict accurate VAD scores even after applying ASP. In this analysis, we focus on LLMs rather than PLMs, because it is unfair to compare LLMs directly with PLMs that have been fine-tuned on the dataset.

#	Context Summary	Current Utterance	No Reasoner (hist)	EmoLASP (hist)	Label (V, A, D)
1	**Summary:** Person B tells Person A that Person C (the father) is still obsessed with missing Larry. Person B insists it's dishonest to pretend Larry is alive and reveals he wants to marry Annie, who is Larry's former girlfriend. However, Person A fears this will devastate Person C and disrupt the family and business. **Two prior turns:** Person B: The business? The business doesn't inspire me. Person A: Must you be inspired?	Person B: Yeah, I'd like one hour a day. If I have to grub for money all day long, I'd like to go home to something beautiful. I'd like a family. I'd like some kids.	gemma3:4b: (4.00, 4.00, 4.00) gemma3:27b: (4.00, 2.00, 3.00) gpt-oss:20b: (5.00, 3.00, 3.00) gpt-oss:120b: (4.00, 2.00, 2.00) llama3.2:3b: (4.00, 3.00, 4.00) llama3.3:70b: (4.00, 3.00, 4.00)	gemma3:4b: (4.00, 4.00, 4.00) gemma3:27b: (4.00, 2.00, 3.00) gpt-oss:20b: (5.00, 3.00, 3.00) gpt-oss:120b: (3.00, 3.00, 2.00) llama3.2:3b: (4.00, 3.00, 3.00) llama3.3:70b: (4.00, 3.00, 4.00)	(1.00, 4.50, 5.00)
2	**Summary:** Person A and Person B are a married couple on a beach at night waiting for grunion. They argue about unmet expectations in their marriage and past disappointments but then reassure each other of their commitment. They share a quiet, intimate moment, and decide to stay together in that spot, hoping the fish might come. **Two prior turns:** Person A: Or not. Person B: Or not.	Person A: Hmm.	gemma3:4b: (2.00, 4.00, 3.00) gemma3:27b: (2.00, 2.00, 2.00) gpt-oss:20b: (2.00, 2.00, 2.00) gpt-oss:120b: (2.00, 2.00, 2.00) llama3.2:3b: (3.00, 3.00, 4.00) llama3.3:70b: (3.00, 2.00, 3.00)	gemma3:4b: (2.00, 4.00, 3.00) gemma3:27b: (3.00, 2.00, 3.00) gpt-oss:20b: (2.00, 2.00, 2.00) gpt-oss:120b: (3.00, 2.00, 3.00) llama3.2:3b: (3.00, 2.00, 3.00) llama3.3:70b: (3.00, 2.00, 3.00)	(5.00, 4.00, 0.50)
3	**Summary:** Person A and Person B are a married couple on a beach waiting to see fish that never seem to appear. They argue about disappointment and unmet expectations in their lives and relationship. They then gradually reassure each other that despite feeling let down by how things have turned out, they still want to be together. **Two prior turns:** Person B: It's not champagne. Person A: I guess we don't need glasses.	Person B: Are you cold, huh? Do you want to go home?	gemma3:4b: (2.00, 4.00, 3.00) gemma3:27b: (3.00, 2.00, 4.00) gpt-oss:20b: (4.00, 2.00, 3.00) gpt-oss:120b: (4.00, 2.00, 3.00) llama3.2:3b: (2.00, 2.00, 2.00) llama3.3:70b: (3.00, 2.00, 4.00)	gemma3:4b: (2.00, 4.00, 3.00) gemma3:27b: (3.00, 2.00, 4.00) gpt-oss:20b: (4.00, 2.00, 3.00) gpt-oss:120b: (3.00, 3.00, 3.00) llama3.2:3b: (2.00, 2.00, 2.00) llama3.3:70b: (3.00, 2.00, 4.00)	(4.50, 4.50, 1.00)
4	**Summary:** Person A and Person B are arguing on a beach because Person B is excited about seeing grunion fish while Person A is frustrated and dismisses it as PR hype. Person A insists the outing isn't really about the fish, and complains that they never actually see them. **Two prior turns:** Person B: We'll see them this year. Person A: No, we won't. It's pointless. It's like waiting up to see Santa Clause.	Person A: God I feel as though my entire life is going to be spent standing on the beach with my eyes wide open and my hands clasped expectantly waiting for the fish to show up.	gemma3:4b: (2.00, 4.00, 3.00) gemma3:27b: (2.00, 3.00, 2.00) gpt-oss:20b: (2.00, 4.00, 1.00) gpt-oss:120b: (1.00, 3.00, 1.00) llama3.2:3b: (4.00, 5.00, 2.00) llama3.3:70b: (2.00, 4.00, 1.00)	gemma3:4b: (2.00, 4.00, 3.00) gemma3:27b: (2.00, 3.00, 2.00) gpt-oss:20b: (1.00, 3.00, 2.00) gpt-oss:120b: (2.00, 4.00, 2.00) llama3.2:3b: (4.00, 5.00, 2.00) llama3.3:70b: (2.00, 4.00, 1.00)	(1.00, 5.00, 5.00)
5	This is the first turn in the conversation; there are no previous turns.	Person A: Look at this, goose bumps .	gemma3:4b: (4.00, 3.00, 4.00) gemma3:27b: (4.00, 4.00, 3.00) gpt-oss:20b: (4.00, 5.00, 3.00) gpt-oss:120b: (4.00, 5.00, 3.00) llama3.2:3b: (5.00, 5.00, 5.00) llama3.3:70b: (5.00, 5.00, 4.00)	gemma3:4b: (4.00, 3.00, 4.00) gemma3:27b: (4.00, 4.00, 3.00) gpt-oss:20b: (4.00, 5.00, 3.00) gpt-oss:120b: (4.00, 5.00, 3.00) llama3.2:3b: (5.00, 5.00, 5.00) llama3.3:70b: (5.00, 5.00, 4.00)	(3.00, 1.50, 2.50)

Table 5: Turns where **No Reasoner** (hist) and **EmoLASP** (hist) predictions both have high error.