Xi’an Jiaotong-Liverpool University
西交利物浦大学

EmoCaustics Series: Causal Chain Analysis in Extended Conversations

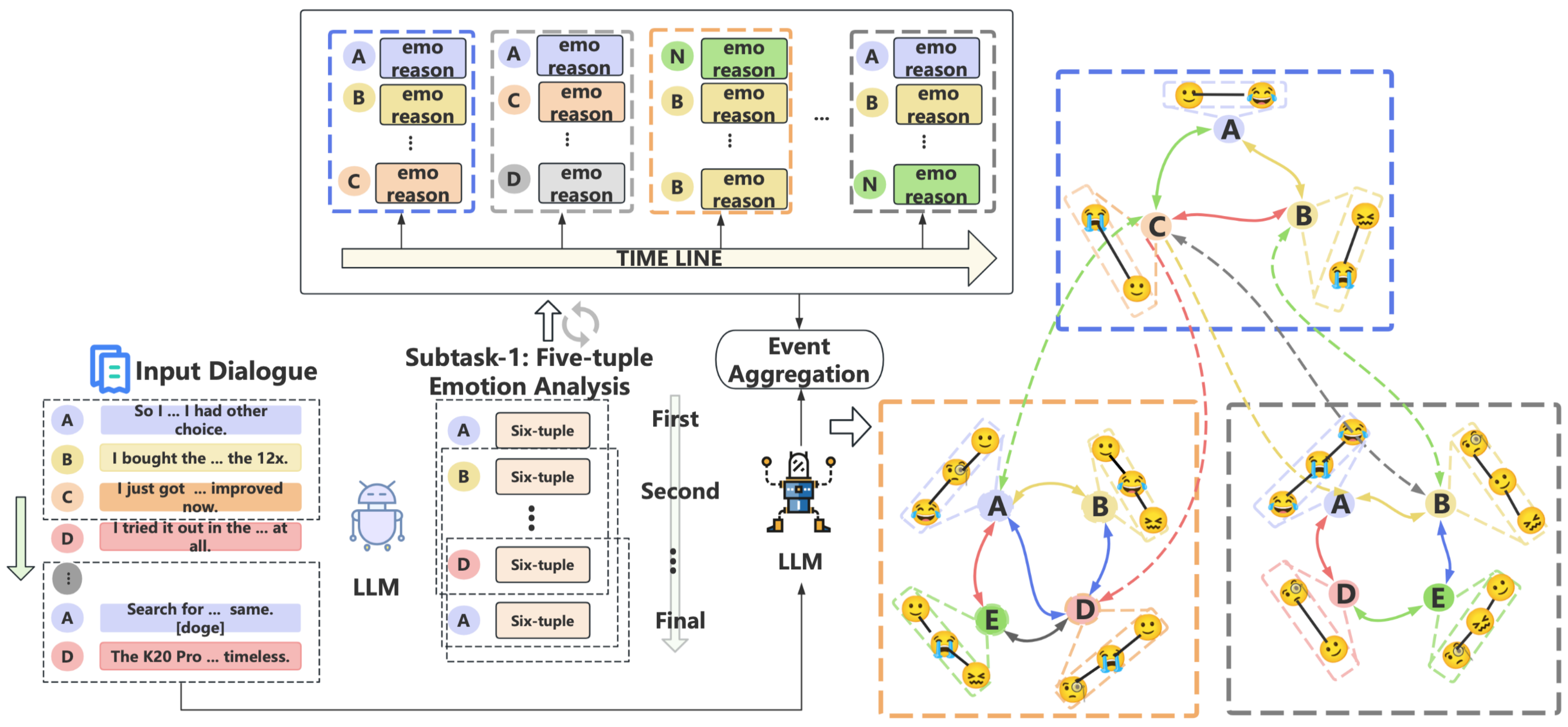
Student: Yuxuan Zhang Supervisor: Mian Zhou

Abstract

Emotion causality reasoning plays a crucial role in understanding complex conversational dynamics, especially in long-dialogue scenarios. However, two major challenges hinder its effectiveness: the difficulty of comprehending ultra-long contexts due to limitations in dependency modeling and the inadequate analysis of emotion causation in psychosocial dialogues. To overcome these issues, we propose the **EmoCaustics** framework, which enhances the modeling of emotion causality chains in extended conversations. This is achieved through innovative memory modules, event aggregation mechanisms, and dynamic context windows. Additionally, we evaluate both mainstream closed-source and open-source large language models on emotion causality reasoning tasks to assess their effectiveness in this domain.

	Model	SA	SIA	RCLLM	RCEM	T	A	O	R	S
	GPT-o1-2024-12-1 [1]	65.51	50.16	44.03	30.07					
LLM	GPT-o1-mini-2024-09-12 [24]	$55.5_{\downarrow 10.01}$	$41.6_{\downarrow 8.56}$	$34.07_{\downarrow 9.96}$	$22.37_{\downarrow 7.70}$	73.52	71.15	71.47	71.61	73.09
	GPT-o3-mini-2025-01-31 [14]	$56.36_{\downarrow 9.15}$	$41.76_{\downarrow 8.40}$	$34.21_{\downarrow 9.82}$	$21.87_{\downarrow 8.20}$	$\underline{76.26}$	$\underline{74.95}$	69.17	66.67	71.53
	GPT-4o-2024-08-06 [23]	$\underline{58.05}_{\downarrow 7.46}$	$39.53_{\downarrow 10.63}$	$\underline{35.56}_{\downarrow 8.47}$	$\underline{23.79}_{\downarrow 6.28}$	70.54	69.12	68.12	70.98	74.45
	GLM-4-Plus [13]	$49.73_{\downarrow 15.78}$	$46.03_{\downarrow 4.13}$	$30.9_{\downarrow 13.13}$	$21.41_{\downarrow 8.66}$	70.12	66.65	68.41	65.32	65.12
	GLM-4-Air-0111 [13]	$57.88_{\downarrow 7.63}$	$43.45_{\downarrow 6.71}$	$32.69_{\downarrow 11.34}$	$21.74_{\downarrow 8.33}$	71.12	68.54	68.11	65.44	62.52
	DeepSeek-V3 [7]	$51.71_{\downarrow 13.80}$	$42.45_{\downarrow 7.71}$	$32.52_{\downarrow 11.51}$	$22.55_{\downarrow 7.52}$	73.25	70.12	$\underline{73.32}$	$\underline{71.88}$	$\underline{75.98}$
	Qwen2.5-7B-Instruct [29]	$53.65_{\downarrow 11.86}$	$\underline{47.75}_{\downarrow 2.41}$	$21.11_{\downarrow 22.92}$	$14.22_{\downarrow 15.85}$	70.55	66.22	67.45	66.32	65.12
	Qwen2.5-14B-instruct [29]	$53.43_{\downarrow 12.08}$	$41.45_{\downarrow 8.71}$	$30.64_{\downarrow 13.39}$	$20.58_{\downarrow 9.49}$	67.11	65.02	66.06	67.85	67.09
	Qwen2.5-72B-Instruct [29]	$48.22_{\downarrow 17.29}$	$33.69_{\downarrow 16.47}$	$28.36_{\downarrow 15.67}$	$19.63_{\downarrow 10.44}$	71.52	68.20	69.12	69.65	70.22
	InternLM3-8B-Instruct [15]	$47.80_{\downarrow 17.71}$	$44.76_{\downarrow 5.40}$	$17.86_{\downarrow 26.17}$	$12.07_{\downarrow 18.00}$	74.08	69.41	66.82	64.96	69.20
	Average	53.23	42.25	29.79	20.02	71.81	68.94	68.80	68.07	69.43
CauseMotion [34]	GLM-4-Plus	$\mathbf{60.85}_{\downarrow 4.66}$	$\mathbf{47.20}_{\downarrow 2.96}$	$34.50_{\downarrow 9.53}$	$23.15_{\downarrow 6.92}$	72.40	72.60	73.50	68.80	72.80
	GLM-4-Air-0111	$59.95_{\downarrow 5.56}$	$40.15_{\downarrow 10.01}$	$\mathbf{36.85}_{\downarrow 7.18}$	$25.75_{\downarrow 4.32}$	71.20	71.30	71.80	65.70	74.20
	GPT-4o-2024-08-06	$57.50_{\downarrow 8.01}$	$35.50_{\downarrow 14.66}$	$33.40_{\downarrow 9.63}$	$\mathbf{26.35}_{\downarrow 3.72}$	75.80	74.30	74.80	72.60	74.90
	DeepSeek-V3	$56.25_{\downarrow 9.26}$	$37.45_{\downarrow 12.71}$	$28.35_{\downarrow 15.68}$	$21.40_{\downarrow 8.67}$	74.30	70.15	72.80	70.40	77.60
	Qwen2.5-7B-Instruct	$54.40_{\downarrow 11.11}$	$44.65_{\downarrow 1.51}$	$30.45_{\downarrow 13.58}$	$20.85_{\downarrow 9.22}$	67.35	65.45	65.80	67.20	67.40
	Qwen2.5-72B-Instruct	$55.75_{\downarrow 9.76}$	$36.30_{\downarrow 13.86}$	$29.65_{\downarrow 14.38}$	$19.95_{\downarrow 10.12}$	68.00	66.15	67.50	62.90	66.10
	Average	57.45	39.88	32.70	23.08	71.51	69.99	71.03	67.93	72.17
EmoCaustics	GLM-4-Plus	$65.79_{+0.28}$	$46.52_{\downarrow 3.64}$	$44.05_{+0.02}$	$30.02_{\downarrow 0.05}$	73.22	73.26	74.32	69.21	73.54
	GLM-4-Air-0111	$\underline{\mathbf{66.10}}_{+0.59}$	$45.16_{\downarrow 5.00}$	$\underline{\mathbf{46.04}}_{+2.01}$	$\underline{\mathbf{31.54}}_{+1.47}$	71.52	72.22	72.32	66.12	75.51
	GPT-4o-2024-08-06	$59.09_{\downarrow 6.42}$	$37.16_{\downarrow 13.00}$	$38.23_{\downarrow 5.80}$	$24.95_{\downarrow 5.12}$	$\underline{\mathbf{77.32}}$	$\underline{\mathbf{77.11}}$	$\underline{\mathbf{75.21}}$	$\underline{\mathbf{73.52}}$	76.21
	DeepSeek-V3	$60.1_{\downarrow 5.41}$	$42.27_{\downarrow 7.89}$	$39.53_{\downarrow 4.50}$	$25.81_{\downarrow 4.26}$	75.22	70.21	73.32	71.33	$\underline{\mathbf{81.27}}$
	Qwen2.5-72B-Instruct	$62.58_{\downarrow 2.93}$	$\underline{\mathbf{56.43}}_{+6.27}$	$38.26_{\downarrow 5.77}$	$25.12_{\downarrow 4.95}$	67.75	65.92	66.15	67.74	67.82
	Qwen2.5-7B-Instruct	$55.62_{\downarrow 9.89}$	$29.51_{\downarrow 20.65}$	$26.44_{\downarrow 17.59}$	$17.63_{\downarrow 12.44}$	68.33	66.31	68.46	63.25	66.26
	Qwen2.5-14B-Instruct	$60.48_{\downarrow 5.03}$	$22.19_{\downarrow 27.97}$	$35.78_{\downarrow 8.25}$	$24.21_{\downarrow 5.86}$	71.32	68.15	67.42	69.11	69.15
	Average	59.96	36.32	35.11	23.41	72.10	70.45	71.03	68.61	72.82

Impact of EmoCaustics on Large Language Models Performance Across Different Evaluation Metrics. The scores are reported in percentage (%). ↓and + indicate performance decrease and increase compared to GPT-01-2024-12-17, respectively. Bold and underlined values represent the best performance in each section.


A
emo reason
B
C
D
N
TIME LINE
Event Aggregation
Subtask-1: Five-tuple Emotion Analysis
Input Dialogue
So I ... I had other choice.
I bought the ... the 12x.
I just got ... improved now.
I tried it out in the ... at all.
Search for ... same. [doge]
The K20 Pro ... timeless.
LLM
Six-tuple
First
Second
Final
E

An overview of our multimodal emotion analysis framework. The framework processes dialogue inputs through LLM-based Five-tuple Emotion Analysis, followed by Event Aggregation and Complex Causal Reasoning to understand emotional relationships. The results are visualized as social interaction graphs showing Value System Alignment, Social-Cultural Structure Analysis, and Critique of Standardized Life, with a timeline view tracking emotional state evolution across participants.

Result

EmoCaustics significantly enhances model performance across key metrics compared to baseline models like GPT-01-2024-12-17. Notable improvements include +0.59% in Emotion State Accuracy for GLM-4-Air-0111, +6.27% in Source ID Accuracy for Qwen2.5-72B-Instruct, and substantial performance boosts for GLM-4-Plus and DeepSeek-V3 in the RCEM task. Overall, EmoCaustics models outperform LLM-only approaches by approximately +14% across key metrics, demonstrating the framework's effectiveness in emotion causality reasoning tasks.

Conclusion

- HELIX-6 addresses a critical gap in modeling emotional causality in ultra-long multi-party dialogues .
- Systematic evaluation of eleven mainstream large language models in emotional causality reasoning
- Significantly enhances emotional causality reasoning capabilities, enabling GLM-4-Air to outperform GPT-01 in emotional causality reasoning tasks

Methodology

My framework comprises three key modules designed to simultaneously address the tasks of emotion quintuple extraction and emotion causal chain inference:

1.**Dynamic Context Windows**: Captures long-range dependencies by effectively maintaining and processing emotional fluctuations throughout multi-turn dialogues. The sliding window at step t is defined as:

$$W_t = \{u_{\max(1,t-K+1)}, \dots, u_t\},$$

Where K is the window size, allowing for adaptive context processing.

2. **Event Storage and Updates**: Manages all events within the dialogue using an event memory bank, classifying new utterances into existing events through similarity measures. Event similarity is calculated using sentence embeddings:

$$\arg\max_{p} \cos(\phi(u_t), \phi(E_p)) \geq \tau,$$

Where τ is the similarity threshold.

3. **Event Aggregation**: Combines multiple related events into a single, coherent event, enabling more consistent modeling of emotional changes in long dialogues. The storage optimization is represented by:

$$\mathcal{E}' = \left\{ E_p \mid \frac{1}{|I_p|} \sum_{i \in I_p} \cos(\phi(u_i), \phi(E_p)) \geq \tau_{\text{retain}} \right\},$$

Where τretain is the threshold.

HELIX-6 Dataset

I constructed the HELIX-6 dataset, which extends ATLAS-6 by adding annotations for ultra-long conversations, multi-party emotional interactions, and emotional causality chains in complex scenarios. Key features include:

- 1,000 sets of longitudinal dialogues from authentic scenarios (each encompassing over 70 conversational turns)
- Complex interactive contexts such as pedagogical discussions and multi-party deliberative debates
- Diverse sources including curated public discourse from social platforms and context-specific dialogue recordings
- Four-stage annotation protocol: identification of pivotal events, documentation of emotional states, establishment of causal relationships, and analysis of participant influence patterns
- Five-category sentiment classification system (positive, negative, neutral, ambiguous, doubt)

Benchmark	**Modality**	**Scenario**	**ABSA**	**Causal chain**	**type**
CR(Blitzer et al., 2007)	Text	Sentence	✗	✗	Short
Yelp(Tang et al., 2015)	Text	Document	✗	✗	Medium
SemEval(Pontiki et al., 2014)	Text	Sentence	✗	✗	Short
TOWE(Fan et al., 2019)	Text	Sentence	✗	✗	Short
ACOS(Cai et al., 2021)	Text	Sentence	✗	✗	Short
ASTE(Peng et al., 2019)	Text	Sentence	✗	✗	Short
DiaASQ(Li et al., 2023)	Text	Dialogue	✗	✗	Very Long
Twitter2015(Ma et al., 2017)	Text, Image	Sentence	✗	✗	Long
CMU-MOSEI(Bagher Zadeh et al., 2018)	Text, Audio, Video	Sentence	✗	✗	Short
IEMOCAP(Busso et al., 2008)	Text, Audio, Video	Dialogue	✗	✗	Long
MELD(Poria et al., 2019)	Text, Audio, Video	Dialogue	✗	✗	Medium
M3ED(Zhao et al., 2022a)	Text, Audio, Video	Dialogue	✗	✗	Short
PanoSent(Luo et al., 2024a)	Text, Image, Audio, Video	Dialogue	✓	✗	Medium
NTCIR-13(Gao et al., 2017)	Text	Clause	✗	✗	Short, Medium
RECCON(Poria et al., 2017)	Text	Dialogue	✗	✓	Medium
ATLAS-6(Zhang et al., 2025)	Text, Audio, Video	Dialogue	✓	✓	Long
HELIX-6(OURS)	Text, Audio, Video	Dialogue	✓	✓	Long, Very Long

The table shows that HELIX-6 is uniquely positioned as the only dataset supporting both ABSA and causal chain analysis for Long and Very Long dialogues, addressing a critical gap in emotion causality research.