2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Emotion Traceability Analysis: A Multi-Strategy Framework for LLM Dialogue Processing

Zichen Yu(✉), Senwei Liang, Aiyu Li , Xiaoyi Zhu, Tianlan Pan,
Jionglong Su (✉)

School of AI and Advanced Computing, Xi'an Jiaotong-liverpool university, Suzhou, 215123, China
Zichen.Yu22@student.xjtlu.edu.cn & Jionglong.Su@xjtlu.edu.cn

Abstract. Recent approaches in emotional causal reasoning leverage Retrieval-Augmented Generation (RAG) and multimodal fusion to enhance the accuracy of large language models (LLMs) in analyzing emotions. As a critical cognitive process for understanding, inferring and predicting the antecedents and consequences of emotional states, emotional causal reasoning primarily involves two components: emotional understanding and causal inference. However, LLMs face two key challenges in analyzing emotional causality: (1) the inability to process ultra-long texts due to input length constraints, and (2) insufficient capability to track emotional dynamics in dialogues. To address these limitations, we propose the Emotion Traceability Analysis Framework (ETAF), which employs RAG-based keyword retrieval to extract critical events from dialogues and dynamically segments conversations according to event progression, enabling LLMs to comprehend contextualized events holistically. In addition, we integrate character analysis and variation correction modules to improve the precision of the model in tracking emotional causal chains between characters and refining the interpretation of emotional shifts. Experimental results on the ATLAS-6 dataset demonstrate that our framework improves the performance of GLM-4-air by 17.79%, outperforming DeepSeek-R1(origin) by 6.49% and achieving state-of-the-art results.

Keywords: Emotional Causal Reasoning, Large Language Models, Retrieval-Augmented Generation, Emotion Traceability.

1 Introduction

Emotion constitutes a core component of human cognition and social interaction. In everyday contexts, emotional expressions conveyed through language exhibit significant complexity, particularly in lengthy texts. While progress has been made in emotion recognition using large language models (LLMs), current approaches predominantly focus on static, discrete emotion classification while neglecting the exploration of temporal evolution and causal relationships in emotions—specifically, the analysis of emotional chains [14]. Although works like CausEmotion have employed Retrieval-Augmented Generation (RAG) and multimodal methods to investigate emotional shifts, achieving notable advancements, emotion recognition remains a challenging frontier in AI research [21].

Three critical challenges persist in emotion-aware causal reasoning. First, the complexity of emotional causality significantly surpasses that of general causal inference, as emotions are inherently volatile and prone to external influences, frequently introducing analytical inaccuracies [10][18]. Second, tracking dynamic emotional shifts in extended conversations remains problematic—while such dialogues typically feature intricate emotional transitions, current LLMs exhibit limited capacity to interpret these evolving dynamics or reconstruct the contextual "flow" of emotions across sequential events [15][16][21]. Third, attribution ambiguity arises from the tendency of the models to conflate different emotional triggers, generating misattribution errors that obscure the context of the event and hinder precise causal reasoning [9][20].

Despite their robust linguistic capabilities, LLMs exhibit notable limitations in emotional chain analysis: they tend to mistake statistical correlations for causal relationships [12][19]; inadequately model temporal dependencies in emotional transitions [2][18]; and face difficulties integrating individual differences and cultural background information [7][13]. Furthermore, existing models often treat emotions as discrete labels rather than structured psychological states with internal logic, overlooking the mutual influences and transformation patterns between emotions.

To address these issues, we propose an **Emotion Traceability Analysis Framework (ETAF)**. This framework extracts keywords from extended dialogues to retrieve critical events, employs dynamic segmentation to enhance LLMs' awareness of emotional flows within events, and incorporates character role analysis to trace the causal evolution of emotional changes.

2 Related work

In emotion analysis research, the emotion six-tuple framework established by Zhang et al. effectively analyzes emotions in dialogues [21]. Majumder et al. develop the DialogueRNN model, which captures emotional interactions between dialogue participants by tracking their emotional states [13].

Recent research focuses on understanding the causal mechanisms behind emotions. Chen et al. introduce the emotion reasoning task, which aims to infer emotional causes from narrative texts [3]. Huang et al. propose the causal emotion entailment task designed to identify utterances that trigger specific emotions in dialogues [6]. Poria et al. enhance emotion cause prediction accuracy by incorporating commonsense knowledge and constructing knowledge-enhanced dialogue graphs [17].

The application of large language models (LLMs) to emotion analysis presents both opportunities and challenges. Poria et al. utilize chain-of-thought prompting to guide LLMs in analyzing dialogue emotions, though reasoning errors persist. Huang generates Emotion Reasoning Chains (ERC) that require supervised learning to correct output inconsistencies.

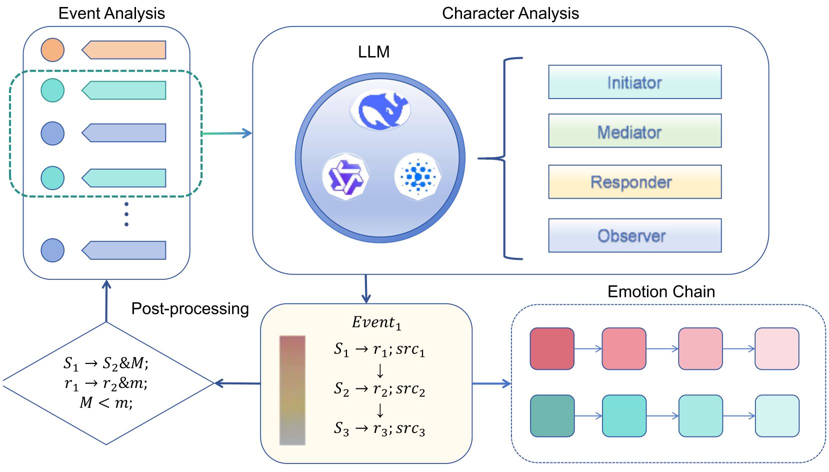


Fig. 1. The Emotion Causal Analysis Framework illustrates the three-module architecture for emotional causality reasoning in dialogues. The Event Analysis module (left), Character Analysis module (top right), Post-processing module (bottom). The diamond shape represents the validation checkpoint.

In evaluating GPT's emotion recognition capabilities, Laban et al. identifies limitations in analyzing dynamic emotion transfer due to context constraints [8]. The CausEmotion-1 framework by Zhang et al.[21] incorporates RAG and multimodal fusion techniques but its limitations include lengthy dialogues and complex emotional logic.

3 Methodology

3.1 Framework Design

The emotion causal analysis framework first extracts key information from long dialogues using event retrieval, then employs large language models to construct character roles and build complete emotion causal chains. **Fig. 1** presents an overview of our approach, illustrating the interconnections between different components of the system. The framework begins by segmenting the input dialogue and detecting event boundaries as described in Section 3.2. These detection results create a structured representation of the dialogue flow, identifying critical transition points where emotional states may change. Subsequently, these results are provided to large language models to construct dialogue participants' role positioning and relationship networks, as detailed in Section 3.3. This enables a richer understanding of the social dynamics that influence emotional expressions and reactions within the dialogue. The core of our method involves logical detection and optimization of the constructed results, with detailed information presented in Section 3.4.

3.2 Event Analysis Module

The event analysis module focuses on precise identification of dialogue event boundaries by integrating semantic change detection, role interaction analysis, and topic transition recognition into a multi-strategy fusion boundary detection approach.

Semantic Change Detection For a given dialogue text set $D = \{u_1, u_2, \ldots\}$, where u_i represents the *i*-th utterance, the system first utilizes an embedding model to calculate the vector representation for each utterance $e_i = f(u_i)$. Subsequently, a sliding window compares the semantic similarity between adjacent segments to compute semantic change scores, allowing us to detect significant shifts in the dialogue's semantic content.

$$\triangle \mathrm{S(i)} = 1 - \frac{\sum_{\mathrm{j=i-w}}^{\mathrm{i-1}} \mathrm{e_j} \cdot \sum_{\mathrm{j=i}}^{\mathrm{i+w-1}} \mathrm{e_j}}{|\sum_{\mathrm{j=i-w}}^{\mathrm{i-1}} \mathrm{e_j}| \cdot |\sum_{\mathrm{j=i}}^{\mathrm{i+w-1}} \mathrm{e_j}|}, \tag{1}$$

where $\triangle S(\mathrm{i})$ represents the semantic change score at position i; e_j is the embedding vector of the j-th utterance, and w is the sliding window size that determines how many utterances are considered in each segment. The dot product in the numerator captures the alignment between the semantic content of adjacent windows, while the denominators normalize for the magnitude of the pooled embedding vectors. $| \cdot |$ denotes the Euclidean norm of the vector.

Role Interaction Change Analysis This analysis focuses on transitions in interaction patterns between dialogue participants and changes in interaction roles.

We use Jensen-Shannon Divergence (JSD) to calculate the difference in interaction patterns between the previous and subsequent windows, by means of the interaction change score, i.e.,

$$\triangle R(i) = JSD\big(P_{\text{before}}(i), P_{\text{after}}(i)\big), \tag{2}$$

where $\triangle R(i)$ represents the role interaction change score at position i, *JSD* denotes Jensen-Shannon divergence, and $P_{before}(i)$ and $P_{after}(i)$ represent the role interaction probability distributions in the windows before and after position i.

The interaction probability distribution at position i is calculated by normalizing the interaction frequency matrix for the corresponding window:

$$P_{\text{before/after}}(i)(j,k) = \frac{m_{jk}^{\text{before/after}}}{\sum_{j',k' \in P} m_{j'k'}^{\text{before/after}}}\ , \tag{3}$$

where m_{jk}^{before} and $\mathrm{m}_{\mathrm{jk}}^{\mathrm{after}}$ represent the number of interactions between role j and role k in the windows $[i-w, i-1]$ and $[i, i+w-1]$, respectively.

Topic Change Detection This component focuses on analyzing topic changes in dialogue content, locating event boundaries by identifying significant transitions in discussion topics. Topic transitions frequently trigger emotional responses as participants react to new information or shifts in conversational focus. We employ the Dirichlet process mixture model to uncover latent topics in the dialogue and calculate topic distribution change scores.

The topic distribution vectors are calculated using the Gibbs sampling update formula from the Latent Dirichlet Allocation (LDA) model, computing the probability of words being assigned to corresponding topics:

$$p\left(z_{i,j}=k\middle|\boldsymbol{z}_{-(i,j)},\boldsymbol{w}\right)\propto\frac{n_{j,-(i,j)}^{(k)}+\alpha_k}{n_{j,-(i,j)}+\sum_{k'=1}^{K}\alpha_{k'}}\cdot\frac{n_{-(i,j),w_{i,j}}^{(k)}+\beta_{w_{i,j}}}{n_{-(i,j)}^{(k)}+\sum_{v=1}^{V}\beta_v},\tag{4}$$

where p represents the conditional probability that the i-th word in the j-th window is assigned to topic k, given all other topic assignments and all observed words; z_{ij} denotes the topic of the i-th word in window j, $z_{-(i,j)}$ represents all topic assignments except this one, w represents the observed words, $n_{j,-(i,j)}^{(k)}$ is the number of words in window j assigned to topic k (excluding the current word), $n_{-(i,j),w_{i,j}}^{(k)}$ is the number of times word $w_{i,j}$ is assigned to topic k (excluding the current instance), and α and β are Dirichlet prior parameters that control the smoothing of topic and word distributions, respectively.

Multi-strategy Fusion Boundary Detection To improve the accuracy and robustness of event boundary detection, we propose a multi-strategy fusion method, which comprehensively utilizes the change boundary nodes B_{sem}, B_{role} and B_{topic} of the above three dimensions (sentiment, topic and role) of analysis, i.e.,

$$S_{boundary}(i)=w_{sem}\cdot I(i\in B_{sem})+w_{topic}\cdot I\left(i\in B_{topic}\right)+w_{role}\cdot I(i\in B_{role}),\tag{5}$$

where $S_{boundary}(i)$ represents the comprehensive boundary score at position i, $I(\cdot)$ is an indicator function that returns 1 when the condition is true and 0 otherwise, and $w_{sem}, w_{topic},$ and w_{role} represent the weights for each detection method.

When the boundary score exceeds the threshold $\left(S_{boundary}(i)\ >\ \tau_B\right)$, position i is defined as an event boundary. Since dialogue events typically occur within a certain range of sizes, we perform post-processing analysis after dialogue segmentation to handle cases where events are too large or too small. We set maximum (max) and minimum (min) event size thresholds, removing boundaries with lower scores when event length is less than *min*, and further analyzing events exceeding the maximum size by selecting one or more intermediate boundary points with the highest boundary scores.

$$B_{final}=\{i|S_{boundary}(i)>\tau\_B,1\le i\le n-1\}.\tag{6}$$

The final set B_{final} represents the optimized event boundaries.

3.3 Character Analysis Module

The character analysis module builds upon the event segmentation provided by the event analysis module, focusing on role positioning of dialogue participants, relationship network construction, and emotion tracking.

Character Relationship Recognition To construct character relationships in dialogues, we quantify interaction patterns between participants to structure the character relationship network. We first analyze speakers in the dialogue, extract all speaker IDs, and construct an initial character set $P=\{per_1,\ldots,per_m\}$, where m represents the number of speakers. We then analyze character relationships along three dimensions: interaction frequency, linguistic features, and topic preferences.

Interaction Frequency We calculate character interactions and volume relationships to construct a complete interaction frequency matrix for the dialogue. The interaction frequency is calculated as:

$$f_{jk} = \sum_{t=1}^{n} I(\text{Holder}_t = j \wedge \text{Target}_t = k) + \lambda \sum_{t=1}^{n} I\big(\text{Holder}_t = j \wedge \text{Mention}_t(k)\big). \quad (7)$$

Here, Holder_t and Target_t represent the speaker and target object in round t of the dialogue, $\text{Mention}_t(k)$ indicates whether character k is mentioned in round t of the dialogue, and λ is the mention weight coefficient that determines the relative importance of direct interactions versus mentions.

Relationship Types Based on the interaction frequency matrix, we define additional metrics to characterize the nature of relationships between characters: Relationship Intensity (RI) and Asymmetry Index (AI).

$$RI_{jk} = \frac{f_{jk} + f_{kj}}{\sum_{a,b} f_{ab}} \cdot m^2, AI_{jk} = \frac{|f_{jk} - f_{kj}|}{f_{jk} + f_{kj}}. \quad (8)$$

RI_{jk} represents the relative strength of the relationship between characters j and k, normalized by the total interaction volume in the dialogue and multiplied by m^2 to adjust to a reasonable range ($[0 \sim 1]$). Higher values indicate stronger connections between characters. AI_{jk} represents the degree of asymmetry in interactions, with values closer to 0 indicating balanced relationships and values closer to 1 indicating highly unbalanced relationships where one character dominates the interaction.

Character Roles We calculate three centrality indices that capture different aspects of a character's position in the social network: Degree Centrality (DC), Betweenness Centrality (BC), and Eigenvector Centrality (EC):

$$DC(j) = \sum_{k \neq j} I\big(f_{jk} + f_{kj} > 0\big), BC(j) = \sum_{s \neq j \neq t} \frac{\sigma_{st}(j)}{\sigma_{st}}, EC(i) = \frac{1}{\lambda} \sum_{j} a_{ij} \cdot EC(j) (9)$$

where σ_{st} is the number of shortest paths from character s to character t on the interaction graph, and $\sigma_{st}(j)$ is the number of paths that pass-through character j. λ is the eigenvector, and F is the largest interaction matrix, EC is the eigenvector centrality of node i, and $EC = [EC_1, EC_2, \dots, EC_n]$. EC_i is the first element in the eigenvector centrality vector. a_{ij} represents the connection strength.

Emotion Tracking and State Transitions. The system analyzes each person's emotional trajectory across events to construct emotional state sequences. For each character p in event v, we construct an emotional state sequence $E_p^v = \{e_1, \dots, e_k\}$, where each emotion e_i includes three key attributes:

- State type: $s_i \in s = \{positive, negative, neutral, ambiguous, doubt\}$, representing the valence of the emotional state
- Change reason: r_i, textual description of the reason for the emotion change, providing a causal explanation
- Influence source: src_i, the object or event causing the emotional change, identifying the trigger

When we assign change reasons r_i to each emotional state e_i, we systematically consider three key factors: the previous state (s_{i-1}), the current state (s_i), and contextual information from the relationship network (G_R), as well as two main factors: the utterance triggering the current emotion (u_i) and the set of historical emotional states (H_i). The cause of emotional change is the combined effect of these five factors, calculated through a reasoning function R:

$$r_i = R(s_{i-1}, s_i, G_R, u_i, H_i). \tag{10}$$

This function generates a natural language explanation of the emotional transition, identifying key triggers and contextual factors that contributed to the change.

3.4 Post-processing Module

The post-processing module integrates the results of event analysis and character analysis, preliminarily constructs complete emotional causal chains, and performs logical detection and optimization.

Initial Causal Chain Detection Our constructed emotional causal chains adopt a hierarchical structure, organized into three levels according to the logical combination of characters, events, and emotions.

$$C = \{(p_1, E_1), (p_2, E_2), \dots, (p_m, E_m)\}, \tag{11}$$

where C represents the complete emotion causal chain; p_i represents the i-th character; E_i represents the set of events in which character p_i participates, defined as $E_i = \{(v_i^1, S_i^1), (v_i^2, S_i^2), \dots, (v_i^{n_i}, S_i^{n_i})\}$; v_i^j represents the i-th person in j-th event description; S_i^j represents the emotion sequence of character p_i in event v_i^j, defined as $S_i^j = \{(s_i^{j,1}, r_i^{j,1}, src_i^{j,1}), \dots, (s_i^{j,k_j}, r_i^{j,k_j}, src_i^{j,k_j})\}$; $s_i^{j,k}$, $r_i^{j,k}$, $src_i^{j,k}$ respectively represent emotion state, change reason and influence source.

The post-processing module validates the logical consistency and psychological reasonability of the generated emotional causal chains. If inconsistencies are detected, the system returns to the beginning of the emotion chain construction process.

4 Experiments

Dataset and Experimental Setup

Dataset We conduct experiments on the ATLAS-6 dataset, which consists of long-form dialogues annotated with emotion six-tuples [12]. Each annotation includes target, aspect, opinion, rationale, holder, and sentiment information. To minimize the impact of large language model errors and hallucinations, we randomly sample 100 dialogues from ATLAS-6 and perform repeated tests, reporting average results to ensure generalizability.

Experimental Details In our experiments, we employ multiple large language models, including Qwen-7B [1], GLM-4-Plus [4], GLM-4-Air [4], and DeepSeek-V3 [11], with DeepSeek-R1 [5] serving as the baseline. To simulate real-world usage conditions, we set the temperature parameter to 0.3 across all models. For event boundary detection, we pre-configure the event threshold and event matching threshold to 0.3. For evaluation purposes, we utilize Embedding-3 as the embedding model and DeepSeek-R1 as the judge model.

Evaluation Metrics To comprehensively assess the correctness of emotion causal chains predicted by large language models, we adopt the following four evaluation metrics: **SA, SIA, RCEM, RCLLM**.

$$\mathrm{SA} = \frac{1}{N}\sum_{i=1}^{N} I\left(S_{pi} = S_{ti}\right), \mathrm{SIA} = \frac{1}{N}\sum_{i=1}^{N} I\left(ID_{p_i} = ID_{t_i}\right),$$

$$RCEM = \frac{1}{N}\sum_{i=1}^{N} cos\left(E\left(r_g\right), E\left(r_p\right)\right), RCLLM = \frac{1}{N}\sum_{i=1}^{N} I_{LLM}\left(r_g, \mathrm{r_p}\right),$$

where $I(\cdot)$ is the indicator function, N is the count of dialogs, $S_{pi}\ and\ S_{ti}$ is the i-th predict state and true state. $ID_{pi}\ and\ ID_{ti}$ is the source id of i-th dialogs. $E(\cdot)$ is the embedding function, $r_g\ and\ r_p$ is the true reason and predict reason, $I_{LLM}(\cdot)$ is a binary judgment function based on a large language model.

4.1 Main Experimental Results

Baseline Comparison Table. 1 below and **Fig. 2** present the performance of different large language models on the ATLAS-6 dataset across our four evaluation metrics.

The results demonstrate that our framework significantly improves performance across all metrics. Particularly noteworthy is the improvement observed with GLM-4-Air, where the average score increases from 45.29 to 63.08, with a remarkable 27.91 point improvement in the *RCLLM* metric. Compared to the baseline (DeepSeek-R1) average score of 56.59, our approach achieves a 6.49 point improvement.

4.2 Ablation Studies

To validate the effectiveness of each component in our proposed framework, we conduct a series of ablation studies on the ATLAS-6 dataset. In all experiments, we use GLM-4-Air as the base model and progressively add different functional modules to evaluate each component's contribution.

Component Effectiveness Analysis We systematically examine the three main components of our framework: RAG event segmentation, character analysis, and post-processing. **Table. 2 below** presents the results of these experiments.

The Event Analysis + Emotional Chain Post-processing combination (f) demonstrates superior overall performance (avg: 61.16, +15.87), with exceptional improvement in SA (88.40, +25.48). This combination significantly enhances the model's ability to understand emotional expressions. Character Analysis + Emotional Chain Post-processing (g) achieves outstanding results in SIA (65.21, +25.77), confirming its effectiveness for identifying emotional subject-object relationships. All configurations, including the Emotional Chain Post-processing module, show substantial improvements, while those relying solely on Event Analysis (b) or Event + Character Analysis (e) show performance decreases in SA (-9.80 and -13.91). This indicates that event segmentation without emotional context may impair sentiment understanding.

Table 1. LLM Performance Comparison: OnlyLLM vs. ETAF Framework.

Model	Framework	SA	SIA	RCLLM	RCEM	AVG
DeepSeek-r1	OnlyLLM	58.85	70.28	32.85	64.40	56.59
	ETAF	77.36	**72.92**	41.80	**65.34**	**64.34**
	Diff	+18.51 ↑	+2.64 ↑	+8.95 ↑	+0.94 ↑	+7.75 ↑
DeepSeek-v3	OnlyLLM	64.93	43.68	38.75	66.91	53.57
	ETAF	83.61	49.65	32.57	62.92	57.19
	Diff	+18.68 ↑	+5.97 ↑	−6.18 ↓	−3.99 ↓	+3.62 ↑
Qwen2.5-7B-Instruct	OnlyLLM	33.82	25.62	2.29	48.64	30.09
	ETAF	59.17	24.17	26.04	58.96	42.08
	Diff	+25.35 ↑	−1.45 ↓	+23.75 ↑	+10.32 ↑	+11.99 ↑
GLM-4-Plus	OnlyLLM	56.25	49.72	17.29	65.47	47.18
	ETAF	83.12	50.83	41.95	63.36	59.82
	Diff	+26.87 ↑	+1.11 ↑	+24.66 ↑	−2.11 ↓	+12.64 ↑
GLM-4-Air	OnlyLLM	62.92	39.44	17.64	61.18	45.29
	ETAF	**85.21**	56.53	**45.55**	65.03	63.08
	Diff	+22.29 ↑	+17.09 ↑	+27.91 ↑	+3.85 ↑	+17.79 ↑

These findings establish Emotional Chain Post-processing as the critical component in our framework, functioning most effectively when paired with complementary analytical modules for specific tasks.

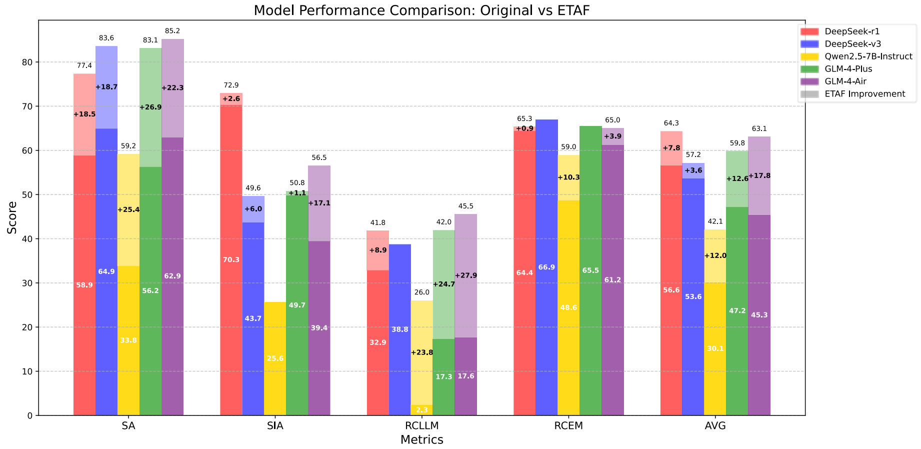
4.3 Comparison with State-of-the-Art Methods

To comprehensively evaluate our proposed framework, we compare it with existing state-of-the-art methods for emotion causal chain analysis.

The ETAF framework employs a three-tier architecture with event, character, and post-processing modules. The event module detects dialog emotional shifts via multi-feature analysis, while the character module maps relationship networks. The post-processor integrates these outputs and ensures logical consistency.

In contrast, CauseMotion focuses on six-tuple recognition using RAG and audio features, prioritizing element identification over emotional progression analysis.

Experiments show ETAF achieves 59.82 on ATLAS-6, outperforming CauseMotion (57.4) by 2.42 points. ETAF's strength lies in holistic emotional trajectory understanding and relational analysis, enabling superior emotional causality modeling. This advancement enhances intelligent systems' capacity to interpret complex affective processes.



Darker color: Original model performance, Lighter color: Additional improvement with ETAF

Fig. 2. Model Performance Comparison: Original vs ETAF Framework. The bar chart illustrates the performance of different models (DeepSeek-r1, DeepSeek-v3, Qwen2.5-7B-Instruct, GLM-4-Plus, and GLM-4-Air) across four evaluation metrics (SA, SIA, RCLLM, and RCEM) and their average (AVG).

Table 2. Ablation Study Results: Performance Metrics of Different Module Configurations.

Type	Model	SA	SIA	RCLLM	RCEM	AVG
Baseline	BASE[A]	62.92	39.44	17.64	61.18	45.29
Model Variants	EVENT[B]	53.12	44.37	26.18	63.94	46.91
		−9.80 ↓	+4.93 ↑	+8.54 ↑	+2.76 ↑	+1.62 ↑
	PERSON[C]	63.26	40.69	23.13	64.48	47.89
		+0.34 ↑	+1.25 ↑	+5.49 ↑	+3.30 ↑	+2.60 ↑
	EMOTION[D]	64.17	45.83	11.67	62.47	46.10
		+1.25 ↑	+6.39 ↑	−5.97 ↓	+1.29 ↑	+0.81 ↑
	EVENT+PERSON[E]	49.01	41.88	25.90	64.78	49.01
		−13.91 ↓	+2.44 ↑	+8.26 ↑	+3.60 ↑	+3.72 ↑
	EVENT+EMOTION[F]	**88.40**	53.33	**33.89**	**69.02**	**61.16**
		+25.48 ↑	+13.89 ↑	+16.25 ↑	+7.84 ↑	+15.87 ↑
	PERSON+EMOTION[G]	76.88	**65.21**	30.63	62.72	58.86
		+13.96 ↑	+25.77 ↑	+12.99 ↑	+1.54 ↑	+13.57 ↑

5 Conclusions

In this work, we propose a novel Emotion Traceability Analysis Framework for emotional causal reasoning, which aims to address the limitations of current LLMs in processing ultra-long texts and tracking emotional dynamics in dialogues. This framework integrates multi-strategy event boundary detection, character relationship analysis, and emotion state tracking to enhance the accuracy of emotion causal chains. The core of our method is the multi-dimensional approach, which is achieved by event segmentation, character role positioning, and hierarchical causal chain validation to accurately capture emotional transitions and their underlying causes. Experimental results on the ATLAS-6 dataset demonstrate that our framework improves the performance of GLM-4-air significantly, outperforming state-of-the-art models like DeepSeek-R1/V3, and ablation studies confirm the effectiveness of each component in our proposed framework.

References

[1] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[2] Chen, Y., Chen, S., Li, Z., Yang, W., Liu, C., Tan, R., Li, H.: Dynamic transformers provide a false sense of efficiency. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st An- nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7164–7180. Association for Computational Linguis- tics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.395, https://aclanthology.org/2023.acl-long.395/

[3] Chen, Y., Hou, W., Cheng, X., Li, S.: Joint learning for emotion classification and emotion cause detection. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsu- jii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Nat- ural Language Processing. pp. 646–651. Association for Computational Linguis- tics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1066, https://aclanthology.org/D18-1066/

[4] GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas,D., Feng, G., Zhao, H., et al.: Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 (2024)

[5] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

[6] Huang, Z., Zhao, J., Jin, Q.: Ecr-chain: Advancing generative language mod- els to better emotion-cause reasoners through reasoning chains. arXiv preprint arXiv:2405.10860 (2024)

[7] Kumar, C.O., Gowtham, N., Zakariah, M., Almazyad, A.: Multimodal emotion recognition using feature fusion: An llm-based approach. IEEE Access (2024)

[8] Laban, P., Vig, J., Hearst, M., Xiong, C., Wu, C.S.: Beyond the chat: Executable and verifiable text-editing with llms. In: Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. pp. 1–23 (2024)

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)

[10] Li, B., Fei, H., Li, F., Wu, Y., Zhang, J., Wu, S., Li, J., Liu, Y., Liao, L., Chua, T.S., et al.: Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. arXiv preprint arXiv:2211.05705 (2022)

[11] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437(2024)

[12] Luo, M., Fei, H., Li, B., Wu, S., Liu, Q., Poria, S., Cambria, E., Lee, M.L., Hsu, W.: Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7667–7676 (2024)

[13] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E.: Dialoguernn: An attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6818–6825 (2019)

[14] Ni, J., Young, T., Pandelea, V., Xue, F., Cambria, E.: Recent advances in deep learning based dialogue systems: A systematic survey. Artificial intelligence review **56**(4), 3055–3155 (2023)

[15] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: International workshop on semantic evaluation. pp. 19–30 (2016)

[16] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1050, https://aclanthology.org/P19-1050/

[17] Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S.Y.B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., et al.: Recognizing emotion cause in conversations. Cognitive Computation **13**, 1317–1332 (2021)

[18] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H., Wu, F.: Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635 (2020)

[19] Wang, Y., Wu, S., Zhang, Y., Wang, W., Liu, Z., Luo, J., Fei, H.: Multimodal chain- of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605 (2025)

[20] Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., Lam, W.: Aspect sentiment quad prediction as paraphrase generation. arXiv preprint arXiv:2110.00796 (2021)

[21] Zhang, Y., Li, Y., Yu, Z., Tang, F., Lu, Z., Li, C., Dang, K., Su, J.: Decoding the flow: Causemotion for emotional causality analysis in long-form conversations. arXiv preprint arXiv:2501.00778 (2025)