ECoK: Emotional Commonsense Knowledge Graph for Mining Emotional Gold

Zhunheng Wang[1*] **Xiaoyi Liu**[1*] **Mengting Hu**[2†] **Rui Ying**[1] **Ming Jiang**[1]
Jianfeng Wu[1] **Yalan Xie**[1] **Hang Gao**[3] **Renhong Cheng**[1]
[1]College of Computer Science, Nankai University, [2]College of Software, Nankai University
[3]College of Artificial Intelligence, Tianjin University of Science and Technology
{zh.wang, xyiliu}@mail.nankai.edu.cn, mthu@nankai.edu.cn

Abstract

The demand for understanding and expressing emotions in the field of natural language processing is growing rapidly. Knowledge graphs, as an important form of knowledge representation, have been widely utilized in various emotion-related tasks. However, existing knowledge graphs mainly focus on the representation and reasoning of general factual knowledge, while there are still significant deficiencies in the understanding and reasoning of emotional knowledge. In this work, we construct a comprehensive and accurate emotional commonsense knowledge graph, ECoK. We integrate cutting-edge theories from multiple disciplines such as psychology, cognitive science, and linguistics, and combine techniques such as large language models and natural language processing. By mining a large amount of text, dialogue, and sentiment analysis data, we construct rich emotional knowledge and establish the knowledge generation model COMET-ECoK. Experimental results show that ECoK contains high-quality emotional reasoning triples, and the performance of our knowledge generation model surpasses GPT-4-Turbo, which can help downstream tasks better understand and reason about emotions. Our data and code is available from https://github.com/ZornWang/ECoK.

1 Introduction

In the field of artificial intelligence, despite the remarkable progress we have made, a core challenge remains: how to endow machines with commonsense reasoning abilities akin to those of humans (Lake et al., 2017). When humans observe an event, they can effortlessly leverage their background knowledge, experience, and intuition to predict and reason about the unobserved causes and effects related to that event. This capability comes naturally to humans but poses a significant challenge for current AI systems. Currently, the majority of AI systems are trained on datasets and objectives tailored for specific tasks (Marcus, 2018). As these models are solely focused on completing their designated tasks, they often lack an understanding and application of the simple and interpretable commonsense knowledge that is widespread in the human world (Davis and Marcus, 2015).

Figure 1: An illustrative example of emotional commonsense knowledge reasoning in ECoK.

To address this limitation, commonsense knowledge graphs have been proposed as a solution. Commonsense knowledge graph is a graph structure that encompasses a wide range of general knowledge, providing AI systems with rich background information and contextual relationships. By integrating commonsense knowledge graphs into the training of AI systems, we can endow models with enhanced commonsense reasoning capabilities. Examples of previously proposed commonsense knowledge graphs include ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019),

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 8055–8074
August 11-16, 2024

which contain representations and inferences of numerous factual knowledge, thereby serving as valuable information sources for AI systems.

However, these knowledge graphs often emphasize the representation of factual knowledge while neglecting the description of emotions. Emotion is a multifaceted and complex psychological phenomenon that involves cognition, physiology, behavior, and the environment. The absence of knowledge in this domain leads to limited capabilities in understanding and reasoning about affective aspects. For example, as shown in Figure 1, a YouTuber may need to shoot a video. If the individual lacks the necessary equipment, they may feel *disappointed* and seek solutions such as *purchasing equipment*. If the video editing is completed, they may *sound delighted* and *feel grateful* due to *praise received*.

Therefore, there is a need for the development of emotional knowledge graphs that can capture and represent the multifaceted nature of emotions. Emotional commonsense knowledge graphs not only pay attention to factual knowledge but also emphasize the description and expression of emotions. Compared to general commonsense knowledge graphs, emotional commonsense knowledge graphs place more emphasis on portraying emotional states, emotional triggers, and emotional reactions. This more nuanced representation of emotional knowledge will greatly facilitate artificial intelligence's understanding of emotional complexity, thereby enhancing its performance in areas such as human-computer interaction, natural language processing, and intelligent recommendation.

In this work, we propose a dedicated **E**motional **Co**mmonsense **K**nowledge graph, **ECoK**. Based on research on the theory of emotional components and psychological analysis of emotions, we construct an emotional knowledge framework that focuses on seven related aspects of emotions centered around specific events: ***group***, ***causes***, ***bodily symptoms***, ***feelings***, ***expression***, ***action tendencies***, and ***advice***. We extract and generate emotional knowledge from large language models, and use professionals to check the data to build a large-scale emotional commonsense knowledge graph. Our final knowledge graph contains over 140K high-quality emotional reasoning knowledge. Additionally, we train ECoK using the COMET framework (Bosselut et al., 2019) and evaluate the performance of the generated commonsense knowledge and ECoK in downstream tasks.

In summary, our main contributions are as follows:

- We propose a novel commonsense knowledge graph, ECoK, which comprises over 140K high-quality tuples and encompasses various aspects of emotional inference knowledge.
- We establish knowledge generation model COMET-ECoK for more extensive knowledge generation. Our model outperforms all baseline models, including the current largest pretrained language model, GPT-4-Trubo.
- We conduct extensive experiments to validate the knowledge expression capabilities of ECoK. In the tasks of Emotion Recognition in Conversation (ERC) and Casual Emotion Entailment (CEE), the comprehensive performance of our knowledge graph also surpasses that of other commonsense knowledge graphs.

2 Related Work

Commonsense Knowledge Graphs Commonsense knowledge graphs can be referenced by machine learning models to handle a variety of challenging tasks, playing an important role in the field of natural language processing. Among them, the ConceptNet (Speer et al., 2017) knowledge graph consists of 36 relations and mainly contains 3.4M taxonomic and lexical knowledge and physical commonsense knowledge. The ATOMIC (Sap et al., 2019) knowledge graph consists of 9 relations and 880K tuples, containing a large amount of social commonsense. The ATOMIC$^{20}_{20}$ (Hwang et al., 2021) extends to 23 relations based on ATOMIC, containing 1.33M tuples of daily reasoning knowledge. However, among these widely used knowledge graphs, there is no explicit focus on emotions. To extend the research on emotional commonsense knowledge graphs, we propose ECoK, which can be better applied to natural language processing-related tasks.

Psychological Study of Emotion Emotion has been a fundamental area of inquiry within psychology, serving as a key factor in understanding human behavior, cognition, and social interactions (Izard, 1977). In the realm of emotional research, numerous theories and frameworks have emerged, each offering a unique perspective on the emotional landscape. For instance, the basic emotion theory posits that emotions are categorical, discrete, and universal, with a set of basic emotions such as happiness, sadness, anger, fear, surprise, and disgust serving as the foundation for more complex emotional experiences (Izard, 2007). On the other hand, the dimensional approach views emotions as varying along continuous dimensions, such as valence and arousal, allowing for a more nuanced representation of emotional states (Mehu and Scherer, 2015). The study of emotion in psychology has also examined the role of cognitive processes in emotional experiences. Cognitive appraisal theories, for example, emphasize the importance of individuals' interpretations and evaluations of events in determining their emotional responses (Frijda, 1993). These theories suggest that emotions arise from the cognitive evaluation of the personal significance and implications of events, highlighting the interplay between cognition and emotion. By incorporating these psychological perspectives, we can create more comprehensive and nuanced representations of emotional knowledge, enabling AI systems to better understand, reason about, and respond to the emotional dynamics of human experience.

Data Collection from LLMs The evolution of pre-trained generative models has paved the way for the widespread adoption of expansive Large Language Model (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Achiam et al., 2023) across numerous downstream tasks. Increasingly, researchers are leveraging these large-scale models for the purpose of dataset construction. For instance, Ruida et al. (2023) have demonstrated the effectiveness of employing larger models for the annotations of datasets aimed at smaller models, achieving superior outcomes compared to those annotated manually. Furthermore, West et al. (2022) have developed the ATOMIC^{10x} corpus, which boasts heightened accuracy and variety in comparison to the human-curated ATOMIC$^{20}_{20}$ (Hwang et al., 2021) dataset by extrapolating causal commonsense knowledge from GPT-3. Similarly, Gao et al. (2023) improved the consistency and engagement of narrative systems by prompting InstructGPT-3 (Ouyang et al., 2022) to extract personality commonsense knowledge.

3 ECoK Framework

Emotion is a multi-dimensional and multi-level psychological phenomenon, involving cognition, physiology, behavior, and environment. Firstly, the cognitive aspect of emotion includes our evaluation of things, attitude, and thinking style. Different cognitive evaluations lead to different emotional reactions (Izard, 1993). For example, positive evaluation of the same event may lead to pleasant emotions, while negative evaluation may lead to sadness or anger. Secondly, the physiological aspect of emotion involves the body's physiological response to emotions, such as accelerated heartbeat, elevated blood pressure, and sweating (Levenson, 2003). These physiological responses are closely related to emotional states, but they are also influenced by individual differences and health status. Additionally, the behavioral aspect of emotion manifests in facial expressions, body movements, and verbal expressions (Dael et al., 2012). Different emotions have different behavioral manifestations, and these behavioral expressions can also affect the emotional reactions of others. Finally, emotions are also influenced by environmental factors such as social and cultural background, interpersonal relationships, and specific situations (Thoits, 1989). Different cultures and social environments have different ways of expressing and understanding emotions, which requires us to have corresponding knowledge and abilities in different communication environments.

In this field, how to define emotion has always been a thorny problem. The number of scientific definitions proposed has reached the point of almost incalculable. Nowadays, what is generally accepted is the "component theories of emotion". In the framework of component process model, emotion is defined as a set of interrelated events (Scherer, 2001), the components of an emotional event consist of the respective states of multiple subsystems and coordinated changes over time. Generally, we do not have emotions for people or events that we do not care about. Therefore, the key point of the definition of emotional components is that the events and consequences that cause them must be relevant to the main concerns of the organism. Therefore, five components of emotion are derived: Cognitive component (appraisal), Neurophysiological component (bodily symptoms), Motivational component (action tendencies), Motor expression component (facial and vocal expression), and Subjective feeling component (emotional experience) (Scherer, 2005).

Based on the theoretical research on emotions mentioned above, we have ultimately identified seven relations for ECoK: ***group, causes, bodily symptoms, feelings, expression, action tendencies,*** and ***advice***. We will provide a detailed description of these relations below:

Group The classification of crowds that are most likely associated with the generation of certain emotions is outlined. For example, individuals likely to *experience irritability due to staying up late* include the demographic of *middle-aged* and *young adults*, *students*, as well as *working professionals*.

Causes Describe the potential factors that led to the occurrence of the emotional descriptions event. For example, the reason for *feeling irritable due to staying up late* may stem from factors such as *project deadlines* and *procrastination*.

Bodily symptoms Describe the physiological aspect of the emotional experience. For example, when *experiencing irritability*, physiological responses such as *elevated blood pressure*, *increased heart rate*, and *digestive issues* may manifest.

Feelings Describe the subjective experience of emotional state once it has occurred. For example, following *a night of staying up late*, individuals often experience feelings of *remorse*.

Expression Describe the responses and behavioral intentions accompanied by emotional states, such as facial expressions and vocal tone. For example, when individuals *experience anxiety due to staying up late*, manifestations such as *verbal tension*, *irritability*, and a *furrowed brow* commonly arise.

Action tendencies Describe the motivations that arise from emotions. For example, when individuals *experience anxiety*, actions such as *engaging in destructive behavior*, *shouting loudly* may be observed.

Advice The recommendations on how to appropriately respond to emotions are outlined. Note that the advice are generated based on the group of the target audience. For example, in the case of *students experiencing anxiety*, *formulation of a well-structured study plan* and *seeking of support from peers or elders* would be beneficial in alleviating anxiety.

4 ECoK Construction

In order to construct our emotional knowledge graph, we devise an end-to-end framework where events describing emotions are regarded as *head* entities in the graph, frame relations constitute *edge type relations*, with emotional attributes serving as *tails* in the triples (*head, relation, tail*). In Figure 2, we present an overview of our framework to construct ECoK. The head entity of ECoK is derived from two components: a) extraction from existing commonsense KGs, b) acquisition from open source social media dataset[1]. Following this, we collected *tail* entities that related to emotion attributes from pretrained LMs. Subsequently, a final knowledge graph was formed through manual verification.

4.1 Event Selection

Emotions need to be associated with specific events in some way, whether external or internal (Scherer, 2005). The head entities in ECoK consist of events containing emotional descriptions. In contrast to merely using emotions as head entities, this approach establishes connections with real-life events. In everyday situations, individuals often do not explicitly articulate their emotions. For instead, they express their emotions through complete sentences or actions. This results in emotions being frequently concealed within the descriptions of a sentence. Based on this observation, we employ events containing emotional descriptions as head entities to provide emotional context. We choose to utilize events from ATOMIC$^{20}_{20}$ (Hwang et al., 2021) that offer emotional descriptions as a part of our data source. This collection includes a significant number of coarse-grained events. Additionally, to ensure a more diverse representation of people's daily expressions, we acquire an open-source dataset[1] from Twitter via Kaggle as another data source. First, we extract emotion-oriented descriptions from ATOMIC$^{20}_{20}$ and social media datasets. Then, to reduce redundancy, we perform clustering on the merged entities. Finally, we manually review the selected head entities and extract 4.9K descriptions as the head entities of ECoK.

Social media datasets We obtain this dataset from Kaggle and process with prompting Llama-2-70B (Touvron et al., 2023). This dataset[1] is collected from Twitter (now referred to as X) for sentiment analysis, containing 1,600,000 user expressions of emotions on the platform. In this work we do not use all of them, but some of them by random

[1]https://www.kaggle.com/datasets/kazanova/sentiment140

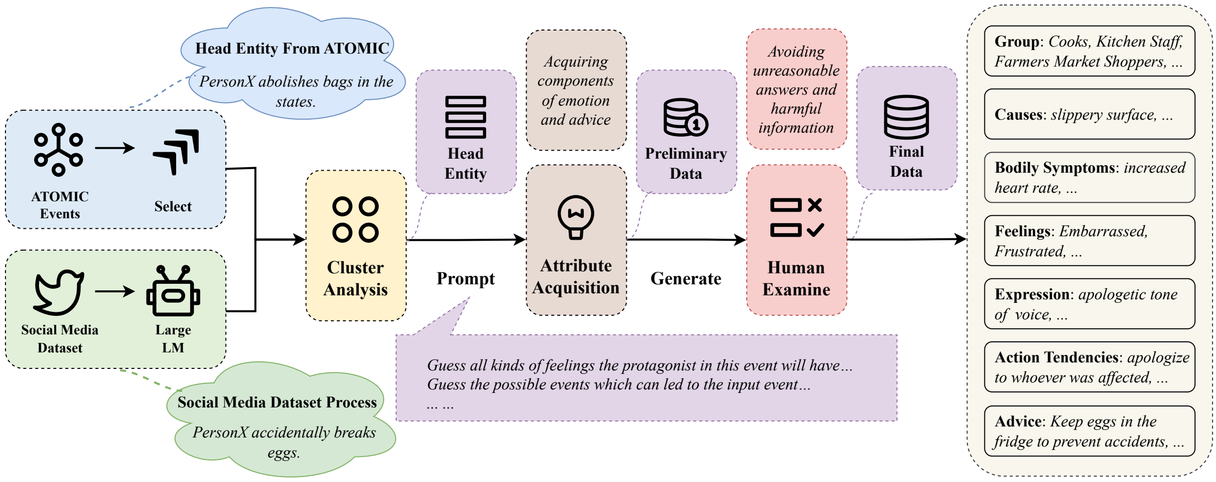

Figure 2: Overview of our steps to construct emotional commonsense knowledge graph.

sampling. We also need to take further steps due to the different dataset construction purposes. Initially, we employ a large language models (LLMs) to determine whether a given sentence included an emotional event, and in this step, we filter out harmful data, resulting in candidate raw sentences such as "*@Zyber17 i wish i didnt miss the guy anymore*". Subsequently, we instruct the LLMs to extract emotional events from these sentences, and the data underwent desensitization to transform it into a format consistent with the head entities in ATOMIC$^{20}_{20}$, "*PersonX wishes not to miss someone*".

Commonsense knowledge graph ATOMIC$^{20}_{20}$ (Hwang et al., 2021) is a commonsense knowledge graph containing 1.33M daily inferential knowledge tuples. We extract emotion-involving event entities through the *xReact* and *isBefore* relations in ATOMIC$^{20}_{20}$. The *xReact* relation captures how people react emotionally to events, while the *isBefore* relation helps in understanding the temporal sequencing of events. By exploring these relations, we can identify and extract event entities that involve emotions. Entities with the same meaning may exist in multiple ways, such as "*PersonX achieves PersonX's aims*" and "*PersonX achieves PersonX's goals*". Clustering can effectively reduce this redundancy. We use BERT (Devlin et al., 2019) to calculate the vector representation of entities and adopt a hierarchical clustering algorithm to cluster the entities. Each cluster has multiple entities, and we need to extract the content expressed by these descriptions. We calculate the centroid of each cluster and select the entity closest to the centroid as the representative of the cluster.

4.2 Attribute Acquisition

Recent studies indicate that large language models trained on extensive corpora can serve as sources for data collection. Previously, Gao et al., 2023 have successfully utilized InstructGPT-3 (Ouyang et al., 2022) to construct a persona commonsense knowledge graph. Inspired by these endeavors, we design suitable prompt[2] to guide Llama-2-70B (Touvron et al., 2023) in the construction of the ECoK. Next, we outline the detailed process of constructing ECoK through the design of prompts and human examine. Detailed templates and examples for prompts are provided in Appendix C.

Acquiring components of emotion Based on the preceding description, we devise six relationships to deconstruct emotions: group, causes, bodily symptoms, feelings, and expression. For these six relationships, we employ a consistent approach when prompting Llama-2-70B. As shown in Figure 3, each prompt is comprised three components: Instruction, Examples, and Input Text. The Instruction provides a detailed description of the relationships with additional requirements to constrain the output format and content. In the Examples section, we present instances in JSON format. Given that the specific meanings of the relationships are already outlined in the Instruction, the examples primarily serve to control the model's output format for ease of information extraction. Lastly, we provide processed events containing emotional descriptions as input.

[2]We follow the prompt format given by Hugging Face at https://huggingface.co/blog/llama2 to use Llama 2

Example Prompt for acquiring feelings

[Instruction] *Guess all kinds of feelings the protagonist in this event will have.*

[Description] *Feelings are subjective emotions experienced by people. For example: after the death of a pet, the owner's body and mind produce a series of reactions, and the subjective consciousness detects these changes and collectively refers to these reactions as "sadness"*

[Input] *PersonX is about to land in Dallas*

Large LM response

[Output] *Excited, Nervous, Hopeful, Tired*

Figure 3: Prompting LLMs for components of emotion.

Acquiring advice For advice, the purpose of designing this relations is to enable individuals to receive appropriate guidance for managing emotions and events through ECoK when they become aware of their emotions. In addition, as LLMs tend to provide generic and uninteresting responses, obtaining advice directly through prompting emotional events can result in the loss of effectiveness and diversity in recommendations. Therefore, we aim to enhance the relevance of advice by introducing *group* as an additional input, ensuring that the recommendations take into account the specific audience associated with the *group* as shown in Figure 4. Our comparative analysis indicates that the results obtained by introducing *group* as a supplementary input surpass those obtained solely by using emotional events as input in terms of effectiveness and diversity.

Prompt for acquiring advcice

[Instruction] *Give some short, useful and exclusive advice to PersonX who belongs to a specific group is meeting the input event.*

[Event] *PersonX plans to play guitar but their video editing plan was unsuccessful due to lack of necessary equipment.*

[Group] *YouTubers*

Large LM response

[Output] *Invest in a good quality camera and microphone*

Figure 4: Prompting LLMs for advice.

Relation	Tail	Distinctive Tail	Words
Group	16332	3386	1.8
Causes	23997	20126	5.4
Bodily symptoms	15717	1480	2.1
Feelings	24203	950	1.0
Expression	23699	3537	2.1
Action tendencies	16148	9546	3.8
Advice	23909	21367	8.8
Total	144005	60392	3.7

Table 1: Statistics of ECoK. Words represents the average number of words per tail.

4.3 Human Examine

To reduce redundancies and errors, and further improve data quality, we adopt a combination of prompting and manual inspection. We prompt the LLMs to avoid generating controversial content and ensure that the answers do not contain any harmful, unethical, racist, sexist, or illegal content. We generate answers in batches and filter out duplicate answers after generation. For each batch, we sample 1k tuples and manually inspect them according to the following criteria: a) relevance of the answer, avoiding unreasonable answers, b) safety of the answer, avoiding harmful information, c) authenticity of the answer, avoiding false information. We fine-tune the prompt based on the inspection results until the manual sampling inspection is qualified.

4.4 ECoK Statistics

Table 1 presents the statistical information of ECoK. Our knowledge graph comprises over 140K emotional inference tuples, encompassing a wide range of emotions from the commonly recognized ones like happiness, anger, sadness, and joy to more complex psychological states such as anxiety and depression. This diversity mirrors the richness and complexity of human emotions. Furthermore, the *Advice* relationship within ECoK offers over 20K distinct personalized suggestions for coping with different emotions. These recommendations range from seeking support and adjusting cognitions to relaxation techniques. In conclusion, ECoK showcases the complexity and diversity of emotions through its multifaceted relationships. It serves as a comprehensive and systematic knowledge graph of emotional commonsense, providing valuable insights into the intricate world of human emotions.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	SPICE	SkipThought
Llama-2-13B	77.99	74.76	72.46	70.04	56.31	74.26	41.48	72.38
Llama-2-70B	80.88	78.25	76.25	74.64	61.41	77.80	47.04	76.54
GPT-3.5-Turbo	79.13	76.10	74.08	72.02	57.10	75.54	44.98	72.53
GPT-4-Turbo	80.19	77.17	75.02	72.87	59.17	76.32	43.21	75.23
COMET-ECoK	**85.98**	**83.50**	**81.63**	**79.83**	**69.39**	**83.62**	**49.06**	**79.82**

Table 2: Automatic evaluation results of emotion knowledge generation. The optimal outcomes are indicated in bold. All baseline model results are based on 5-shot settings, utilizing five randomly sampled facts from the training set as prompts.

Models	Accept	Reject	No Judgement
Llama-2-13B	85.02	11.24	3.75
Llama-2-70B	84.27	11.99	3.75
GPT-3.5-Turbo	86.14	10.86	3.00
GPT-4-Turbo	88.39	8.24	3.37
COMET-ECoK	**90.26**	**7.12**	**2.62**

Table 3: Human evaluation results of emotion knowledge generation. The optimal outcomes are indicated in bold. The crowdworkers evaluate each inference by providing three options: *Accept* (%), *Reject* (%), and *No Judgement* (%) to determine.

5 Experiments

5.1 Emotion Knowledge Generation

While various approaches are employed in the construction of ECoK to collect events describing emotions through multiple ways, the acquisition of emotional knowledge is still limited to querying situations similar to those represented in the knowledge graph. However, this approach has its inherent limitations, as ECoK cannot encompass the infinite diversity of the real world. Following the methodology proposed by Hwang et al., 2021, we opt to partition the ECoK dataset and employ a training set($\sim$100K inferences) to train the COMET (Bosselut et al., 2019) knowledge generator based on BART (Lewis et al., 2020) to handle unknown situations and expand the application scope of ECoK. After training, the model is capable of generating corresponding *tail* attributes based on the given *head* containing emotional descriptions and the *relation* in the ECoK.

To evaluate our model, we conduct a knowledge generation comparison on the ECoK test set with two versions of the currently prominent open-source LLMs Llama 2 (Touvron et al., 2023), versions 13B and 70B, as well as two widely acknowledged powerful LLMs, GPT-3.5 turbo, and GPT-4 turbo. We use five facts randomly sampled on the training set to prompt both two versions of Llama 2 and GPT. We conduct automatic and human evaluations on the generated results, and the results are presented in Table 2 and 3.

In cases where the model scale is significantly smaller than the baseline, COMET trained on ECoK, demonstrates superior performance across various Natural Language Generation (NLG) metrics, as assessed through automated evaluations. In addition, we sought to ascertain the acceptability of content generated by our trained COMET. Therefore, we sample the output of each model for manual evaluation. In the results of the manual assessment, COMET-ECoK exhibits the highest *Accept* rate and the lowest *Reject* rate. This indicates that the results generated by COMET-ECoK we trained are more likely to be accepted by humans. All the aforementioned findings collectively suggest that ECoK can serve as a reliable emotional knowledge base, enabling light-scale LMs to rival LLMs in the generation of emotional knowledge.

5.2 Enhancing Emotional Reasoning

ECoK is rich in emotional knowledge, so we will test whether it can provide better results in downstream tasks. We apply ECoK to enhance the emotional reasoning model MHGT based on the BHG method (Yang et al., 2023) and evaluate its effectiveness in two tasks: Emotion Recognition in Conversations (ERC) and Casual Emotion Entailment (CEE). By incorporating ECoK into MHGT, we aim to improve the model's ability to understand and reason about emotions in natural language texts, particularly in dialogue settings where emotions play a crucial role.

The main steps of emotional knowledge enhancement are as follows: First, knowledge extraction, target sentences are queried to obtain emotional knowledge. This process involves searching and retrieving relevant emotional knowledge from the ECoK-trained COMET model (Bosselut et al., 2019).The retrieved knowledge can include

Task	ERC					CEE		
Dataset	IEMOCAP	MELD	DailyDialog		EmoryNLP	RECCON		
Model	Weighted F1	Weighted F1	Micro Fl	Macro F1	Weighted Fl	Neg.Fl	Pos.Fl	Macro F1
MHGT+ConceptNet	70.55	65.99	61.41	52.70	38.07	89.53	68.60	79.07
MHGT+ATOMIC	71.07	65.57	61.50	53.00	38.62	90.26	68.85	79.55
MHGT+ATOMIC$^{20}_{20}$	71.20	65.54	62.37	54.11	39.06	90.34	69.13	79.73
MHGT+ECoK	**71.50**	**66.30**	**62.98**	**54.17**	**39.37**	**90.59**	**69.44**	**80.02**

Table 4: Test results of all models on the five datasets of ERC task and CEE task. We use the model provided by the MHTG paper and leverage ECoK, which has been trained using COMET, to augment four datasets. For ConceptNet, ATOMIC and ATOMIC$^{20}_{20}$, we employ the pre-trained data from the MHTG paper.

information about emotional states, causes, effects, and relationships with other concepts or entities. For example, by querying the sentence "*I'll get married.*", we obtain emotional knowledge "*to be loved*" and "*feel happy*". Second, knowledge filtering, selecting the most relevant part of the extracted knowledge. This step employs semantic-aware knowledge filtering method, which analyzes the semantic relationships between the extracted knowledge and the target utterance to determine which pieces of information are most pertinent to the task. By filtering out irrelevant or redundant knowledge, we can ensure that the model focuses on the most useful and informative aspects of the emotional knowledge. Finally, knowledge interaction involves integrating the filtered knowledge into the model to enhance its reasoning capabilities. This step aims to effectively combine the extracted and filtered emotional knowledge with the model's internal representations. The enhanced model can leverage external knowledge sources during inference to improve their understanding and reasoning about emotional content and make more informed decisions.

For the ERC task, we use ECoK to enhance four datasets: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), DailyDialog (Li et al., 2017), and EmoryNLP (Zahiri and Choi, 2018). We select the Weighted F1 score as the evaluation metric for IEMOCAP, MELD, and EmoryNLP. These datasets are widely used in emotion recognition research, and the Weighted F1 score is a commonly accepted metric for evaluating performance in multi-class classification tasks. Since DailyDialog contains a large number of "neutral" emotions, we choose the Micro F1 score to evaluate the performance of non-neutral emotions and calculate the Macro F1 score to evaluate overall performance.For the CEE task, we enhance the RECCON (Poria et al., 2021) dataset and use the F1 score on positive and negative utterances, as well as the Macro F1 score, as evaluation metrics.

The enhanced model is denoted as MHGT+ECoK. To compare the capabilities of emotional knowledge graphs with other commonsense knowledge graphs, we selected MHGT+ConceptNet (Speer et al., 2017), MHGT+ATOMIC (Sap et al., 2019) and MHGT+ATOMIC$^{20}_{20}$ (Hwang et al., 2021) as baseline models. The results of our MHGT+ECoK and all baseline models on the two tasks is shown in Table 4. ECoK achieves the best comprehensive performance on five datasets while having a lower construction cost than other commonsense knowledge graphs with artificial annotations. These results demonstrate the effectiveness of incorporating emotional knowledge from ECoK into the MHGT model and highlight the importance of emotional commonsense knowledge in enhancing the performance of emotional reasoning models.

6 Conclusion

In this work, we have successfully constructed ECoK, a comprehensive and accurate emotional commonsense knowledge graph. ECoK comprises over 140K high-quality emotional inference knowledge tuples, representing the complexity and diversity of emotions through seven distinct relationships: ***group***, ***causes***, ***bodily symptoms***, ***feelings***, ***expression***, ***action tendencies***, and ***advice***. Furthermore, we leverage ECoK to build a knowledge generation model called COMET-ECoK, designed to facilitate more extensive knowledge generation. Through rigorous evaluation, we have shown that incorporating emotional commonsense knowledge from ECoK enhances the performance of Knowledge Generation, Emotion Recognition in Conversations and Casual Emotion Entailment tasks. These results validate the effectiveness of our approach and its potential to revolutionize human-machine interaction. In the future, we will continuously expand the emotional knowledge graph to explore more nuanced emotional representations and further enhance the capabilities of artificial intelligence systems in emotion-related tasks.

Limitations

While our work has demonstrated the potential of constructing a comprehensive emotional commonsense knowledge graph, it is important to acknowledge several limitations.

Firstly, our approach relies heavily on the quality and diversity of the text, dialogue, and sentiment analysis data used for mining emotional knowledge. Any biases or inconsistencies in these data sources can potentially affect the accuracy and completeness of ECoK.

Secondly, although we have integrated cutting-edge theories from multiple disciplines, the current version of ECoK may not capture the full complexity and nuances of human emotions. Future iterations could benefit from incorporating more fine-grained emotional representations and models.

Finally, the evaluation of ECoK's performance in downstream tasks is limited to specific benchmarks and may not generalize to all possible scenarios.

Despite these limitations, we believe that ECoK represents a valuable step towards enhancing emotional understanding and reasoning in artificial intelligence systems.

Ethics Statement

In the construction of our ECoK, we have adhered strictly to ethical principles and best practices. We recognize the sensitivity of emotional data and the importance of protecting privacy. Therefore, we have taken careful measures to filter and exclude any data containing potentially private or identifying information from our analysis. Furthermore, we have ensured that all data sources used in this work are publicly available and have been obtained through legitimate means, respecting copyright and intellectual property rights. We are committed to transparency in our methods and have documented our data collection, preprocessing, and filtering procedures comprehensively. Additionally, we acknowledge the potential impact of our work on society and strive to use our findings ethically, promoting positive outcomes for individuals and communities. We emphasize that the use of ECoK should adhere to ethical guidelines and be subject to appropriate oversight and accountability mechanisms.

Acknowledgements

We sincerely thank all the anonymous reviewers for providing valuable feedback. This work is supported by the youth program of National Science Fund of Tianjin, China (Grant No. 22JC-QNJC01340).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, pages 1–100.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam,The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 382–398.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*, 42:335–359.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Nele Dael, Marcello Mortillaro, and Klaus R Scherer. 2012. Emotion expression in body action and posture. *Emotion*, 12(5):1085.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Nico H Frijda. 1993. The place of appraisal in emotion. *Cognition & Emotion*, 7(3-4):357–387.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6569–6591, Toronto, Canada. Association for Computational Linguistics.

Silin Gao, Jena D Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. Comfact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, pages 1–21.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Carroll E Izard. 1977. *Human emotions*. Springer Science & Business Media.

Carroll E Izard. 1993. Four systems for emotion activation: cognitive and noncognitive processes. *Psychological review*, 100(1):68.

Carroll E Izard. 2007. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3):260–280.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28:3294–3302.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Robert W Levenson. 2003. Blood, sweat, and fears: The autonomic architecture of emotion. *Annals of the New York Academy of Sciences*, 1000(1):348–366.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. Association for Computational Linguistics.

G Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631:1–27.

Marc Mehu and Klaus R Scherer. 2015. Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, 15(6):798.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

WANG Ruida, Wangchunshu Zhou, and Mrinmaya Sachan. 2023. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11831.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Klaus R Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, 92(120):57.

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, pages 4444–4451.

Peggy A Thoits. 1989. The sociology of emotions. *Annual review of sociology*, 15(1):317–342.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, pages 1–77.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Shaoxiong Ji, and Sophia Ananiadou. 2023. A bipartite graph is all we need for enhancing emotional reasoning with commonsense knowledge. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2917–2927.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*, pages 44–52.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14020–14028.

A ECoK Construction Details

A.1 ECoK Head Selection

The head entities of ECoK originate from two parts: a) the commonsense knowledge graph, b) social media datasets. Table 6 displays some examples of the prompts used to select data from social media datasets and format transformation described in Sec. 4.1 using LLMs. Specifically, we first let the LLMs judge if the sentence contains emotional description, which allows us to filter out irrelevant data from social media datasets, such as data we consider useless like "*watching 'House'*". These types of data are filtered out at this step. Then, we use the LLMs to convert the selected sentences into the data format of ATOMIC$^{20}_{20}$ (Hwang et al., 2021); that is, we replace the agent of action in the sentences with "PersonX", and use "Person(Y to Z)" to represent other individuals mentioned in the sentence, for example, transforming the original data "*need a huge*" into "*PersonX needs a huge*". During this process, as the data comes from social media, it includes the "@" symbol to mention individuals. We address this by adding words that start with the "@" character to the LLMs bad words list and by including explanations in the prompts to ensure that the processed sentences do not contain the "@" symbol.

During this extraction process, we encounter entities with semantically similar meanings but expressed in different ways. For instance, "*PersonX achieves PersonX's aims*" and "*PersonX achieves PersonX's goals*" convey the same concept but use different terminologies. To address this redundancy, we employ clustering techniques. Before clustering, we need to represent the entities in a form that captures their semantic meaning. We utilize BERT (Devlin et al., 2019) to calculate the vector representations of the extracted entities. BERT's ability to understand contextualized word meanings makes it an ideal choice for this task.With the vector representations obtained from BERT, we proceed to cluster the entities. We adopt a hierarchical clustering algorithm for this purpose. Hierarchical clustering algorithms group data points into a hierarchy of clusters, where each cluster at a given level is derived from the clusters at the lower level. Once the clustering is complete, we have multiple clusters, each containing multiple entities.

To summarize the content expressed by these entities within a cluster, we calculate the centroid of each cluster. The centroid is a representative point that minimizes the distance to all other points in the cluster. We then select the entity closest to this centroid as the representative of the cluster. This representative entity captures the essential meaning expressed by the entire cluster, serving as a concise summary.

A.2 Advice Process Detail

For *advice* generation, we introduce the concept of a *group* as a reference, allowing LLMs to generate suggestions for specific *group*. During this process, there may be some suggestions that are semantically redundant across different groups, which is due to certain suggestions being applicable to more than one *group*. To address this, for a given head, we first gather all the *group*, then collect all the *advice* for these *group*. Finally, we compile these pieces of *advice* and use the prompt from Table 11, prompting the LLMs to deduplicate them to avoid the occurrence of advisories with similar meanings.

B Parameters for LLMs

During the data construction process, we read the license terms[3] and follow the acceptable use policy[4] to obtain a commercial license for Llama 2 (Touvron et al., 2023), and implemented Llama2-70B-Chat using Hugging Face. The details of the parameters we used for Llama2-70B-Chat in constructing ECoK and in our experiments are given in Table 5. For the GPT series of models, we read the terms of service[5] and follow the usage policies[6], and we use the default hyperparameters provided by the OpenAI platform.

C Prompts for Attribute Acquisition

In Tables 10 and tables 12 to 18, we present the original input we used to prompt Llama2-70B-Chat, with an input format consistent with the templates used for model training as given in the Llama 2 (Touvron et al., 2023) paper. Table 10 displays the prompts used for selecting data from social media datasets and for format transformation. Table tables 12 to 18 shows the prompts used to obtain relations.

[3]https://ai.meta.com/llama/license/.

[4]https://ai.meta.com/llama/use-policy/.

[5]https://openai.com/api/policies/service-terms/.

[6]https://beta.openai.com/docs/usage-policies.

Parameter	Construction	Experiments
max_new_tokens	2048	2048
num_return_sequences	1	1
repetition_penalty	1.2	1.2
temperature	0.1	1
top_p	0.75	0.75
top_k	40	40
num_beams	4	1
do_sample	True	True

Table 5: Parameters for Llama2-70B-Chat

D Emotion Knowledge Generation Details

D.1 Evaluatin Details

We divided the facts of ECoK into three sets, with size 101086, 20018 and 20006 for training, validation and testing, respectively. Please note that there are no overlapping head entities among these three sets of data. The automatic evaluation employed the following several NLG metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) , METEOR (Banerjee and Lavie, 2005), SPICE (Anderson et al., 2016), and SkipThought (Kiros et al., 2015). For human evaluation, we employed three experts from various regions in the pertinent domain as reviewers, and they were compensated with wages exceeding the local minimum wage standard. Each generated tail entity was assessed by the three experts according to the guidelines displayed in Figure 5. The *Accept* rate, *Reject* rate, and *No Judgement* rate were calculated individually by each expert, and the average value of these was taken as the final result.

D.2 Model Training

We trained COMET-ECoK using the code from the COMET repository[7], along with the default hyperparameter configuration. We utilize a single Nvidia RTX A6000 GPU with a batch size set to 32, conducting training for 50 epochs, which costs about 6 hour to get the highest ROUGE-L score on the validation set. In addition, in line with the practice from PEACOK, we also trained a DeBERTa (He et al., 2020) discriminator to re-rank the results generated by COMET-ECoK, Llama2-13B, Llama2-70B, GPT-3.5-turbo, and GPT-4-turbo. For the training process, we created negative examples

[7]https://github.com/allenai/comet-atomic-2020.

Select raw data from social media datasets	
Input: Hey hey de e inte okey.. Answer: false	Input: previewing this meth and red album Answer: false
Input: I love the fact that I am already up Answer: True	Input: Is sad that jake is leaving Answer: True
Input: bed. have to be at the courthouse at 10am Answer: false	Input: @lalavazquez no don't wear it Answer: false
Input: @CityLightFade I'm so jealous i can't be there Answer: Ture	Input: I'm excited that Dad said he would buy me a mac Answer: Ture
Format transformation	
Input: id sad that jake is leaving Answer: PersonX feels sad that PersonY is leaving	Input: I love the fact that I am already up Answer: PersonX love getting up
Input: @CityLightFade I'm so jealous i can't be there Answer: PersonX is jealous that can't be somewhere	Input: I'm excited that Dad said he would buy me a mac Answer: PersonX is excited about getting a mac
Input: @gilneas ooh that's going to be fun! enjoy Answer: PersonX tells PerosnY to enjoy the fun event	Input: @tuftedpuffin Oh dear.. that's no good Answer: PersonX responds negatively to PersonY
Input: I dont wanna care what others think of you Answer: PersonX doesn't care what other think	Input: @michaelalacey Thanks for your support Answer: PersonX thanks PersonY for support

Table 6: Some examples of select header entities and format handling

by replacing the *tail* entity of one *head* entity with a randomly sampled *tail* entity from the same *relations* of another *head* entity. We trained DeBERTa using the hyperparameters recommended by the ComFact (Gao et al., 2022) benchmark and based on a binary classification loss, enabling it to distinguish between real and negative data. The training was performed using two Nvidia RTX A6000 GPUs, and the highest F1 score was obtained at the 8th epoch, with about 18 hours. We had all models generate 5 *tail* entities for a *head* entity and a relation, and we used the DeBERTa discriminator to rank these 5 generated *tail* entities, selecting the first one for evaluation.

D.3 Case Study

Table 7 presents 5 examples generated by COMET-ECoK on the test dataset data. Upon observation, it can be found that COMET, when trained with ECoK as the knowledge base, is capable of generating effective emotional reasoning data.

E Enhancing Emotional Reasoning Details

In this experiment, we aim to enhance the emotional reasoning model MHGT, which is based on the Bipartite Heterogeneous Graph (BHG), by leveraging ECoK rich in emotional knowledge (Yang et al., 2023).

E.1 Datasets

We conducted evaluations on two tasks: Emotion Recognition in Conversations (ERC) and Casual Emotion Entailment (CEE). A total of five benchmark datasets were employed for comparison against baseline models.

IEMOCAP (Busso et al., 2008) This is a multimodal dataset. The content is derived from the lines in the scripts of the two actors, and the emotional tags included are excited, neutral, frustrated, sad, happy, and angry.

MELD (Poria et al., 2019) This dataset comes from the dialogue content of the characters in the American drama Friends. The pre-defined emotions are neutral, sad, anger, disgust, fear, happy, surprise.

DailyDialog (Li et al., 2017) Manually compiled data sets about daily communication. The pre-defined emotion labels are neutral, happy, surprise, sad, anger, disgust, fear.

EmoryNLP (Zahiri and Choi, 2018) This dataset also comes from Friends, and the difference from MELD is the annotation of utterance's

Head	Relation	Tail
PersonX agree with PersonY's logic	Group	Colleagues
	Causes	PersonY presents compelling arguments
	Bodily symptoms	relaxed breathing
	Feelings	Relieved
	Expression	nodding
	Action tendencies	approach PersonY
	Advice	Be aware of potential biases in data and assumptions
PersonX is late for work and hungry	Group	Employees
	Causes	difficulty sleeping
	Bodily symptoms	stomach growling
	Feelings	Stressed
	Expression	rushing footsteps
	Action tendencies	rush to get ready
	Advice	Leave earlier than usual to account for unexpected delays
PersonX starts school	Group	Children
	Causes	turned 5 years old
	Bodily symptoms	excitement-induced trembling
	Feelings	Nervous
	Expression	hesitant tone of voice
	Action tendencies	get excited about school clothes
	Advice	Establish a routine for homework and study time
PersonX feel tired from rain or oversleeping	Group	Night owls
	Causes	alarm clock didn't go off
	Bodily symptoms	yawns
	Feelings	Irritable
	Expression	rubbing eyes
	Action tendencies	take a shower
	Advice	Use caffeine strategically to boost energy levels
PersonX working late and camping out	Group	Outdoor enthusiasts
	Causes	deadline for project is nearing
	Bodily symptoms	fatigue
	Feelings	Accomplished
	Expression	contented sigh
	Action tendencies	get a snack or drink
	Advice	Invest in a portable stove for cooking meals

Table 7: Case study of Emotion Knowledge Generation. By providing the head entity, the trained model COMET-ECoK generates the 7 types of relations in ECoK.

emotional label category. The emotional tags contained in this dataset are: joyful, neutral, powerful, mad, sad, scared, and peaceful.

RECCON (Poria et al., 2021) A dataset sampled from ERC dataset DailyDialog with both utterance-level emotion labels and binary emotion cause labels.

E.2 Emotional Knowledge Enhancement

To further improve the quality and adaptability of ECoK, we employ the COMET framework for training enhancement. During the enhancement process, we treat the emotional knowledge entries in ECoK as inputs and utilize the COMET model for reasoning and generating emotional knowledge. Through this approach, we can expand the emotional knowledge in ECoK, adding new emotional relationships, attributes, and scenarios, thus making the dataset more extensive and diverse.

Then we integrated the ECoK dataset enhanced by COMET training into the MHGT model. Specifically, we added the enhanced emotional knowledge as extra nodes or edges to the BHG, enriching the model's input information. This approach allowed the model to consider both textual information and emotional knowledge during the reasoning process, improving the accuracy of emotional inference.

BHG provided two additional knowledge aggregation node types designed to automatically perform knowledge filtering and interaction, ensuring effective screening and integration of relevant information during knowledge infusion. The MHGT model retained the ability to preserve consistent feature spaces and unequal dimensions for heterogeneous node types during inference, preventing

Dataset	IEMOCAP	MELD	DailyDialog	EmoryNLP
Model	Weighted F1	Weighted F1	Micro F1	Weighted F1
Llama-2-7B	34.99	30.56	27.03	13.98
Llama-2-7B+ECoK	36.59	34.72	37.04	17.36
MHGT+ECoK	**71.50**	**66.30**	**62.98**	**39.37**

Table 8: Test results of all models on the four datasets of ERC task.

Model	Neg. F1	Pos. F1	Macro F1
KBCIN+ATOMIC$^{20}_{20}$	-	67.51	78.43
KBCIN+ECoK	89.21	69.39	79.39
MHGT+ECoK	**90.59**	**69.44**	**80.02**

Table 9: Test results of all models on the dataset of CEE task.

unnecessary loss of information. These advantages facilitated the model's easy extension to multi-type and multi-grained knowledge sources, making the integration of ECoK relatively straightforward and efficient.

Through knowledge aggregation nodes, the emotional knowledge from ECoK was effectively combined with other information in the model. During training, we closely monitored the model's ability to retain feature spaces and unequal dimensions, ensuring that the enhancement of emotional reasoning capabilities did not sacrifice the information retention advantages of the original model. Additionally, we employed appropriate optimization strategies to improve the model's training efficiency and performance.

The experimental results demonstrated that the ECoK dataset enhanced by COMET training significantly improved the performance of the MHGT model on both ERC and CEE tasks. This indicated that the enhanced dataset not only enriched the model's input information but also effectively boosted its emotional reasoning capabilities.

E.3 Additional Experimental Data

On the basis of the initial experiments, we conducted additional experiments to compare with larger models and other new model performances. The results of the experiments are shown in the table 8 and 9.

We applied ECoK over Llama-2-7b (Touvron et al., 2023) and KBCIN (Zhao et al., 2023). From the results, it can be seen that after applying ECoK as an external knowledge source, Llama-2-7B achieves a considerable performance improvement on all four datasets of the ERC task. At the same time, in comparison with ATOMIC$^{20}_{20}$, KBCIN+ECoK also achieved better performance on the CEE task. Combined with the existing experiments in Section 5.2 of this thesis and these two supplementary experiments, it is well demonstrated that ECoK can be applied as an external knowledge base to enhance the performance of downstream tasks.

Prompt for data selection from social media datasets
<s>[INST] «SYS» You are a data analysis assistant, do a sentence filter job. Pick out sentences that reflect human emotions and human behavior, return true in output. If this condition is not met, return false in output. And exclude the sentences include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content by return false in output json string. Only output one json string with original input text and filter flag follow the examples format, no need explain! Don't include examples in your answer! Examples: {"text": "I'm lonely", "filter": true}, {"text": "@twittera que me muera ? ", "filter": false}, {"text": "It's gonna be like 95 and no rain all week.", "filter": false} «/SYS» text: {{ content1 }} [/INST]
Prompt for format transformation
<s>[INST] «SYS» Use a concise and short sentences to highly summarize just one main events in the input text. In the summary sentence, you can ignore some specific events and only focus on the events that contain emotions. To desensitize sentences, use Person(X Z) to anonymize the names or persons in the sentences, do not include specific names or persons in summarize. Notice that this sentences may come from social media, so it may contain @ symbol, which is used to mention someone. Don't use the words after @. You may replace it with PersonY in some situation. For example, "@richstep51 Thanks" can be summarized as "PersonX thanks PersonY" where "@richstep51" is considered a person's name and replaced by "PersonY". Return both the input text and the summarize text in json string follow the examples format. Only output json string, no need explain! Examples: {"text": "i wanna drink..", "summarize": "PersonX wanna drink"}, {"text": "I'm lonely", "summarize": "PersonX feel lonely"}, {"text": "@RoisinMcK good idea, i'll do that ", "summarize": "PersonX agrees with PersonY's idea"}, {"text": "@craigeryowens wow have fun...I wishi could go to the warped tour this summer! But I can't! I hope to see u in concert soon! Luv Katelyn!", "summarize": "PersonX want to go to the music festival"} «/SYS» text: {{ content1 }} [/INST]

Table 10: Prompts for selecting data from social media datasets and for format transformation

Prompt for remove duplicates of *Advice*
<s>[INST] «SYS» Remove items from the list that have a similar or same meaning to other items. For similar items, just keep one. Please only output json string follow the examples format, no need explain! Don't include examples in your answer! Output format: {"deduplication": array} «/SYS» Input: {{ content1 }} [/INST]

Table 11: Prompt for remove duplicates of *Advice*

Prompt for acquiring *Group*
<s>[INST] «SYS» Guess which crowds of people who have some common factors that are more likely to experience the input events. Find the groups of people who have some common factors that can lead to this kind of events. Only output one json string follow the examples format, no need explain! Don't include examples in your answer! Examples: {"event": "PersonX got his computer dead", "groups":["IT professionals", "Staff", "Student", "Programmer"]}, {"event": "PersonX will miss this place", "groups":["Traveller", "Graduates"]} «/SYS» event: {{ content1 }} [/INST]

Table 12: Prompt for acquiring *Group*

Prompt for acquiring *Causes*
<s>[INST] «SYS» Guess the possible events which can led to the input event. In the causes returned, Use sentences without subjects(e.g. PersonX) and objects(e.g. PersonY, PersonZ) as much as possible. Only output one json string follow the examples format, no need explain! Don't include examples in your answer! Examples: {"event": "PersonX wants to have a drink", "causes":["eating something salty or spicy", "it was hot outside", "feel bored", "just woke up", "exercised or did physical activity"]}, {"event": "PersonX will miss this place", "causes":["moves to a new city/country", "somewhere goes out of business", "health issues or mobility challenges"]} «/SYS» event: {{ content1 }} [/INST]

Table 13: Prompt for acquiring *Causes*

Prompt for acquiring *Bodily symptoms*
<s>[INST] «SYS» Guess all kinds of physical reaction the protagonist in the input text will have. Bodily symptoms also means physical reaction. Physical reaction is The physiological component of emotion, the automatic reaction of the body that allows the subject to adapt to the unexpected situation. Return without subjects(e.g. PersonX) and objects(e.g. PersonY, PersonZ) as much as possible. Only output one json string follow the examples format, no need explain! Don't include examples in your answer! Examples: {"event": "PersonX realizing that the pet's death is irreversible", "bodily_symptoms":["reduced nervous system arousal", "general weakness", "slowed heart rate"]}, {"event": "PersonX want to eat schnitzel in December", "bodily_symptoms": ["increased appetite", "salivation", "stomach grumbling"]} «/SYS» event: {{ content1 }} [/INST]

Table 14: Prompt for acquiring *Bodily symptoms*

Prompt for acquiring *Feelings*

<s>[INST] «SYS»
Guess all kinds of feelings the protagonist in this event will have.
Feelings are subjective emotions experienced by people. for example: after the death of a pet, the owner's body and mind produce a series of reactions, and the subjective consciousness detects these changes and collectively refers to these reactions as "sadness").
Return without subjects(e.g. PersonX) and objects(e.g. PersonY, PersonZ) as much as possible.
Only output one json string follow the examples format, no need explain! Don't include examples in your answer!
Examples:
{"event": "PersonX is about to land in Dallas", "feelings": ["Excited", "Nervous", "Hopeful", "Tired"]},
{"event": "PersonX spend entire break on Twitter with no tweets", "feelings": ["Bored", "Lonely", "Restless", "Disappointed", "Anxious"]}
«/SYS»
event: {{ content1 }} [/INST]

Table 15: Prompt for acquiring *Feelings*

Prompt for acquiring *Expression*

<s>[INST] «SYS»
Guess all kinds of expression the protagonist in input event will have.
Expression is facial and voice changes show the person's emotions. This is to convey the emotional subject's view of an event and his action intentions to the people around him. The expression of emotions has elements that are common to all humans, as well as elements that are unique to each place.
Return without subjects(e.g. PersonX) and objects(e.g. PersonY, PersonZ) as much as possible.
Only output one json string follow the examples format, no need explain! Don't include examples in your answer!
Examples:
{"event": "PersonX realizing that the pet's death is irreversible", "expression":["frowning", "mouth corners downward", "crying"]},
{"event": "PersonX want to play game","expression": ["smiling","eyes wide open","excited tone of voice","energetic body language","eagerness","enthusiasm"]}
«/SYS»
event: {{ content1 }} [/INST]

Table 16: Prompt for acquiring *Expression*

Prompt for acquiring *Action tendencies*
<s>[INST] «SYS» Guess all kinds of action tendencies the protagonist in the input text will have. Action tendencies: Emotions create motivations (for example: when you are sad, you want someone to talk to, and when you are angry, you do something you would not normally do.). Think more about physical actions rather than mental activities. Treat the text as one event for analyze. Return without subjects(e.g. PersonX) and objects(e.g. PersonY, PersonZ) as much as possible If there is an object, use "someone" to instead. Only output one json string follow the examples format, no need explain! Don't include examples in your answer! Examples: {"event": "PersonX feel sad", "action_tendencies": ["talk to someone", "want to be alone"]}, {"event": "PersonX feel angry", "action_tendencies": ["do something he would not normally do", "swear loudly"]} «/SYS» event: {{ content1 }} [/INST]

Table 17: Prompt for acquiring *Action tendencies*

Prompt for acquiring *Advice*
<s>[INST] «SYS» Give some short, useful and exclusive advice to PersonX who belongs to a specific group is meeting the input event. Please try to only give advice that is specific to this group of people and exclude general advice. Only output one json string follow the examples format and no need explain! Examples: {"event":"PersonX lost train of thought", "group":"Elderly","advice":["Ask young people for help","Try retracing your steps to remember","Stay mentally active by doing crossword puzzles or reading"]} «/SYS» event: {{ content1 }}, group: {{ content2 }} [/INST]

Table 18: Prompt for acquiring *Advice*

Emotion Knowledge Generation Evaluate Guide

Thanks for participating！We need you to evaluate the given data for us. We will use this data to assess our model.

The table consists of three parts in the knowledge graph (head, relations, tail). **The head represents the event, relations are the relationships, and tail represents the data to be evaluated.**

Explanation of relations:

1. **Group**: The most likely audience to experience this event. For example, individuals likely to *experience irritability due to staying up late* include the demographic of *middle-aged* and *young adults*, *students*, as well as *working professionals*.
2. **Cause**: The cause that leads to the event. For example, the reason for *feeling irritable due to staying up late* may stem from factors such as *project deadlines* and *procrastination*.
3. **Bodily symptoms**: The physiological aspects of emotions, the physical reactions a person would have after experiencing the event. For example, when *experiencing irritability*, physiological responses such as *elevated blood pressure*, *increased heart rate*, and *digestive issues* may manifest.
4. **Feelings**: The subjective emotions experienced by people. For example, following *a night of staying up late*, individuals often experience feelings of *remorse*.
5. **Expression**: Facial expressions, voice, and other forms of emotional expression. For example, when individuals *experience anxiety due to staying up late*, manifestations such as *verbal tension*, *irritability*, and a *furrowed brow* commonly arise.
6. **Action tendencies**: Motivations generated by emotions. For example, when individuals *experience anxiety*, actions such as *engaging in destructive behavior*, *shouting loudly* may be observed.
7. **Advice**: Advice about the event for specific audience groups. For example, in the case of *students experiencing anxiety*, *formulation of a well-structured study plan* and *seeking of support from peers or elders* would be beneficial in alleviating anxiety.

Please judge whether the tail corresponds to the description of the event's relationship. **Choose one of the following options**:

1. **Accept**: Accept as correct.
2. **Reject**: Reject as being far-fetched.
3. **No Judgment**: Unable to judge.

Figure 5: Screenshot of our human evaluation instruction for Emotion Knowledge Generation.