COKE: A Cognitive Knowledge Graph for Machine Theory of Mind

Jincenzi Wu[1,2*†] **Zhuang Chen**[1†] **Jiawen Deng**[3] **Sahand Sabour**[1]
Helen Meng[2] **Minlie Huang**[1‡]
[1]CoAI Group, DCST, IAI, BNRIST, Tsinghua University
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
[3]University of Electronic Science and Technology of China
jincenziwu@gmail.com zhchen-nlp@mail.tsinghua.edu.cn aihuang@tsinghua.edu.cn

Abstract

Theory of mind (ToM) refers to humans' ability to understand and infer the desires, beliefs, and intentions of others. The acquisition of ToM plays a key role in humans' social cognition and interpersonal relations. Though indispensable for social intelligence, ToM is still lacking for modern AI and NLP systems since they cannot access the human mental state and cognitive process beneath the training corpus. To empower AI systems with the ToM ability and narrow the gap between them and humans, in this paper, we propose COKE: the first cognitive knowledge graph for machine theory of mind, formalizing cognitive processes as a chained structure. Specifically, COKE formalizes ToM as a collection of 45k+ manually verified cognitive chains that characterize human mental activities and subsequent behavioral/affective responses when facing specific social circumstances. In addition, we further generalize COKE using LLMs and build a powerful generation model COLM tailored for cognitive reasoning. Experimental results in both automatic and human evaluation demonstrate the high quality of COKE, the superior ToM ability of COLM, and its potential to significantly enhance social applications. We release our code and data at https://github.com/jincenziwu/COKE.

1 Introduction

In social environments, human beings must be able not only to react to what others are doing, but also to anticipate what they will do. This ability to understand and infer human goals is typically described as Theory of Mind (ToM) (Premack and Woodruff, 1978). One way of accomplishing ToM is to observe what others do in various situations, and derive a set of affective and behavioral rules.

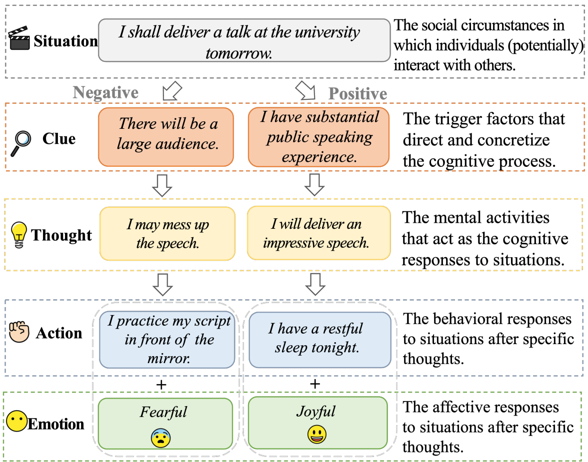

Figure 1: COKE instantiates Theory of Mind as positive and negative cognitive chains in social situations. *Situation⇒ Clue ⇒ Thought⇒ (Action + Emotion).*

When the same or highly similar things arise again, we can bring out plausible predictions accordingly (Call and Tomasello, 2011). Figure 1 presents an example that *someone will deliver a talk at the university tomorrow* (a social circumstance), and *he has substantial public speaking experience* (a trigger factor). We can plausibly anticipate that *he will deliver an impressive speech* (a mental activity), *feels joyful* (an affective response), and *has a restful sleep tonight* (a behavioral response). Here ToM is instantiated as a chained cognitive process that derives from our knowledge, experiences, and memories (Harris et al., 1989). ToM is indispensable to humans since it allows us to leverage our own minds to simulate others', so as to achieve efficient communication (Rabinowitz et al., 2018).

Despite its importance for social intelligence, ToM is not well internalized by modern AI and NLP systems. Shapira et al. (2023) illustrates a significant decline in performance and outright failure of Large Language Models (LLMs) in ToM tasks, particularly evident when confronted with adversarial samples. The main reason is that learning-based systems are usually trained on superficial text cor-

* Work done during internship at the CoAI Group.
† Equal Contribution.
‡ Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15984–16007
August 11-16, 2024 ©2024 Association for Computational Linguistics

pora, while lacking access to the underlying human mental state and cognitive process (Sap et al., 2022). In other words, NLP systems rely on the maximum likelihood to understand and generate texts, but do not go beneath the surface to the desires, beliefs, and intentions of humans.

In this paper, we introduce COKE: the first ***CO**gnitive **K**nowledg**E** graph* formalizing cognitive process as a chained structure for machine theory of mind. Our goal is to formalize ToM and make it accessible and learnable for AI systems. In COKE, we instantiate ToM as a collection of manually verified cognitive chains that characterize humans' mental activities in specific social circumstances along with their behavioral and affective responses (Meinhardt-Injac et al., 2018; Mehl et al., 2020). Each cognitive chain involves five types of nodes: **1)** ***situations*** denote the social circumstances; **2)** ***clues*** denote the trigger factors; **3)** ***thoughts*** denote the mental activities; **4)** ***actions*** denote the behavioral responses; **5)** ***emotions*** denote the affective responses. Moreover, as shown in Figure 1, individuals react differently to the same situation due to the diversified cognitive processes. Therefore, for each situation, we derive multiple cognitive chains and further label them as ***positive*** (means optimistic) or ***negative*** (means pessimistic) to mark the chain polarity. We propose to induce the raw data from LLMs, and then recruit educated workers majoring in psychology for manual selection and revision. The resulting knowledge graph constitutes 62,328 nodes and 45,369 cognitive chains.

The construction of COKE offers the basic ToM ability to understand and infer the human goals in already collected situations (Call and Tomasello, 2011). But obviously, it is impossible to enumerate all situations in the real world. Thus we move one step further and build a cognitive language model COLM to cope with unseen situations that have not appeared in the knowledge graph. Specifically, we decompose the construction of cognitive chains into four cognitive generation tasks, then finetune LLMs using the manually collected data in COKE. By this means, we combine the commonsense knowledge embedded in LLMs and the ToM ability provided by COKE, enabling COLM to infer cognitive chains for unseen situations.

We summarize our contributions in this work as follows. **1)** We propose the first cognitive knowledge graph for machine theory of mind, formalizing cognitive process as a chained structure. We instantiate human theory of mind as a collection of 45k+ manually verified cognitive chains, which provides a basic ToM ability for accessing and learning. **2)** We build a powerful cognitive language model COLM by associating COKE with LLaMA-2 (Touvron et al., 2023), so as to predict cognitive chains for out-of-KG situations. **3)** We conduct extensive experiments to evaluate the ToM ability of COLM and typical LLMs. The results show that COLM outperforms strong baseline models such as GPT-4 (Achiam et al., 2023) in both zero-shot and few-shot settings, proved by automatic and human evaluations in all cognitive generation tasks, which in turn demonstrates the high quality of COKE. **4)** We further substantiate the potential of COKE in enhancing social applications, and prove its effectiveness on downstream emotional support conversation task.

2 COKE: Cognitive Knowledge Graph

2.1 Preliminaries

What is Theory of Mind? Theory of Mind refers to the ability of humans to understand and infer other people's desires, beliefs, and intentions (Premack and Woodruff, 1978). Under specific social circumstances (***situations***), the core mechanism by which ToM empowers us is to ascribe others' mental activities (***thoughts***), and predict their corresponding behavioral responses (***actions***) and affective responses (***emotions***) (Leslie et al., 2004; Apperly, 2010). Furthermore, the acquisition of ToM enables us to realize and appreciate that people can have different cognitive responses in the same situation due to various trigger factors (***clues***) like personality and experience (Meinhardt-Injac et al., 2018).

Why Do AI Systems Need ToM? ToM has been a persistent yet elusive goal of artificial intelligence for decades (Choi, 2022). AI systems need ToM to understand a user's situations and predict subsequent reactions, so as to provide effective responses or operations that meet the user's needs (Le et al., 2019; Dhelim et al., 2021; Langley et al., 2022). However, recent studies (Shapira et al., 2023) show that today's LLMs still lack ToM and social intelligence. The main reason is that AI systems can only learn from text corpora in training, but cannot access the human mental state and cognitive process that determine what and why to say. Now, there are neither public resources that contain all the concepts in ToM, nor commonsense knowledge graphs that depict the structure of cognitive chains in ToM. This motivates us to propose the first cognitive knowledge graph for machine Theory of Mind.

2.2 Data Structure of COKE

According to the above-mentioned psychology research on theory of mind, we specify five types of nodes in COKE: ***situations***, ***clues***, ***thoughts***, ***actions***, and ***emotions***. We here define the basic unit of COKE as the following cognitive chain: *Situation⇒Clue⇒Thought⇒(Action+Emotion)*. This structure depicts the intact cognitive process of ToM: When a person faces a situation, some clues trigger his/her thoughts, along with his/her actions and emotions. Furthermore, to distinguish whether a cognitive chain is optimistic or pessimistic under the specific situation, we further define its polarity as ***positive*** or ***negative***. In practice, the polarity of the cognitive chain is determined by its thought node. Notice that, in COKE, we omit the definition of edges (i.e., the connections between nodes) since they can be easily inferred when the types of nodes are already known. We then illustrate the nodes in detail.

Situations in COKE denote the social circumstances in which individuals (potentially) interact with others. By referring to DailyDialog (Li et al., 2017), a widely used daily social dialogue dataset, we select the five most common social topics: *School* (what happened at school), *Work* (what happened at work), *Tourism* (travel and entertainment), *Relationship* (social activities between individuals), *Ordinary Life* (what happened in families). Detailed information about each topic can be found in Appendix A.

Clues in COKE denote the trigger factors that direct and concretize the cognitive process. In a specific situation, humans' mental activities are triggered and directed by relevant subjective and objective factors (Meinhardt-Injac et al., 2018). According to the taxonomy from (Baldwin, 1992), clues mainly involve the particular information about personality, knowledge, experience, education, objective facts, social rules, and so on.

Thoughts in COKE denote the mental activities that act as the cognitive responses to situations. Thoughts serve as the bridge between the external environment and individual cognition, thus can be considered as the core of ToM (Westbrook et al., 2011). As mentioned before, the polarity of a cognitive chain is anchored to its thought node. In other words, an optimistic thought marks the entire cognitive chain as positive, and a pessimistic thought marks it as negative.

Actions in COKE denote the behavioral responses to situations after specific thoughts. Notice that the semantic meaning of actions may not conform to their polarity annotations. For example, in Figure 1, the action "*I practice my script in front of the mirror*" is a neutral sentence. However, it is the consequence of the negative thought "*I may mess up the speech*", so it is still labeled as negative.

Emotions in COKE denote the affective responses to situations after specific thoughts. Without loss of generality, we restrict the emotions to six basic categories (Phillip et al., 1987): *Love*, *Surprise*, *Joyful*, *Sad*, *Angry*, and *Fearful*. The first three appear in positive cognitive chains, while the last three appear in negative cognitive chains.

2.3 Data Creation and Selection

Recent studies have proved that LLMs trained on huge text corpora can naturally serve as a repository for data collection (West et al., 2023; Kim et al., 2023). Inspired by their successful attempts, we propose a two-step data collection approach for constructing COKE. As shown in Figure 2, **1)** we first manually design suitable few-shot prompts to induce `GPT-3.5` (i.e., text-davinci-002) Ouyang et al. (2022) to automatically generate raw data for five types of nodes in a pipeline manner. **2)** We then recruit and train eight graduate students majoring in social psychology as annotators to select and revise the outputs of `GPT-3.5`. Next, we illustrate the details of how to prompt `GPT-3.5`, then introduce the data statistics after human annotation. More detailed parameters and templates used for prompting are provided in Appendix B and C.

Prompting LLMs for Situations Situations in COKE denote the social circumstances in which individuals (potentially) interact with others, which exhibit a strong correlation with social events and describe the social environment of daily lives. Unfortunately, to the best of our knowledge, there is no public dataset that contains a wide variety of social situations. Therefore, we choose the events in ATOMIC (Sap et al., 2019), a frequently-used commonsense knowledge graph for if-then reasoning, as an alternative. However, the events from ATOMIC are still not qualified to become the situations in COKE. The reasons are two-fold, and we here take the event "*PersonX misses calls from PersonY*" as an example. **1)** ATOMIC events replace the characters with "*PersonX*" and "*Per-*

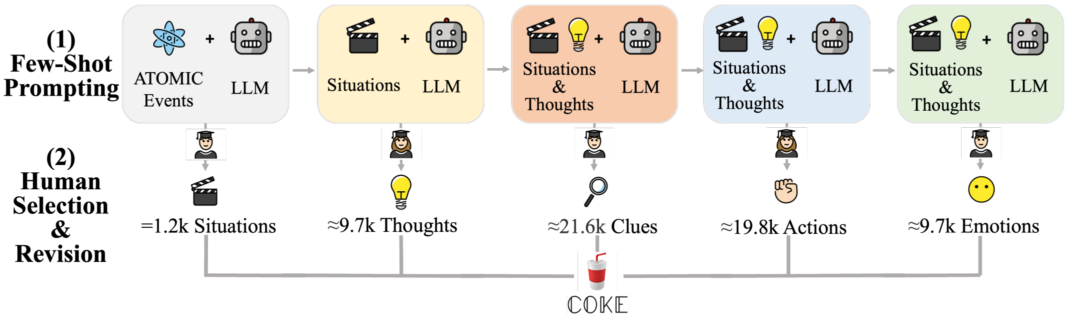

Figure 2: The two-step data collection approach for constructing COKE.

sonY", thereby losing most of the interpersonal information that is indispensable for social situations. For example, if someone misses calls from his employer, he may worry that something is wrong with his work. But if the caller changes to his girlfriend, he may sense love because she cares about him. **2)** ATOMIC events omit most of the social context information, making the background environments where the events occur not available. For example, we don't know if the calls in the above event happen at work or on vacation, so we have no idea how to reify subsequent cognitive processes.

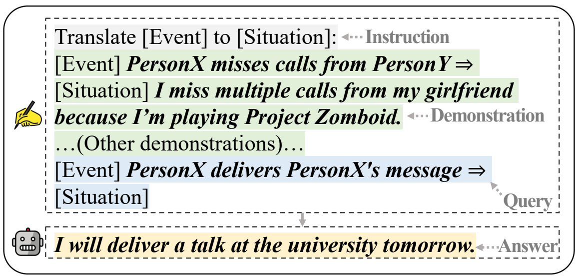

Figure 3: Prompting LLMs for situations.

To address the problem, we choose to prompt `GPT-3.5` to rewrite ATOMIC events to qualified COKE situations. Specifically, we first manually create several demonstration (input, output) pairs such as ("*PersonX misses calls from PersonY*", "*I miss multiple calls from my girlfriend because I'm playing Project Zomboid*"). Then we wrap the demonstrations with the task instruction ("*Translate [Event] to [Situation]*") and an arbitrary event query ("*PersonX delivers PersonX's message*") to form an input template as shown in Figure 3. As a result, we collect the output situation from `GPT-3.5` like "*I will deliver a talk at the university tomorrow*". In practice, we manually select 400 ATOMIC events that are general and easy to adapt, then rewrite each event to 5 situations with different topics. Finally, we obtain **2,000** raw situations.

Prompting LLMs for Thoughts Thoughts in COKE denote the mental activities that act as the cognitive responses to situations. As mentioned above, thoughts serve as the bridge between the external environment and individual cognition, and can be seen as the core of ToM. Since thoughts are triggered by certain clues, they come after clues in cognition chains. However, our pilot experiments show that directly using situations to prompt LLMs can better stimulate their generation ability and get more diverse thoughts, which is beneficial to manual selection and revision. Therefore, here we temporarily reverse the order of thoughts and clues, i.e., we first prompt LLMs using situations to generate thoughts, and then prompt LLMs using thoughts to generate clues.

Similar to the instruction prompts for collecting situations, we manually construct the demonstration as "*When [Situation], I feel great/terrible since I think [Thought]*" and again wrap it with the task instruction and an arbitrary situation query to form the input template. Notice that we use different sentiment terms ("*great*" and "*terrible*") to control the polarity (*positive* and *negative*) of the generated thoughts. Since the polarity of a cognitive chain is anchored to its thought, we have also controlled the polarity of the entire cognitive chain that will be generated. After prompting, we obtain **14,400** raw thoughts, half of which are positive and the other half are negative.

Prompting LLMs for Clues, Actions, and Emotions Clues in COKE denote the trigger factors that direct and concretize the cognitive process. Thus we construct the demonstrations as "*Complete the sentence: When [Situation], I think [Thought] since [Clue]*" to prompt LLMs for clues. Actions and emotions in COKE denote the behavioral and affective responses to situations after specific thoughts. We hence construct the demonstrations as "*When [Situation], I think [Thought] so [Action]*" and "*When [Situation], I think [Thought] and I feel [Emotion]*" to prompt LLMs for actions and emotions. Since emotion belongs to predefined categories, we further modify its corresponding prompt to the question-answering form. Finally, we obtain **29,364** raw clues, **29,364** raw actions, and **14,400** raw emotions. Their polarities are already determined by previously generated thoughts.

Human Selection and Revision After collecting the raw data, we manually annotate a small amount of data and formulate several rules to distinguish good data from bad data. Subsequently, we use the detailed definition of nodes in COKE and the filtering rules as a tutorial to train the eight annotators. After passing an annotation qualification test, they are asked to select and revise the raw data of five types of nodes. More details about human annotation can be found in Appendix D.

As shown in Table 1, the final reserved data contains **1,200** situations, **9,788** thoughts, **21,677** clues, **19,875** actions, and **9,788** emotions, resulting in an overall retention rate of around 70%. This statistic proves that ToM ability exhibited by most powerful LLMs like `GPT-3.5` is still not satisfactory enough even with delicate prompting, thus further emphasizing the necessity of our construction of COKE. After linking and ordering the obtained nodes, we instantiate ToM with a total of **45,369** cognitive chains in COKE, containing **23,252** positive chains and **22,117** negative chains in English. Example COKE data can be found in Appendix E.

Dimension	Raw	Final	Avg. Len.	Retention Rate
Situation	2,000	1,200	11.5	60.00%
Thought	14,400	9,788	6.6	67.97%
Clue	29,364	21,677	7.3	73.82%
Action	29,364	19,875	6.8	67.68%
Emotion	14,400	9,788	1.0	67.97%

Table 1: The statistics of COKE.

3 Cognitive Language Model COLM

By consulting the constructed cognitive knowledge graph COKE, we can obtain the basic ability of theory of mind (ToM) via matching the faced situation to a similar situation in KG, and then inspecting the involved cognitive chains (a.k.a, entity linking). But obviously, COKE only collects finite situations and cannot cover the infinite and diverse situations in the real world. Inspired by the methods for automatic knowledge graph completion like COMET (Bosselut et al., 2019), we propose a cognitive language model COLM to cope with unseen situations and expand the scope of application of COKE. COLM is built upon LLMs, aiming to integrate the commonsense knowledge inside LLMs and the ToM ability from COKE. To this end, we first decompose the cognitive process into a sequence of cognitive generation tasks, and then finetune LLaMA-2 (Touvron et al., 2023) using the collected data from COKE.

3.1 Cognitive Generation Tasks

In COKE, a cognitive process towards theory of mind (ToM) is instantiated as a sequenced cognitive chain containing situation, clue, thought, action, and emotion. Therefore, given a situation, we can decompose the cognitive process into four generation tasks as shown in Table 2. These four tasks work in a pipeline manner, and the complete cognitive chain can be restored by linking their generated results. 1) **Clue Generation**. When facing a specific situation, humans can automatically distinguish the factors that may influence beliefs and trigger thoughts. 2) **Thought Generation**. In a specific situation, the related clues trigger and arouse diversified human mental activities, i.e., thoughts. 3) **Action Generation**. Driven by specific thoughts, we humans will take corresponding actions to realize our beliefs and achieve our goals. Notice that we omit clues here since their impacts are largely covered by the triggered thoughts. 4) **Emotion Generation**. After forming a specific thought in a certain situation, humans will naturally generate corresponding emotions to express attitudes and views on the situation. Since emotions are limited to 6 categories, this task is a classification task.

By linking the above four tasks in a pipeline manner, we can restore the cognitive chain "*Situation $\Rightarrow$ Clue $\Rightarrow$ Thought $\Rightarrow$ (Action+Emotion)*", thus preserving the complete cognition process of ToM. Since each cognitive chain in COKE is labeled with polarity, each cognitive generation task can be further divided into positive and negative subtasks (e.g., positive thought generation and negative thought generation).

3.2 Training COLM

After decomposing the cognitive chain into four cognitive generation tasks, we can process the data in COKE accordingly to obtain training samples. For computational efficiency, we design COLM as a multi-task controllable generation model, so that it can simultaneously accomplish four cognitive generative tasks and further control the polarity (i.e., positive or negative) of the outputs. As shown

Task	Generation Setting	Input X	Output Y
Clue Generation	Situation⇒ Clue	Situation [NegClue]	Clue
Thought Generation	Situation + Clue ⇒ Thought	Situation [NegClue] Clue [NegThought]	Thought
Action Generation	Situation + Thought ⇒ Action	Situation [NegThought] Thought [NegAction]	Action
Emotion Generation	Situation + Thought ⇒ Emotion	Situation [NegThought] Thought [NegEmotion]	Emotion

Table 2: Decomposition of the cognitive process to four cognitive generation tasks for training COLM. The complete outputs of four tasks can be restored to cognitive chains. Here [Neg.*] denotes special tokens for controllable generation in ***negative*** cognitive chains. When coming to ***positive*** cognitive chains, we use [Pos.*].

in Table 2, for each task and polarity, we insert specific tokens (i.e.,[NegClue],[PosClue]) in the input X to guide the generation process of COLM. For implementation, COLM is built as a decoder-only architecture and initialized with the LLaMA-2 (Touvron et al., 2023). The model is trained with each input-output pair X-Y from any task.

4 Experiments

In this section, we construct a dataset from COKE to evaluate ToM ability of our cognitive language model COLM, and compare it with advanced LLMs including GPT-3.5 Turbo, GPT-4, LLaMA-2-7B (our backbone) and Mistral-7B. We first illustrate the experimental setup, then analyze the experimental results for automatic and human evaluation, and finally validate COKE's effectiveness to empower social applications.

4.1 Experimental Setup

To evaluate the ToM ability of different models, we randomly split 1,200 social situations into 1,080 (90%) training and 120 (10%) validation situations. Then we can automatically split samples for different cognitive generation tasks according to the situations, and obtain training/validation splits as 19,409/2,268 (clue), 8,746/1,042 (thought), 17,982/1,893 (action), and 8,746/1,042 (emotion). Based on this setting, we ensure that the cognitive chains in the validation set all occur in **UNSEEN** situations, which is crucial for ToM evaluation.

For COLM, we use the Hugging Face implementation[1] of LLaMA-2, and train it for 20 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) and LoRA (Hu et al., 2021) with learning rate 3e-4 and batch size 32 in Five Tesla V100 GPUs. The whole training process costs about 9 GPU hours. For baseline models, we construct manual prompts to enable them to complete the cognitive generation tasks. We evaluate 0-shot, 2-shot, and 5-shot performance of all baseline LLMs. More details can be found in Appendix F. Due to the data deficiency, we report the performance on the validation set.

4.2 Main Results

Automatic Evaluation To evaluate the clue/thought/action generation tasks, we use METEOR (Lavie and Denkowski, 2009), ROUGE (Lin, 2004), BLEU-1, BLEU-2 (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) as metrics. In these three tasks, each input may be mapped to multiple ground truth outputs. Therefore, following Mostafazadeh et al. (2020), we compute the average scores between all predicted sequences and all ground truth outputs[2]. Moreover, to evaluate the emotion generation (classification) task, we compute the classification accuracy as the metric. We present all automatic evaluation results in Table 3.

Compared to baseline models like LLaMA-2 and Mistral, which face challenges in following task instructions, COLM shows significant performance enhancements across various cognitive generation tasks with COKE. While leveraging additional prompts boosts performance for powerful models like GPT-3.5 Turbo and GPT-4 via in-context learning in clue/thought/action generation tasks, these LLMs still struggle with complex emotional understanding (Wang et al., 2023) and fail in the emotion classification task. Compared to these powerful LLMs, COLM maintains substantial advantages across all evaluation metrics. We hereby have two observations: 1) The data we collected in COKE is of high quality and can empower the model with strong ToM ability. 2) COLM, which is designed as a controllable generative model for multiple cognitive tasks, can effectively internalize the ToM ability and cope with unseen situations.

Human Evaluation Beyond automatic evaluation, we also wonder how humans perceive the cognitive chains generated by different models since ToM is essentially a human ability. Therefore, for each model, we sample a cognitive chain for each

[1]https://huggingface.co/meta-llama/Llama-2-7b-hf.

[2]Here we provide all models with the required polarity (positive or negative) via special tokens or prompts.

Task	Model	GPT-3.5 Turbo			GPT-4			LLaMa-7B			Mistral-7B			COLM
		0-shot	2-shot	5-shot	0-shot	2-shot	5-shot	0-shot	2-shot	5-shot	0-shot	2-shot	5-shot	
Clue Gen.	METEOR↑	0.215	0.268	0.293	0.211	0.286	0.292	0.251	0.247	0.246	0.233	0.238	0.224	**0.370**
	ROUGE↑	0.222	0.245	0.266	0.190	0.260	0.270	0.183	0.181	0.177	0.172	0.178	0.184	**0.381**
	BLUE-1↑	0.200	0.227	0.259	0.183	0.249	0.269	0.115	0.117	0.119	0.117	0.116	0.117	**0.340**
	BLEU-2↑	0.007	0.007	0.008	0.005	0.008	0.009	0.003	0.003	0.003	0.003	0.003	0.004	**0.011**
	BertScore↑	0.862	0.868	0.872	0.859	0.871	0.872	0.840	0.842	0.844	0.829	0.832	0.841	**0.876**
Thought Gen.	METEOR↑	0.166	0.266	0.274	0.131	0.244	0.250	0.144	0.200	0.207	0.142	0.184	0.190	**0.305**
	ROUGE↑	0.132	0.205	0.215	0.109	0.198	0.203	0.094	0.132	0.141	0.152	0.203	0.144	**0.344**
	BLUE-1↑	0.195	0.259	0.313	0.196	0.321	0.326	0.121	0.115	0.112	0.117	0.164	0.160	**0.371**
	BLEU-2↑	0.017	0.024	0.029	0.018	0.030	0.031	0.011	0.010	0.009	0.010	0.015	0.015	**0.037**
	BertScore↑	0.874	0.890	0.896	0.871	0.894	0.896	0.835	0.847	0.855	0.833	0.840	0.854	**0.902**
Action Gen.	METEOR↑	0.206	0.253	0.291	0.209	0.248	0.247	0.134	0.207	0.217	0.168	0.202	0.210	**0.342**
	ROUGE↑	0.186	0.255	0.284	0.193	0.243	0.257	0.104	0.172	0.178	0.133	0.171	0.182	**0.378**
	BLUE-1↑	0.153	0.271	0.302	0.168	0.223	0.257	0.121	0.130	0.149	0.121	0.122	0.130	**0.378**
	BLEU-2↑	0.010	0.018	0.020	0.011	0.014	0.017	0.008	0.008	0.009	0.007	0.008	0.008	**0.027**
	BertScore↑	0.872	0.882	0.886	0.866	0.879	0.882	0.821	0.851	0.842	0.832	0.841	0.844	**0.892**
Emotion Class.	Accuracy↑	0.693	0.437	0.474	0.634	0.613	0.649	Failed	0.068	0.062	Failed	Failed	Failed	**0.793**

Table 3: Automatic evaluation results for cognitive generation tasks. The highest scores are highlighted in color, and second best results are underlined. All results are average scores of 3 runs with random seeds.

situation in the validation set (resulting in 90 samples per model) for human evaluation. We present three experts with the cognitive chains generated by COLM and top-performing LLMs in automatic evaluation, and then ask them to score on a 5-point Likert scale based on *relevance*, *correctness*, and *logicality*. Specifically, they were instructed to rate the chains as follows: "completely irrelevant (1 point)", "contains irrelevant content (2 points)", "relevant but does not follow the task definition correctly (3 points)", "follows the task definition correctly but does not logically continue the cognitive chain (4 points)" and "continues the cognitive chain correctly and logically (5 points)". We calculate the final score for each model by averaging these ratings. A rating of 5 indicates **valid** content.

The results in Table 4 illustrate that the cognitive chains generated by COLM are more acceptable to humans, which further demonstrates the superiority of its ToM ability. Even with clear task instructions and demonstrations, baseline LLMs failed to generate inferences acceptable to humans. GPT-3.5-Turbo received a significant number of 3-point ratings, indicating responses that were "relevant but does not follow the task definition", such as generating thought in clue task. Additionally, there were a few 2-point ratings, indicating "contains irrelevant content", such as generating content unrelated to ToM definition. GPT-4 received a small number of 3-point and 4-point ratings. COLM received only a small number of 4-point ratings. As a further illustration, we present ten exemplary cognitive chains generated by COLM across five diverse situations, covering various topics, in Appendix G.

Task	Evaluation	GPT-3.5 Turbo	GPT-4	COLM
Clue Gen.	Overall	4.22	4.72	4.83
	Valid.	61%	85%	90%
Thought Gen.	Overall	4.84	4.86	4.87
	Valid.	82%	88%	90%
Action Gen.	Overall	4.70	4.79	4.82
	Valid.	70%	78%	85%

Table 4: Human evaluation results of COLM and baseline LLMs with 5 shots. "**Overall**" determines if the response is relevant, correct, and logical to continue the cognitive chains. "**Valid.**" refers to the validation percentage of outputs.

4.3 Empowering Social Application

As COKE is a cognitive knowledge graph for machine theory of mind, we further substantiate its effectiveness in empowering social applications with ToM capabilities. Our chosen testing ground is the emotional support conversation (ESC) task (Liu et al., 2021). The prime objective of ESC is to generate empathetic and effective responses in social dialogues, with the aim of mitigating users' emotional distress and fostering improved mental states. ESC requires ToM because it involves understanding and responding to a user's emotional state and underlying challenges by understanding their mental processes and intentions.

Integration of COKE in ESC The ESC dataset has already provided the situation behind each dialogue. We then treat each user's utterance in the dialogue as a clue, and employ COLM to infer two thoughts (1 positive and 1 negative) via thought generation. Based on the generated thoughts, we can accordingly infer actions for each thought via action generation. We simply extract verbs and nouns in thoughts or actions as the ToM knowledge keywords, and append them to the end of dialogue history, serving as an enhanced context. Based on the context, the ESC system is trained to generate the corresponding response. We here use the Blenderbot (Roller et al., 2021) as the generation model, which is the same as the original ESC paper.

ESC Performance with ToM Knowledge In our experiments, we compare three models: Vanilla ESC, ESC with COLM Actions, and ESC with COLM Thoughts. We conduct both automatic and human evaluations, following the same methodology as the original ESC paper. We sample 30 dialogues in the validation set and let two experts rate *Fluency*, *Identification*, *Comfort*, and *Suggestions* on a 3-point Likert scale. The final score is the average of these ratings. The experimental results shown in Table 5 demonstrate that, when incorporating the ToM knowledge into the response generation process, both the ESC with COLM Actions and ESC with COLM Thoughts models significantly outperform the Vanilla ESC model across all metrics. For clarity, we further present a case study and how COKE offers a more nuanced and empathetic approach to AI-driven emotional support in Appendix H .

Metric	Vanilla	+ Action	+ Thought
PPL ↓	16.96	15.96	15.88
BLEU-2 ↑	6.93	7.18	7.34
ROUGE-L ↑	15.01	15.73	15.89
Extrema ↑	50.28	50.22	50.43
Fluency ↑	2.30	2.45	2.60
Identification↑	1.95	1.95	2.30
Comforting ↑	2.10	2.35	2.50
Suggestion ↑	1.35	1.40	2.05

Table 5: Automatic and human evaluation on ESC.

5 Related Work

Cognitive knowledge is essential for modeling intuitive tasks, such as reasoning, in language models. Previous work mainly focuses on realizing cognitive knowledge as an accessory to commonsense knowledge. They construct text-based knowledge graphs that include numerous instances of Head-Relation-Tail triplets regarding different events and objects. One of the widely-used examples of such knowledge graphs is ConceptNet (Speer et al., 2016), which provides a graph of concepts connected by relations, covering a variety of taxonomic facts. Sap et al. (2019) proposed ATOMIC, a graph of if-then inferences that models social commonsense in daily life events. Hwang et al. (2021) expanded upon ATOMIC by incorporating two additional categories of commonsense relations: physical-entity commonsense relations and event-centered commonsense relations. West et al. (2021) prompted GPT-3 to distill commonsense knowledge and create a machine-generated corpus ATOMIC^{10x}. Kim et al. (2023) further leveraged it to generate a large-scale dataset focused on socially-grounded conversations. Furthermore, NOVATOMIC (West et al., 2023) employs natural language queries as open relations to link the commonsense knowledge, enabling its application in general reasoning tasks.

Generally, existing KGs consider event-centered commonsense relations, such as temporal and causal relationships, represented as $Event \xleftarrow{\text{IsBefore}} Event \xrightarrow{\text{Causes}} Event$. However, they have not explicitly addressed ToM concepts and relations, which are crucial for accessing and interpreting human mental states and cognitive processes. In contrast, COKE delineates ToM concepts and structures them as a cognitive chain: *Situation⇒ Clue⇒ Thought⇒(Action+ Emotion)*. This chained structure is designed to mirror human cognitive processes, enabling a deeper understanding of how individuals infer others' mental states in specific social circumstances along with their behavioral and affective responses. Moreover, the ATOMIC family, which uses short phrases with generic placeholders like "PersonX", significantly constrains the social and interpersonal information critical for complex ToM reasoning (Zalla and Korman, 2018). COKE addresses this limitation by providing a richer and more nuanced context for each cognitive concept in the chain. This approach broadens the scope of cognitive knowledge accessible to machine systems and enhances their capability in social cognition, hereby fostering downstream applications.

6 Conclusion

In this work, we present COKE, the first cognitive knowledge graph for machine theory of mind. We instantiate ToM as a series of cognitive chains to describe human mental activities and behavioral/affective responses in social situations. Through prompting GPT-3.5 and manual annotation, we collect 62k+ nodes and construct 45k+ cognitive chains. Based on COKE, we build a powerful cognitive language model COLM. COLM can handle unseen situations and predict complete cognitive chains in a pipeline manner. Automatic and human evaluations show that COLM effectively internalizes the ToM ability and outperforms strong baseline models like GPT-4.

We further demonstrate that COKE can empower LLMs with a nuanced understanding of human affective and cognitive reasoning, resulting in superior performance in ESC. By integrating COKE into LLMs, they gain the ability to comprehend and predict human cognitive processes with greater depth. We believe that COKE and COLM can facilitate social applications such as dialogue systems and autonomous agents.

Acknowledgments

This work was supported by the National Science Foundation for Distinguished Young Scholars (No. 62125604) and the NSFC Key Project (No. 61936010). This work was also supported by Tsinghua Precision Medicine Foundation, Tsinghua University - Beijing Tsingshang Architectural Decoration Engineering Co., Ltd. Joint Institute for Smart Scene Innovation Design, and China Postdoctoral Science Foundation (No. 2023M741944).

Limitations

In this work, we introduce COKE, the first cognitive knowledge graph for machine Theory of Mind, which aims to empower AI systems with cognitive capabilities. However, we acknowledge the following omissions and inadequacies in our work.

Data Coverage We acknowledge that there is a limitation in topic coverage of proposed COKE, which only covers five social topics that are commonly discussed. Besides, COKE has a relatively small data scale in comparison with other relative knowledge graphs, due to its more specific node contents and more complicated construction process. Consequently, COKE cannot cover all situations in deployment, and the cognitive reasoning models constructed on this basis may have unreliable predictions in out-of-domain situations.

Cognitive Inference Ability As a pioneer in integrating *Theory of Mind* into AI systems, the proposed cognitive language model COLM still has a lot of room for improvement in inferring cognitive chains when deployed in practice. In COLM, we take LLaMA-2 as the backbone model to validate the gain of COKE on the cognitive inference ability of language models. We acknowledge that adopting larger backbone models would contribute to a more powerful inference model, which is worth exploring in more depth in the future.

Ethics Statement

We present the first cognitive knowledge graph COKE for machine theory of mind. Our graph is built based on public datasets and model generations. We strictly adhere to data source usage protocols and ensure that the proposed COKE can be released and used LEGALLY.

The construction of COKE has gone through the steps of manual selection and revision. We manually filtered the data containing potentially private information, such as phone numbers and email addresses, to protect user PRIVACY further. We also carefully delete abusive, offensive, biased, and other inappropriate content to avoid unpredictable ETHICAL hazards.

Our knowledge graph is designed to empower AI systems with the ability of cognitive inference. We are aware that this capability could be misused in malicious scenarios in the future. However, we believe the value of this work outweighs the risks, and we also call for more socially responsible research in this field.

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*.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*.

Mark W Baldwin. 1992. Relational schemas and the processing of social information. *Psychological bulletin*, page 461.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779.

Josep Call and Michael Tomasello. 2011. Does the chimpanzee have a theory of mind? 30 years later. *Human Nature and Self Design*, pages 83–96.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, pages 139–155.

Sahraoui Dhelim, Huansheng Ning, Fadi Farha, Liming Luke Chen, Luigi Atzori, and Mahmoud Daneshmand. 2021. Iot-enabled social relationships meet artificial social intelligence. *IEEE Internet of Things Journal*, pages 17817–17828.

Paul L Harris, Carl N Johnson, Deborah Hutton, Giles Andrews, and Tim Cooke. 1989. Young children's theory of mind and emotion. *Cognition & Emotion*, pages 379–400.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. On symbolic and neural commonsense knowledge graphs.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *EMNLP*.

Christelle Langley, Bogdan-Ionut Cirstea, Fabio Cuzzolin, and Barbara Jacquelyn Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Mach. Transl.*, 23(2-3):105–115.

Matt Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *EMNLP*.

Alan M Leslie, Ori Friedman, and Tim P German. 2004. Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, pages 528–533.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, pages 986–995.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *ACL/IJCNLP 2021*, pages 3469–3483.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*. OpenReview.net.

Stephanie Mehl, Klaus Hesse, Anna-Christine Schmidt, Martin W Landsberg, Daniel Soll, Andreas Bechdolf, Jutta Herrlich, Tilo Kircher, Stefan Klingberg, Bernhard W Müller, et al. 2020. Theory of mind, emotion recognition, delusions and the quality of the therapeutic relationship in patients with psychosis–a secondary analysis of a randomized-controlled therapy trial. *BMC psychiatry*, pages 1–13.

Bozana Meinhardt-Injac, Moritz M Daum, Günter Meinhardt, and Malte Persike. 2018. The two-systems account of theory of mind: Testing the links to social-perceptual and cognitive abilities. *Frontiers in human neuroscience*, page 25.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *EMNLP*, pages 4569–4586.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Shaver Phillip, Schwartz Judith, Kirson Donald, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, page 1061.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, pages 515–526.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew M. Botvinick. 2018. Machine theory of mind. In *ICML*, pages 4215–4224.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *EACL 2021*, pages 300–325.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *ArXiv*.

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 *AAAI*, pages 3027–3035.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*.

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*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, et al. 2023. Novacomet: Open commonsense foundation models with symbolic knowledge distillation. *arXiv preprint arXiv:2312.05979*.

David Westbrook, Helen Kennerley, and Joan Kirk. 2011. *An introduction to cognitive behaviour therapy: Skills and applications*.

Tiziana Zalla and Joanna Korman. 2018. Prior knowledge, episodic control and theory of mind in autism: Toward an integrative account of social cognition. *Frontiers in psychology*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Topic Keywords

We identify the top 8 keywords in five topics using TF-IDF. As indicated in Table 6, the information provided by situations in each topic is distinct and relevant to the theme. According to the keywords "*boss*" or "*work*", we can locate information such as "*I attend my boss's meeting*" in the Work topic. The project deadline dilemma like "*I bet my friend that I can finish my project before the deadline*" is shared in the School topic. Besides, we can find the travel schedule like "*I asked my best friend for advice on what to pack for my upcoming trip to Europe*" in the Tourism topic. Date sharing like "*I brought my date back to my place after the movie*" can be found in the Relationship topic. Daily diary like "*My friend is helping me plan my surprise party*" appears in the Ordinary Life topic.

Topic	**Keywords extracted**
School	asked, friend, friends, dad, project, mom, help, go
Work	boss, ask, work, asked, new, friend, job, meeting
Tourism	asked, trip, travel, go, friend, friends, sister, brother
Relationship	asked, new, date, friends, best, girl,guy, girlfriend
Ordinary Life	asked,day, friend, new, friends, go, wanted, going

Table 6: Keywords in five topics.

B Parameters for GPT-3.5 API Utilization

During the data collection process, we use the GPT-3.5 API offered by OpenAI. We read the terms of service[3] and follow the usage policies[4] . We give the parameter details of the GPT-3.5 API utilized in data collection and the experiments in Table 7. Our data collection was finalized before the release of GPT-3.5 Turbo and GPT-4. Therefore, we opted for the powerful model *text-davinci-002*.

C Prompts for Data Collection

In Table 11, we present the detailed prompts for how to extend the base event to the situation in our COKE. When we design the prompt template, we find that the base events prefixed with the token **[Sentence]** in the prompt lead to better generation than those with the token **[Event]**. Table 12 and Table 13 show the specific prompts used to collect data in the negative and positive chains for each generation task.

[3] https://openai.com/api/policies/service-terms/.
[4] https://beta.openai.com/docs/usage-policies.

Parameter	Situation	Clue	Thought	Action	Emotion
n	1	3	3	3	1
best_of	1	3	3	3	1
model	text-davinci-002	text-davinci-002	text-davinci-002	text-davinci-002	text-davinci-002
temperature	1	1	1	1	1
max_tokens	256	256	256	256	256
top_p	1	1	1	1	1
frequency_penalty	0	0	0	0	0
presence_penalty	0	0	0	0	0

Table 7: Parameters for GPT-3.5 generation.

D Instruction for Human Selection and Revision

Since our cognitive knowledge graph is closely related to psychology, we choose and train eight graduate students majoring in social psychology as annotators. They are evenly distributed by gender (four males and four females) and come from various regions. We pay the annotators approximately $12 per hour, which exceeds the local minimum wage, and the total cost for annotation is about $6,000. We illustrate the relevant background and how the data would be used clearly in the first beginning.

To construct the knowledge graph for machine Theory of Mind (ToM), we have researched relevant papers and books and discussed it with the social psychology professor several times over three months. As a result, we decompose the ToM inference into four inference tasks and clearly define the five types of nodes. In the annotation instructions, explicit definitions of nodes are provided.

In order to effectively train the annotators, we first have a testing annotation on 160 situations for each task ourselves. Afterwards, we determine what types of incorrect data annotators may encounter and utilize these annotation examples to provide specific guidance. Several common types of bad data in all tasks are 1) repetitive context, 2) unsafe words, and 3) offensive content. Therefore, we automatically filter out the repetitive context and let the annotator manually remove the unsafe terms and offensive content. Besides, the specific types of incorrect data for each task are depicted in the details instructions in Figure 5, Figure 6, Figure 7, and Figure 8. In order to ensure data diversity, we ask workers to modify repetitive data into different and reasonable data.

During training, the annotators work on ten testing situations and receive feedback after following the instructions. Then, based on their annotations of the second 10 situation examples, we evaluate if students have a solid grasp of the task. With specific guidance, students all do well in the subsequent annotation. In addition, to maintain the high quality of the data, we allow them to highlight data with which they are confused and work as the experts to make the ultimate decision. For example, in emotion data revision, the annotator assigns additional emotion labels to the illogical inference, and the experts make the final decision.

E Example Cognitive Chains in COKE

We present the cognitive chains in COKE from five topics in Figure 9 (School), Figure 10 (Work), Figure 11 (Tourism), Figure 12 (Relationship), and Figure 13 (Ordinary Life), respectively. The data in COKE is in English.

F Prompts for PLMs Generation in Evaluation

As shown in Table 8, we present the prompts that lead the LLaMA-2, Mistral, GPT-3.5 Turbo and GPT-4 to make inferences on the validation dataset.

G Case Study of Generated Cognitive Chains

To have a close look, in Table 9, we present ten cognitive chains generated by COLM in five situations covering different topics. It can be observed that the proposed cognitive generation model COLM can generate smooth and effective cognitive chains in unseen situations.

H Empowering Emotional Support Conversation with COKE

In this section, we discuss how COKE can be used to enhance social applications, specifically focusing on Emotional Support Conversation (ESC) (Liu et al., 2021). This task represents a key arena in which Theory of Mind (ToM) can play a pivotal role, given the need to understand and respond to a user's emotional state and underlying challenges.

Emotional Support Conversation The ESC task is structured around a user in a negative emotional state, potentially due to a specific problem or challenge they are confronting. The user's emotional state is characterized by a negative emotion label

Type	Prompt for testing clue collection
Clue generation for the negative cognitive chain	L1 Complete the sentence with the negative clue: L2 When Situation, I think negatively since
Clue generation for the positive cognitive chain	L1 Complete the sentence with the positive clue: L2 When Situation, I think positively since

Type	Prompt for testing though collection
Thought generation for the negative cognitive chain	L1 Complete the sentence with the negative thought: L2 When Situation and Clue, I feel terrible since I think
Thought generation for the positive cognitive chain	L1 Complete the sentence with the positive thought: L2 When Situation and Clue, I feel great since I think

Type	Prompt for testing action collection
Action generation for the negative cognitive chain	L1 Complete the sentence with the negative action: L2 When Situation, I think Thought, so
Action generation for the positive cognitive chain	L1 Complete the sentence with the positive action: L2 When Situation, I think Thought, so

Type	Prompt for testing emotion collection
Emotion generation for the negative cognitive chain	L1 Choose one word from Sad, Angry, Fearful to describe the given situation: L2 When Situation, I think Thought.
Emotion generation for the positive cognitive chain	L1 Choose one word from Love, Surprise, Joyful to describe the given situation: L2 When Situation, I think Thought.

Table 8: Prompts for LLaMa/Mistral/GPT-3.5-Turbo/GPT-4 generation on the validation data in the negative and positive cognitive chains.

(e), an intensity level of the emotion (l, on a scale of 1 to 5), and the underlying challenge causing their distress. The goal of the supporter (or AI system) is to comfort the user in a conversation, deploying support skills to reduce the intensity of the user's negative emotions. The conversation's effectiveness is gauged by the extent to which the user's emotional intensity is reduced, and the ability of the supporter to accurately identify the problem, comfort the user, and offer constructive solutions or suggestions.

Experimental Results In experiments, we compare three models: Vanilla ESC, ESC w/ COLM Actions, and ESC w/ COLM Thoughts. We conduct both automatic evaluation (PPL, BLEU-2, ROUGE-L, Extrema) and human evaluation (Fluency, Indentification, Comforting, Suggestion), same as the original ESC paper. The experimental results are presented in Table 5. To have a close look, we also present a case study in Table 10. The results demonstrate the potential of COKE to empower social applications, offering a more nuanced and empathetic approach to AI-driven emotional support. By integrating COKE's cognitive chains into dialogue history, AI systems can achieve a more sophisticated understanding of human emotions and behavior, leading to more effective interactions in emotionally charged contexts such as the ESC task. Interestingly, the ESC with COLM Thoughts model shows a slight advantage over the ESC with COLM Actions model. This could be attributed to the fact that the thoughts generated by COLM offer a more direct reflection of the user's mental state.

Topic	Situation	Clue	Thought	Action	Emotion
Work	I accidentally dropped my companies latest phone prototype.	Neg. It was worth more than my car.	The phone is going to be ruined.	I quickly sent an email to my boss to own up to the mistake.	Fearful
		Pos. I have been working so hard on this project.	My boss will forgive me.	I quickly picked it up and acted like it was no big deal.	Joyful
Relationship	I also set up a romantic evening for the two of us.	Neg. He has been acting distant recently.	He may not show up.	I asked for some friends as backup.	Sad
		Pos. We have been getting along very well lately.	He is going to propose to me.	I am going to take him to his favorite restaurant.	Love
School	I am going to adopt a new approach to my studies.	Neg. I have never done things differently before.	I will not be able to understand the material.	I read through the material again.	Fearful
		Pos. I have put in a lot of effort.	It will improve my grades.	I take my study time more seriously.	Joyful
Tourism	I will accompany my brother to Greece for the summer.	Neg. I do not pack enough money.	I will regret not spending my time more wisely.	I asked him to take me to as many places as he could.	Sad
		Pos. We explored different cities and towns.	I will have a great time there.	I pack my bags with the best clothes I have.	Joyful
Ordinary Life	My dad acts as my power of attorney .	Neg. He is not my real father.	He will spend all my money.	I find a more reliable person.	Fearful
		Pos. My dad is really responsible and always helps me out.	He will make good decisions on my behalf.	I follow his instructions.	Joyful

Table 9: Case study. Given situations from different topics, the proposed cognitive generation model COLM can predict corresponding clues, thoughts, actions, and emotions in a pipeline manner.

Dialogue History	SITUATION: I can't seem to make any friends - always left on my own and no-one to talk to. SYSTEM: Hello how are you doing today? I'd love to chat with you! What's been on your mind? USER: oh hello I'm having problems with friends - or more precisely the lack of them. I don't have any. SYSTEM: Oh I see. You feel as though you don't have any friend's currently? I totally get that. I feel like that a lot too, especially in these weird times we are going through. It's so hard to see people and keep relationships right now USER: Well it's more that I think I've never had friends. Not anyone I can confide in or laugh with SYSTEM: That's a really hard thing to be going through. I'm sorry to hear that. Do you have any really close family members that you are able to confide in? USER: I don't think I'd want to let them know how unhappy I am to be honest SYSTEM: Oh I see. Well personally, when I was going through a hard time and I felt like I didn't have any friends, I found a couple pretty cool apps that helped me out! IT sounds crazy, but there are apps specifically for finding friends. They are kind of like dating apps but it's 100% meant for friends so everyone on them has the same intentions you do. I actually found a couple close friends through apps like that. Have you heard about those or looked into them? USER: I've never heard of friendship apps to be honest
Keywords	viewed,negative,person,going,confront,something, alone,need,make,friends
Golden Response	SYSTEM: Yeah! They are a really good way to connect with people that you share a lot of similar interests with.
Vanilla ESC	SYSTEM: it's a very popular app and there are tons of sites like facetime and facetime.
ESC w/ COLM Thoughts	SYSTEM: well, they are apps that provide a way to meet people from different cultures.

Table 10: Case study of emotional support conversation.

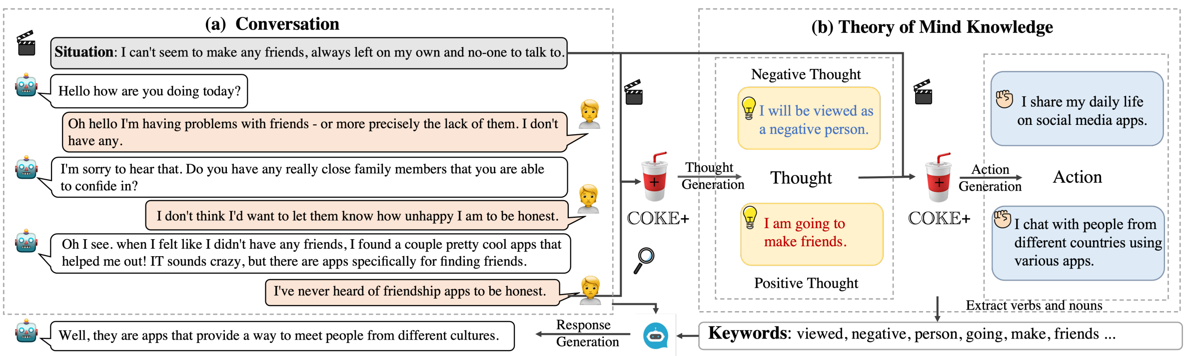

Figure 4: Empowering ESC with COLM. A detailed case study is illustrated in Table 10.

Topic		Prompt inputs for collecting *Situations*
(a) School	L1	Translate [Sentence] to [Situation]:
	L2	[Sentence] PersonX gives a presentation. => [Situation] I will give a presentation at college tomorrow .
	L3	[Sentence] PersonX has PersonX's driving test. => [Situation] My mom has her driving test tomorrow.
	L4	[Sentence] PersonX gives PersonY a book. => [Situation] My professor give me a book for research.
	L5	[Sentence] PersonX pays PersonY's fees. => [Situation] My aunt paid my test fees.
	L6	[Sentence] Selected Atomic Events => [Situation]
(b) Work	L1	Translate [Sentence] to [Situation]:
	L2	[Sentence] PersonX has a meeting . => [Situation]I will have a meeting with my boss .
	L3	[Sentence] PersonX took PersonY . => [Situation] I went out to eat with my boss, and he took me to a really nice restaurant.
	L4	[Sentence] PersonX announces PersonX's decision . => [Situation] I announce my decision to accept a job position that I did not want in order to please my father .
	L5	[Sentence] PersonX becomes PersonY's . => [Situation] My friend becomes my boss.
	L6	[Sentence] Selected Atomic Events => [Sentence]
(c) Ordinary Life	L1	Translate [Sentence] to [Situation]:
	L2	[Sentence] PersonX watches the game . => [Situation] I watched Wimbeldon for the first time .
	L3	[Sentence] PersonX gives PersonY advice . => [Situation] My friend gives me financial advice on stocks .
	L4	[Sentence] PersonX visits PersonX's friends . => [Situation] I went to Cuba to visit some family members and met up with some friends to go to a bar .
	L5	[Sentence] PersonX is cleaning PersonY's house . => [Situation] I was helping clean out my parents' house the other day and found all my old high school year books .
	L6	[Sentence] Selected Atomic Events => [Situation]
(d) Tourism	L1	Translate [Sentence] to [Situation]:
	L2	[Sentence] PersonX travels . => [Situation]My aunt traveled all by herself .
	L3	[Sentence] PersonX takes PersonY to work . => [Situation] My manager has taken my coworker for a work trip to New York.
	L4	[Sentence] PersonX plans PersonX's next trip . => [Situation] We are planning our first cruise .
	L5	[Sentence] PersonX sees PersonY's . => [Situation]I have seen some pictures of friends of mine traveling around Italy.
	L6	[Sentence] Selected Atomic Events => [Situation]
(e) Relationship	L1	Translate [Sentence] to [Situation]:
	L2	[Sentence] PersonX gets married => [Situation] One of old coworkers is getting married .
	L3	[Sentence] PersonX gives PersonY . => [Situation] I gave a house key to a girl i had went on 2 dates with .
	L4	[Sentence] PersonX is on PersonX's . => [Situation] I am on my first blind date .
	L5	[Sentence] PersonX gets PersonY's number . => [Situation] I got the new girl's number at school .
	L6	[Sentence] Selected Atomic Events => [Situation]

Table 11: Prompts for collecting **Situations** under five common daily topics. Each prompt input consists of 6 lines of content. The first line L1 is an **Introduction**, lines L2 - L5 are **Demonstration**, and the last line L6 is **Query**.

Type		Prompt inputs for collecting *Thoughts*
(a) Thought in the negative cognitive Chain	L1	Complete the sentence:
	L2	WhenI will give a presentation at college, I feel terrible since I think I will freeze on stage.
	L3	WhenI was waiting in the waiting room for my last job interview, I felt terrible since I thoughtI would not be selected.
	L4	When I lost the wallet, I felt terrible since I thoughtI would be blamed.
	L5	When Situation, I feel terrible since I think
(b) Thought in the positive cognitive chain	L1	Complete the sentence:
	L2	When I will give a presentation at college, I feel great since I think I am going to ace this presentation.
	L3	When I was waiting in the waiting room for my last job interview, I felt great since I thought I would land my dream job.
	L4	When I gave a house key to the girl I had gone on two dates with, I felt great since I thought she would fall in love with me.
	L5	When Situation, I feel great since I think

Type		Prompt inputs for collecting *Clues*
(a) Clue in the negative cognitive chain	L1	Complete the sentence:
	L2	When I will give a presentation at college, I thinkI will freeze on stagesinceI haven't given a presentation before.
	L3	When I was waiting in the waiting room for my last job interview, I thoughtI would not be selected sincetoo many people were interviewing for the job.
	L4	When my college entrance exam is tomorrow, I thinkI will fail the examsinceI didn't prepare well.
	L5	When Situation, I think Thought since
(b) Clue in the positive cognitive chain	L1	Complete the sentence:
	L2	WhenI will give a presentation at college, I think I will win the appreciation of my teachers and colleagues sincethey clap their hands happily.
	L3	When my boss takes me to a nice restaurant, I think he likes me and respects my opinionsinceI have gone out with him before and he has always seemed to enjoy my company.
	L4	When I will have a meeting with my boss, I thinkit is an opportunity to show myself sinceI have great knowledge about my work.
	L5	When Situation, I think Thought since

Type		Prompt inputs for collecting *Actions*
(a) Action in the negative cognitive chain	L1	Complete the sentence:
	L2	When I will give a presentation at college , I think I will freeze on stage, so I rehearse in front of the mirror .
	L3	WhenI was waiting in the waiting room for my last job interview, I thought I would not be selected, so I took a deep breath repeatedly.
	L4	When my college entrance exam is tomorrow, I think I will fail the exam, so I scan through some old papers to review.
	L5	When Situation, I think Thought, so
(b) Action in the positive cognitive chain	L1	Complete the sentence:
	L2	When I will give a presentation at college, I thinkI am going to ace this presentation, so I take good sleep at night.
	L3	When I was waiting in the waiting room for my last job interview, I thoughtI would land my dream job, so I talked to other candidates confidently.
	L4	When my boss took me to a nice restaurant, I thought it was a good sign of a promotion, so I talked to him about my career plan.
	L5	When Situation, I think Thought, so

Table 12: Prompts for collecting **Thoughts**, **Clues** and **Actions** in negative and positive cognitive chains. Each prompt input consists of 5 lines of content. The first line L1 is an **Introduction**, lines L2 - L4 are **Demonstration**, and the last line L5 is **Query**.

Type		Prompt inputs for collecting *Emotions*
	L1	Situation : I arranged an interview for my friend with the CEO.
	L2	Thought :It will go badly.
	L3	Question : Choose one word to describe my feeling about the Situation when I have the Thought.
	L4	Choice : Angry, Fearful, Sad
	L5	Answer : Fearful
	L6	Situation : The boss asserted his authority by yelling at the employees.
	L7	Thought : He's a terrible person.
	L8	Question : Choose one word to describe my feeling about the Situation when I have the Thought.
	L9	Choice : Angry, Fearful, Sad
(a) Emotion generation in the the negative chain	L10	Answer : Angry
	L11	Situation : I asked my daughter to help me with the dishes.\n
	L12	Thought : she will refuse.
	L13	Question : Choose one word to describe my feeling about the Situation when I have the Thought.
	L14	Choice : Angry, Fearful, Sad
	L15	Answer : Sad
	L16	Situation : Situation Extended from Atomic Event
	L17	Thought : Thought automatically generated and manually evaluated.
	L18	Question : Choose one word to describe my feeling about the Situation when I have the Thought.
	L19	Choice : Angry, Fearful, Sad
	L20	Answer:
	L1	Situation :I arranged an interview for my friend with the CEO.
	L2	Thought :I will help my friend's career.
	L3	Question : Choose one word to describe my feeling about the Thought when I am in the Situation.
	L4	Choice : Joyful, Love, Surprised
	L5	Answer : Joyful
	L6	Situation : I also took out the new girl on a date.
	L7	Thought : She will call me her boyfriend.
	L8	Question : Choose one word to describe my feeling about the Situation when I have the Thought.
	L9	Choice : Joyful, Love, Surprised
(b) Emotion generation in the the positive chain	L10	Answer : Love
	L11	Situation : I broke my leg skiing.
	L12	Thought : I will get a snowboard for my birthday.
	L13	Question : Choose one word to describe my feeling about the Thought when I am in the Situation.
	L14	Choice : Joyful, Love, Surprised
	L15	Answer : Surprised
	L16	Situation: Situation Extended from Atomic Event
	L17	Thought: Thought automatically generated and manually evaluated.
	L18	Question: Choose one word to describe my feeling about the Thought when I am in the Situation.
	L19	Choice: Joyful, Love, Surprised
	L20	Answer:

Table 13: Prompts for collecting **Emotion** labels in negative and positive cognitive chains. Each prompt input consists of 20 lines, and every 5 lines is a case.

Instruction for Thought Selection and Revision

Theory of Mind

Theory of mind refers to humans' ability to understand and infer the desires, beliefs, and intentions of others. People in the same situation will think differently due to their different personal experiences and the different characteristics of the personalities of others involved in the situations. Then, people will have different behavioral and affective responses to the same situation due to their different thoughts.

We want to collect individual **thoughts, clues, actions,** and **emotions** when they are in the situation. Therefore, we can use these data to predict others' mental states and behavioral response affective response to make the Theory of Mind inference.

Thought Definition

Negative Thought: thoughts trigger negative emotion

Positive Thought: thoughts trigger positive emotion

We just find out what they think when people are in a situation.

Task Description:

Find the reasonable thoughts from the 'thought1' 'thought2' 'thought3'

Annotation Examples：

1.

Situation: I will give a presentation to my classmates and teachers tomorrow.

Negative Thought: I think it is difficult for me to give a good speech

2.

Situation: When I am in the waiting room preparing for the next interview

Positive Thought: I feel sure that I will get the job.

3.

Situation: I'm interviewing for a new job today.

Positive Thought: I feel sure I will get the job position.

Negative Thought: I am not favored by the interviewer.

Incorrect Data：

1. Polarity inconsistency: the polarity of the thought candidate is inconsistent with the chain.
2. Containing specific emotions: happy, sad, anxious, etc.
3. A specific action rather than a thought e.g. I sighed (action)
4. Conflicting with the fact:
   Situation： The dummy model is broken.
   Negative Thought: The blood stains clothes.

Figure 5: Instruction for Thought selection and revision.

Instruction for Clue Selection and Revision

Clue Definition

Clue concretizes and differentiates **Thought.**

Task Description:

Find the reasonable Clue from the ‘clue1’ ‘clue2’ ‘clue3’

Categories

Clues can be divided into several categories.

Other’s: others' character, others' experience, others' appearance, others' behavior, etc.

Person: my character, my appearance, my experience.

Event: event characteristics, event flow.

Object: characteristics of the object, etc.

Social: social knowledge, social norms

Annotation Examples:

1.
   Situation: I am doing a presentation to my classmates and teachers.
   Positive Thought: I think they must have enjoyed my presentation.
   Clue: Because they all applauded. (Others’ Action)
2.
   Situation: I'm interviewing for a new job today.
   Positive Thought: I feel sure I will get the job position.
   Clue: I have a lot of work experience. (My Clue)
   Negative Thought: I am not favored by the interviewer.
   Clue: The interviewer kept looking at the computer screen and did not look at me. (Others’ Action)
3.
   Situation: My boss invites me to dinner.
   Positive Thought: He will respect my choice.
   Clue: He is polite person. (Others’ Personality)
4.
   Situation: I dropped my phone on the floor
   Positive Thought: I will have a new phone
   Clue: My phone is still under warranty. (Object Clue)

Incorrect Data:

1. Clues conflict with the Situation or cannot trigger the Thought. (Filter out these Clues)
2. Some irrelevant content with the reasonable Clue content. (Delete the irrelevant part)

Figure 6: Instruction for Clue selection and revision.

Instruction for Action Selection and Revision

Action Definition

The actions denote the humans' behavioral responses which are controlled and guided by the thoughts in specific situations.

Task Description:

Find the reasonable actions from the 'action1' 'action2' 'action3'

Annotation Examples:

1.
   Situation: I will give a presentation to my classmates and teachers tomorrow.
   Negative Thought: I think it is difficult for me to give a good speech
   Action: I keep repeating the speech in front of the mirror.
2.
   Situation: When I am in the waiting room preparing for the next interview
   Positive Thought: I feel sure that I will get the job.
   Action: I am confident talking to the other candidates
3.
   Situation: When I am in the waiting room preparing for the next interview
   Negative Thought: I am not favored by the interviewer.
   Action: I am in the hallway constantly recalling answers to pre-prepared interview questions.
4.
   Situation: I dropped my phone on the floor
   Negative Thought: I need to spend a lot of money to buy a new phone
   Action: I go to check my bank card balance.
5.
   Situation: I have an exam tomorrow.
   Negative Thought: I am worried that I will fail the exam tomorrow
   Behavior: I browse a lot of exam-related past exam questions.

Incorrect Data:

1. Others' action (Keep only my action after having the thought)
2. Action conflict with the Situation or Thought. (Filter out these actions)
3. Some irrelevant content with the reasonable action content. (Delete the irrelevant part)

Figure 7: Instruction for Action selection and revision.

Instruction for Emotion Selection and Revision

Task Description:

Give the emotion label to describe the feeling that the Thought evoke.

There are two class:

Positive (Joyful，Love，Surprised)

Negative(Angry，Fearful，Sad)

The feelings that each label describes:

Angry: Disgusting Envious Angry

Fearful: Horror Nervous

Sad: Disappointed Ashamed Sad

Joyful: Optimistic Proud Relieved Joyful

Love: Loveful

Surprised: Surprised

Annotation Examples：

1. I arranged an interview for my friend with the CEO.
    - It will go badly. **Sad**
    - he may not get the job. **Sad**
2. I asked the cashier for directions to the nearest post office
    - He refuses to let me know. **Angry**
3. I broke my leg skiing.
    - I will have to stay home all winter. **Sad**
    - I will have to have surgery. **Fearful**
4. I broke my leg skiing.
    - I will get a lot of attention. **Love**
    - I will get a snowboard for my birthday. **Surprised**
    - It'll heal soon. **Joyful**

Incorrect Data ：

Filter Out:

1. Inference not in the emotion labels. =====>Give a label
2. Inference cannot describe the feeling of thought. ====> Give a new label

Tips:

The emotion can be evoked by the specific factors:

Fearful:

Negative impact of rejection/ Potential for failure /Loss of control, loss of integrity /Unfamiliar things, new things /Being alone /Being in the dark /Death /Surgery

Sad:

Not Expected Results/ Separation / Rejected by someone /Not getting what you want / Losing power /Sadness of a loved one /Death of a loved one

Angry:

Being treated differently/ Forced to feel pain/Expected to be disappointed

Figure 8: Instruction for Emotion selection and revision.

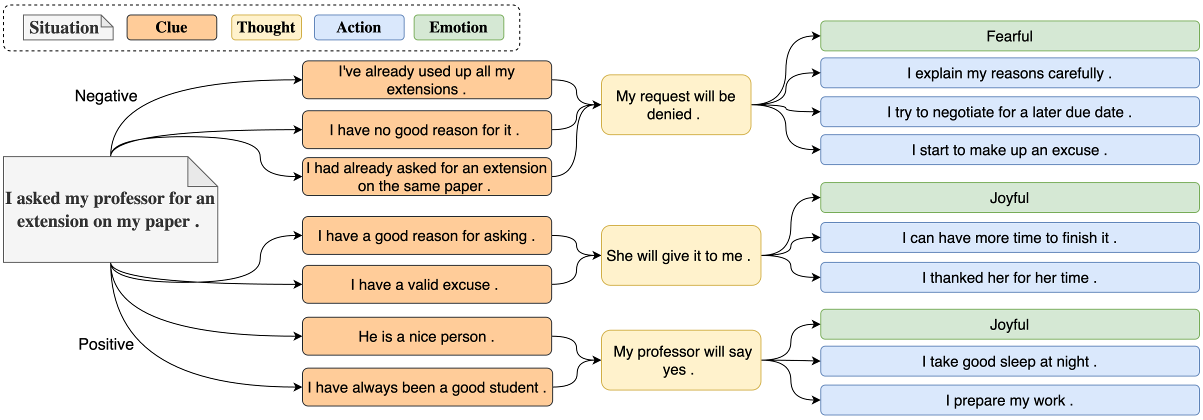

Figure 9: Examples of cognitive chains in the situation from the **School** topic in COKE.

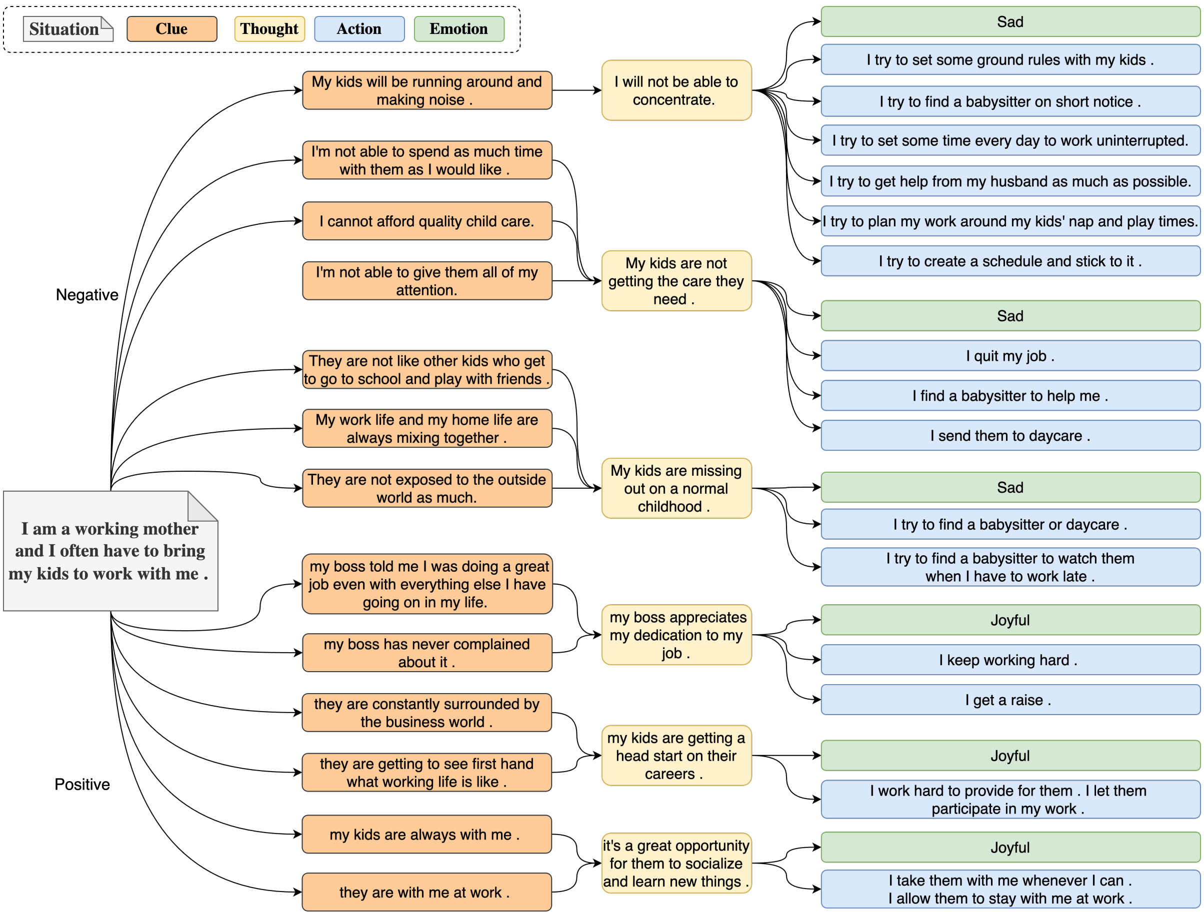

Figure 10: Examples of cognitive chains in the situation from the **Work** topic in COKE.

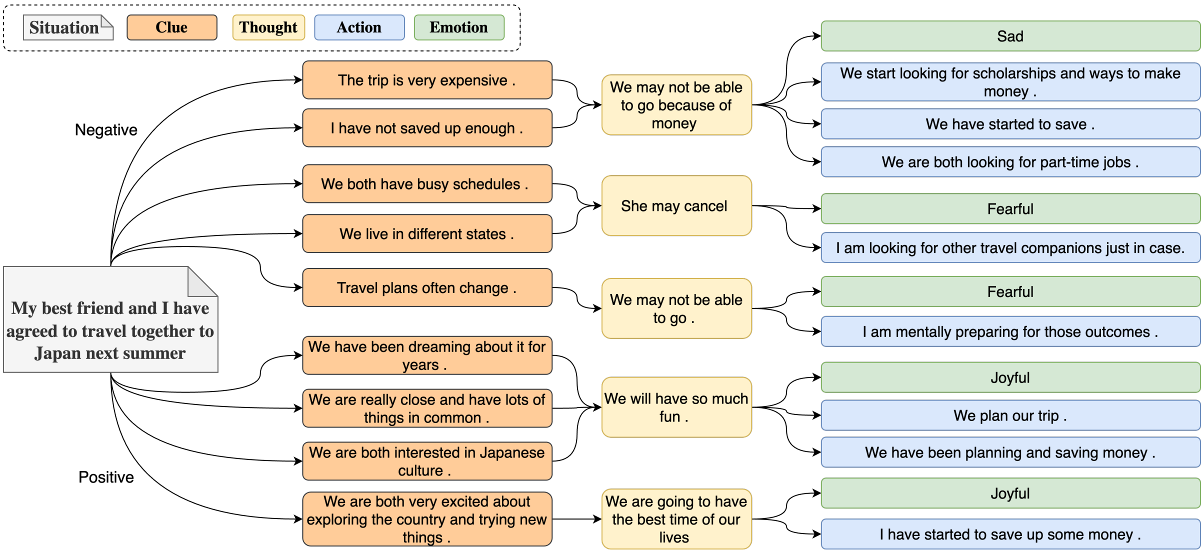

Figure 11: Examples of cognitive chains in the situation from the **Tourism** topic in COKE.

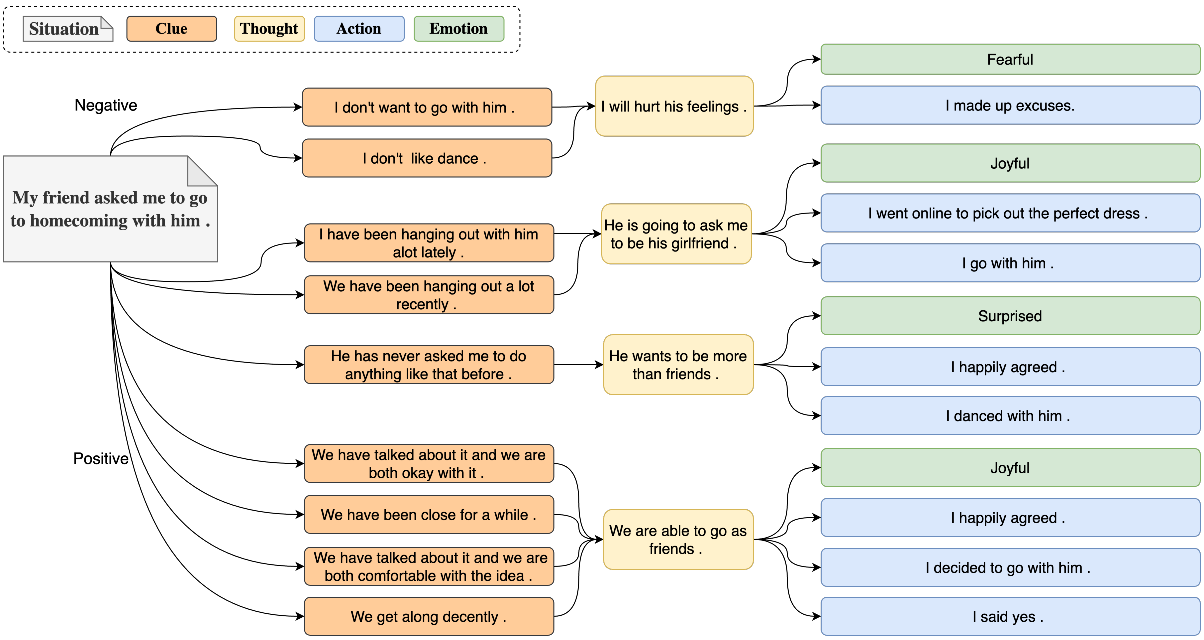

Figure 12: Examples of cognitive chains in the situation from the **Relationship** topic in COKE.

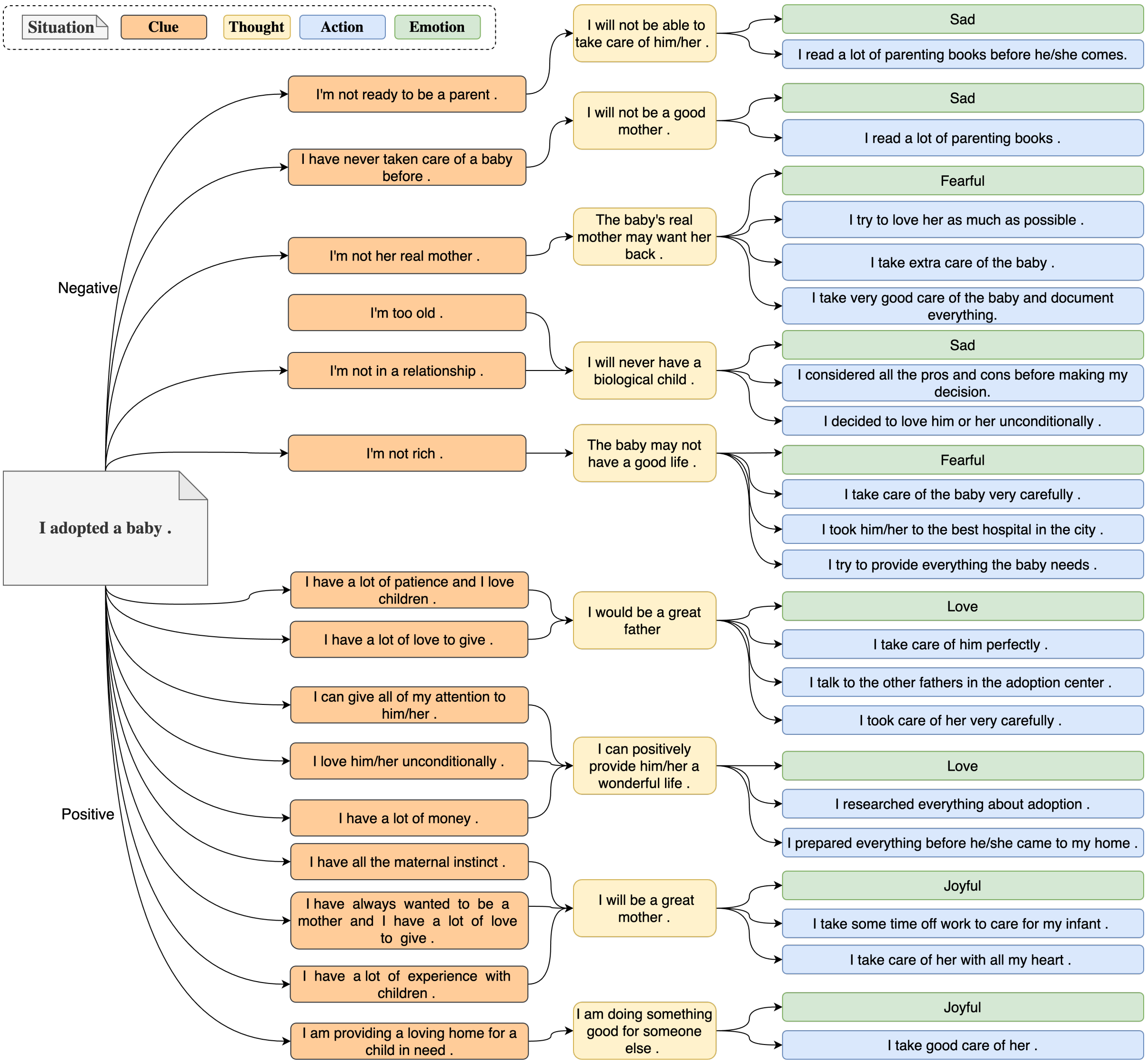

Figure 13: Examples of cognitive chains in the situation from the **Ordinary Life** topic in COKE.