KoACD: The First Korean Adolescent Dataset for Cognitive Distortion Analysis

Anonymous ACL submission

Abstract

Cognitive distortion refers to negative thinking patterns that can lead to mental health issues like depression and anxiety in adolescents. Previous studies using natural language processing (NLP) have focused mainly on small-scale adult datasets, with limited research on adolescents. This study introduces KoACD, the first large-scale dataset of cognitive distortions in Korean adolescents, containing 108,717 instances. We applied a multi-Large Language Model (LLM) negotiation method to refine distortion classification and generate synthetic data using two approaches: cognitive clarification for textual clarity and cognitive balancing for diverse distortion representation. Validation through LLMs and expert evaluations showed that while LLMs classified distortions with explicit markers, they struggled with context-dependent reasoning, where human evaluators demonstrated higher accuracy. KoACD aims to enhance future research on cognitive distortion detection.

1 Introduction

Negative thoughts (Pietromonaco et al., 1985) are a natural part of human cognition, often helping individuals recognize potential dangers, prepare for challenges, or engage in self-reflection. However, when these patterns become rigid and excessive, they can lead to emotional distress and contribute to mental health issues such as depression and anxiety disorders. Adolescents, in particular, may be more vulnerable to these maladaptive thought patterns due to their ongoing cognitive and emotional development.

Globally, one in seven children and adolescents (about 166 million) suffers from mental illness, with 42.9% experiencing anxiety and depression (UNICEF, 2018). The rising prevalence of these conditions during adolescence has become a serious global concern. Since this stage is crucial for self-identity formation and emotional regulation (Pfeifer et al., 2018), understanding how negative thought patterns emerge and persist is essential for early intervention and prevention.

In particular, depression is frequently associated with habitual negative thinking, known as cognitive distortion (Rnic et al., 2016). Adolescents experiencing these distortions may instinctively blame themselves when something goes wrong, thinking, 'I messed up again' or 'I completely failed.' These persistent negative thoughts trigger emotional distress, reinforcing cycles of depression (Chahar et al., 2020). Identifying and analyzing these patterns is essential for developing effective coping strategies. To achieve this, building a comprehensive dataset of adolescent cognitive distortions is necessary to enable more targeted and impactful mental health interventions. Our KoACD dataset is will be publicly released once accepted.

2 Related Work

2.1 Cognitive distortions Detection

Cognitive distortions are closely related to negative thinking patterns and are defined as distorted ways of thinking that reinforce negative emotions (Beck, 1979). With the recent development of natural language processing (NLP), research has been actively conducted to automatically classify cognitive distortions using various datasets.

Previous studies on cognitive distortion classification have primarily relied on small-scale, adult-focused, and English-language datasets. Early research utilized LIWC-based regression models on social media posts (Simms et al., 2017), while later studies adopted deep learning models, including RNNs, CNNs, and BERT, using datasets from counseling platforms and therapist-patient

Dataset	Language	Sample	Target	Data Source	Classification
Tumblr Cognitive Distortion (Simms et al., 2017)*	English	459	nonspecific	Tumblr blogs	Binary (2)
MH-C (Shickel et al., 2020)	English	1,164	Adult	TAO Connect	Multi-class (15)
MH-D (Shickel et al., 2020)	English	1,799	Adult	TAO Connect	Binary (2)
CrowdDist (Shickel et al., 2020)	English	7,666	Adult	Mechanical Turk	Multi-class (15)
Clinician-Client SMS (Tauscher et al., 2023)*	English	7,354	Adult	Clinician-Client SMS	Multi-class (5)
SocialCD-3K (Qi et al., 2024)	Chinese	3,407	nonspecific	Weibo	Multi-label (12)

Table 1: Summary of datasets for cognitive distortion detection: The 'Sample' column indicates the number of instances, and 'Classification' specifies the type (binary, multi-class, or multi-label). *Indicates unofficially named datasets.

conversations (Shickel et al., 2020; Tauscher et al., 2023). More recent approaches have leveraged Large Language Models (LLMs) for cognitive distortion classification, further enhancing performance and adaptability across diverse datasets (Chen et al., 2023; Qi et al., 2024).

Table 1 summarizes existing datasets that deal with cognitive distortions.

Despite these advancements, existing datasets remain limited in scale and predominantly focus on English-speaking adults. To address these limitations, we propose a Korean-language dataset specifically designed for adolescents, filling a crucial gap in research on cognitive distortions in younger populations.

2.2 LLMs-Based Negotiation

Attempts have been made to explore the possibility of models going beyond independent judgment and deriving more sophisticated conclusions through interaction through negotiations between LLMs.

Self-Play and In-Context Learning techniques using AI feedback were applied to improve the negotiation capabilities of LLMs (Yao Fu et al., 2024), and a method for providing feedback by developing an LLM-based Assistant for Coaching nEgotiation (ACE) using negotiation data from MBA students was proposed (Ryan Shea et al., 2024). In addition, research has been conducted on applying negotiation methods to emotional analysis. It has been demonstrated that using LLM negotiation methods to interact between models can outperform the existing single-pass decision (Xiaofei Sun et al., 2024).

Previous studies have shown that LLM negotiations can yield sophisticated results, but challenges remain in balancing negotiation outcomes due to fixed roles and limited structures. To overcome this, we use role-switching and multiple types of LLMs to balance the negotiation process. We also introduce multi-round negotiations to give equal consideration to the 10 cognitive distortions, ultimately arriving at the optimal conclusion. Therefore, this study aims to generate and validate data using LLM negotiation techniques.

3 Constructing KoACD

3.1 Data Source and Preprocessing

We crawled posts on NAVER Knowledge iN[1], a Q&A platform where users can post questions and receive answers, to analyze the cognitive distortions of Korean adolescents. NAVER Knowledge iN remains widely used in Korea, even with the rise of search engines and generative AI (Jang & Kim, 2024). Since Naver Knowledge covers a wide range of age groups, we used only data from five major adolescent counseling organizations and services, covering the years 2011–2024, to focus on adolescent concerns. A total of 69,925 questions were collected, and the distribution of data sources is provided in Appendix A.

A pre-processing step refined the data to align with the research purpose and excluded irrelevant questions. This included removing entries from elementary students or adults, filtering inappropriate content, deleting vague questions, and eliminating duplicates. After applying these criteria, 37,124 questions remained for analysis. Details on preprocessing and removed cases are in Appendix B.

3.2 Definition of Cognitive Distortions

Aaron Beck, a pioneer in cognitive therapy (Beck, 1979), identified 10 cognitive distortions in patients with depression and incorporated them

[1] https://kin.naver.com/

Cognitive Distortion Type	Definition	Examples
All-or-Nothing Thinking	Viewing situations in only two categories (e.g., perfect or failure) instead of on a spectrum.	“If I fail this test, I’m a total failure.”
Overgeneralization	Drawing broad conclusions from a single event or limited evidence.	“My one friend ignored me, so everyone else will hate me too.”
Mental Filtering	Focusing only on the negative aspects of a situation while ignoring the positive.	“I only remember my mistake though I got compliments on my presentation.”
Discounting the Positive	Rejecting positive experiences or compliments by insisting they don’t count.	“People told me I did well, but I was just being polite.”
Jumping to Conclusions	Predicting negative outcomes without evidence.	"She didn’t text back. She must be mad at me."
Magnification and Minimization	Exaggerating negative or risky aspects while minimizing positive or positive aspects.	“One little mistake at work means I’m incompetent.”
Emotional Reasoning	Believing something must be true because you feel it strongly.	"I feel worthless, so I must be worthless."
"Should" Statements	Holding rigid rules about how you or others should behave, leading to guilt or frustration.	"I should always be productive; otherwise, I’m lazy."
Labeling	Assigning negative labels to yourself or others based on one event.	"I made a mistake, so I’m a total failure."
Personalization	Blaming yourself for events outside your control or assuming excessive responsibility.	"My friend looks sad, maybe I did something wrong."

Table 2: Classification of cognitive distortions with definitions and examples

into psychotherapy. He emphasized that reducing these distortions could alleviate stress and anxiety (Beck, 1991). We used these distortions, listed in Table 2, to classify questions reflecting the emotional struggles commonly reported by adolescents.

3.3 Multi-LLMs Negotiation for Identifying Cognitive Distortions

To effectively identify cognitive distortions, this study designed a process for deriving optimal distortions using the multi-LLM negotiation method (Yao Fu et al., 2024), where relevant distortions are gradually derived through LLM interactions.

This study uses a multi-LLM negotiation method based on the interaction between Google’s Gemini 1.5 Flash (Team et al., 2024) and OpenAI’s GPT-4o mini (OpenAI, 2024). The two models work together to identify the most accurate cognitive distortion. One model acts as the Analyzer and the other as the Evaluator. Through their collaboration, cognitive distortions are gradually refined. Negotiation is conducted up to 5 rounds to systematically explore all 10 predefined cognitive distortions, as each round consists of two turns, evaluating one distortion per turn. This structure allows for the possibility that the same sentence may be interpreted in multiple cognitive distortions before reaching a final classification. Table 11 in Appendix E details the LLMs parameters used in this process.

Here’s how the roles work:

- Analyzer: Identifies the most relevant cognitive distortion in a sentence and suggests sentences that match it.
- Evaluator: Reviews the suggestions made by the Analyzer and provides feedback on their accuracy.

The prompts used for these roles are detailed in Appendix H. In each round of negotiation, the models take turns playing the roles of Analyzer and Evaluator.

A round consists of two turns and proceeds in the following structure:

- T1 (Initial Analysis): Identify the most relevant cognitive distortion in the sentence. (Options: one of the 10 cognitive distortions.)
- T1 (Evaluation): Assess whether the proposed cognitive distortion from T1 (Initial Analysis) accurately reflects the distortion present in the sentence. (Options: "Yes" or "No." The evaluator provides a justification.)
- T2 (Reanalysis): If T1 (Evaluation) results in rejection, select the next most relevant cognitive distortion, excluding previously rejected options. (Options: one of the remaining cognitive distortions.)
- T2 (Evaluation): Determine whether the cognitive distortion from T2 (Reanalysis) is appropriate. (Options: "Yes" or "No." The evaluator provides a justification.)

Each step is performed sequentially, incorporating feedback from the previous evaluation. T1 (Evaluation) assesses the distortion proposed in T1 (Initial Analysis), and T2 (Reanalysis) refines the selection based on that feedback. Similarly, T2 (Evaluation) verifies the suitability of the distortion chosen in T2 (Reanalysis).

Throughout the negotiation process, distortions deemed inappropriate are systematically excluded, ensuring the selection of the most fitting cognitive distortion. To maintain fairness, the models

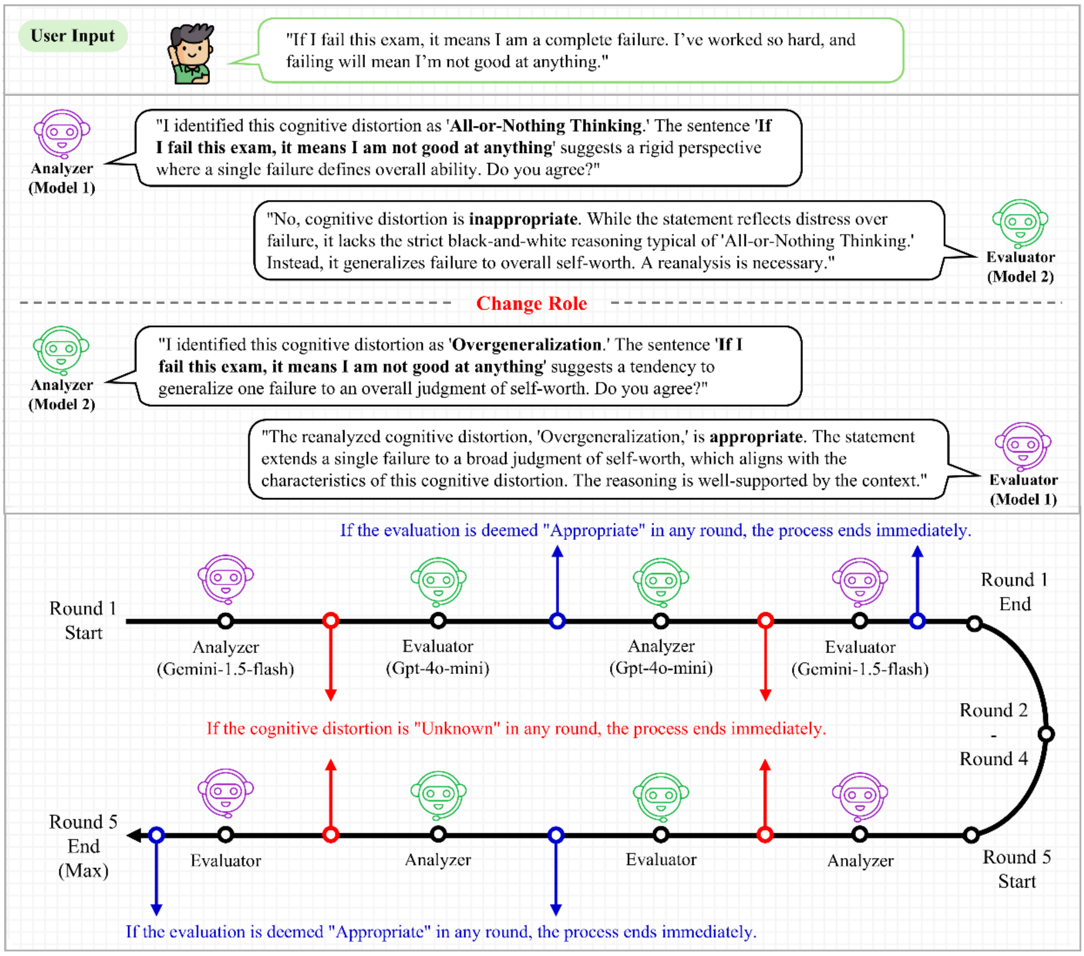

Figure 1: Process for identifying and evaluating cognitive distortions through negotiation

alternate roles in T2 (Reanalysis) so that both contribute equally to the negotiation.

If a consensus is not reached after five rounds, the question is classified as unknown. This indicates that all cognitive distortions proposed during the negotiation process were considered inherently inappropriate.

The number of turns required to identify cognitive distortions or classify the question as unknown varies from dataset to dataset. Some sentences reach conclusions early, while others require multiple turns for final classification. Details of turn counts and classification ratios are given in Appendix C. The overall structure of this negotiation process is illustrated in Figure 1.

3.4 Independent Evaluation

After the negotiation process is complete, Anthropic's Claude 3 Haiku (Anthropic, 2024) is used for an independent validation of the final cognitive distortion and its corresponding sentence. Claude 3 Haiku is not involved in the negotiation process; instead, it evaluates whether the selected cognitive distortion correctly aligns with the given sentence.

Score	Count	Proportion (%)
1	11	0.06
2	874	4.41
3	18,897	95.53
Total	19,782	100.00

Table 3: Distribution of validation scores

During negotiation, the models assess cognitive distortions in the context of the entire original text, whereas Claude 3 Haiku determines the appropriateness based solely on the selected sentence. This additional validation step helps identify potential misclassifications and ensures that the cognitive distortion is properly connected to the sentence.

Claude 3 Haiku assigns a relevance score from 1 to 3, and only cognitive distortion–sentence pairs that receive a score of 3 are used as the final-stage data for generating synthetic data. A summary of the validation score distribution is presented in Table 3. The parameters used for this validation are detailed in Table 11 (Appendix E),

Cognitive Distortion Type	Cognitive Clarification (%)	Cognitive Balancing	Total
All-or-Nothing Thinking	5,949 (10.50%)	4,920 (9.46%)	10,869 (10.00%)
Overgeneralization	11,418 (20.14%)	0 (0.00%)	11,418 (10.50%)
Mental Filtering	2,763 (4.88%)	8,139 (15.64%)	10,902 (10.03%)
Discounting the Positive	822 (1.45%)	9,873 (18.98%)	10,695 (9.84%)
Jumping to Conclusions	10,479 (18.48%)	183 (0.35%)	10,662 (9.81%)
Magnification and Minimization	6,078 (10.72%)	4,836 (9.30%)	10,914 (10.04%)
Emotional Reasoning	10,842 (19.12%)	0 (0.00%)	10,842 (9.98%)
"Should" Statements	2,697 (4.76%)	7,998 (15.37%)	10,695 (9.84%)
Labeling	2,373 (4.19%)	8,463 (16.27%)	10,836 (9.97%)
Personalization	3,270 (5.77%)	7,614 (14.63%)	10,884 (10.01%)
Total	56,691 (100.00%)	52,026 (100.00%)	108,717 (100.00%)

Table 4: Distribution of cognitive distortion types across synthetic data generation methods

and the prompts used for evaluation are provided in Appendix H.

3.5 Synthetic Data Generation

The original data consists of free-form text written by adolescents, often containing spelling errors, excessive use of emoticons, or unclear wording, making it difficult to interpret. Additionally, some texts lack contextual coherence, with disjointed narratives or insufficient background information to accurately assess cognitive distortions. As a result, the data could be difficult to use as is.

Furthermore, the distribution of the 10 cognitive distortion categories we propose was imbalanced, leading to a potential bias in the dataset. To address these issues, we employ two methods to generate synthetic data. The prompts used for both methods are provided in Appendix H.

3.5.1 Cognitive Clarification of Cognitive Distortions

The first approach to generating synthetic data is to identify cognitive distortions and rephrase the text in a clearer and more structured form while preserving the meaning of the original text: maintaining the emotional tone and context.

We used three LLMs—Gemini 1.5 Flash, Claude 3 Haiku, and GPT-4o mini—independently to generate a wide variety of expressions, ensuring greater diversity in the generated content. The parameters of these models used for synthetic data generation are detailed as shown in Table 11 in Appendix E.

3.5.2 Balancing Cognitive Distortions with Context-Preserved Data

The second approach we adopted aimed to address the imbalance of cognitive distortions by utilizing data classified as 'Unknown' or data for which a suitable cognitive distortion could not be identified. First, we analyzed the distribution of cognitive distortions to detect which types were underrepresented. Then, synthetic data was generated by reconstructing and reorganizing the original data, ensuring that the overall context was preserved.

Table 4 summarizes the distribution of cognitive distortions produced through both the cognitive clarification and cognitive balancing methods, along with the overall total after combining both approaches.

4 Validating Synthetic Data with Clustering

To verify the validity of the synthetic data we created, we performed clustering based on two criteria: (1) topics that trigger negative emotions in adolescents and (2) negative emotions and symptoms outlined in the DSM-5[2], a widely used framework for assessing and diagnosing mental disorders (Lee et al., 2023).

4.1 Topic-Based Classification of Adolescent Negative Thinking

The Korea National Youth Policy Institute (NYPI[3]), under the Ministry of Gender Equality and Family, categorized adolescents' concerns into five areas:

[2] https://www.mdcalc.com/calc/10195/dsm-5-criteria-major-depressive-disorder

[3] https://www.nypi.re.kr/

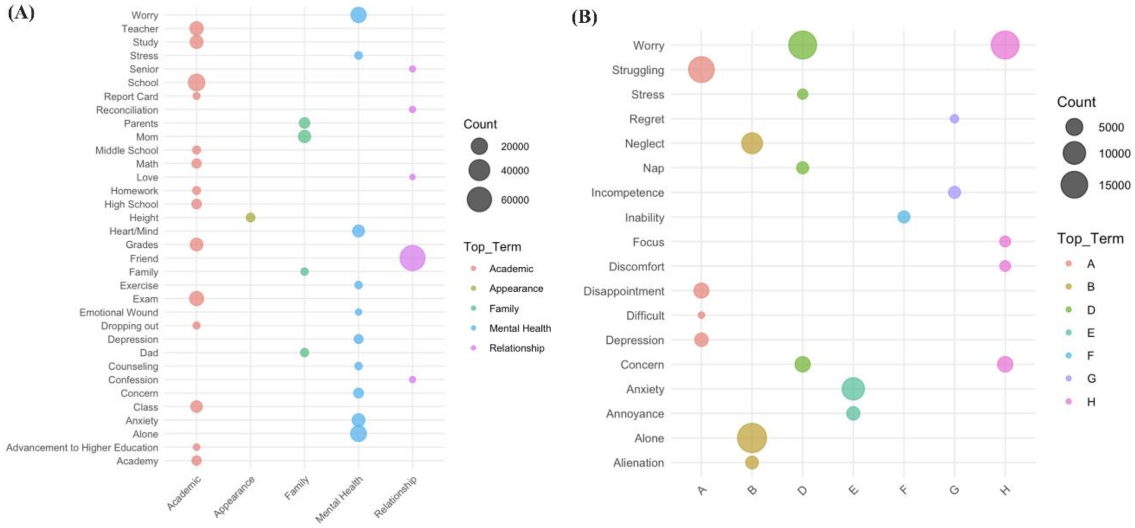

Figure 2: Cluster distribution of high-frequency by (A) negative emotion-triggering topics, (B) DSM-5 symptom keywords (≥1,000 occurrences)

(1) academic and career concerns, (2) relationships (friendships, romance, bullying), (3) physical and mental health, (4) family issues, and (5) appearance and self-image.

To assess the alignment of our synthetic data on adolescent negative thinking with these categories, we applied K-means clustering, an unsupervised machine learning algorithm that partitions data into distinct groups, to keywords extracted from 69,925 adolescents' questions (Section 3.1). This process grouped the data into the five predefined subject clusters, each of which was assigned sub-keywords based on relevance. As a result, a dictionary with five topics and 139 keywords was created, as shown in Table 12 in Appendix F.

We identified the most frequent keywords for each topic. The top topic was *Academic performance and career concerns*, with 99,076 times (36.9%), followed by *Relationships* (73,586, 27.4%), *Physical and mental health* (71,249, 26.5%), *Family issues* (20,532, 7.6%), and *Appearance and self-image* (4,007, 1.5%).

4.2 DSM-5 based Classification of Adolescent Negative Thinking

Cognitive distortions can contribute to depression, so we examined the nine categories of the DSM-5 to determine whether a significant relationship exists. To explore this, we analyzed 69,925 adolescents' questions (Section 3.1). and identified DSM-5-related word distributions using NLTK text mining[4] (3.9.1). These distributions were then used to create dictionaries for DSM classification, resulting in nine categories and 143 keywords, as shown in Table 13 in Appendix F.

For keyword mapping, we used our dataset of 108,717 synthesized data points (Section 3.5.2), allowing multiple keywords per data point. For DSM-based keyword mapping, 69,290 data points (63.7%) were successfully mapped, with 115 unique keywords assigned 1,335,337 times. For negative emotion-triggering topic-based mapping, 103,183 data points (94.9%) were successfully mapped, with 129 unique keywords assigned 268,450 times.

Among the DSM-5 symptom categories, five out of nine categories appeared more than 15,000 times. The most frequent keyword was B. Loss of interest or pleasure (321,157 occurrences, 23.8%), followed by H. Decreased concentration (25,580 occurrences, 18.9%), A. Depressed mood (25,258 occurrences, 18.7%), D. Insomnia or hypersomnia (24,864 occurrences, 18.4%), and E. Psychomotor agitation or retardation (15,235 occurrences, 11.3%).

We found 34 keywords (Table 14 in Appendix F**)** for cognitive distortion-triggering topics and 20 (Table 15 in Appendix F**)** for DSM-5 categories, each with a frequency of 1,000 or more, are listed in Figure 2.

The generated synthetic data mainly highlighted academic and career stress, along with social conflicts like friendships and romantic relationships, while underrepresenting appearance and self-image issues. Additionally, its cognitive distortions were closely linked to five of the nine DSM-5 depression symptom keywords.

[4] https://www.nltk.org/book/ch07.html

Criteria	LLMs Evaluation		Human Evaluation	
	Cognitive Clarification	Cognitive Balancing	Cognitive Clarification	Cognitive Balancing
Consistency	2.400 ± 0.232	2.105 ± 0.173	2.254	2.160
Accuracy	2.708 ± 0.177	2.416 ± 0.270	2.322	2.738
Fluency	2.655 ± 0.219	2.529 ± 0.223	2.904	2.690

Table 5: Evaluation results of synthetic data by LLMs and humans: The LLMs evaluation (left) reports mean ± standard deviation scores assigned by three models, where the standard deviation represents variations across models. The human evaluation (right) presents the average scores given by two experts after cross-validation.

Original Distress Question	*"Whenever I meet my cousins, my mom asks me why I'm not tall. Even my friends, who are shorter than me, stay up past 2 a.m., and when I see them grow taller, I can only think about why I'm not tall."*
[Type of Cognitive Distortion] Synthetic Data	**["Should" Statement]** *"My mom often compares me to her cousins and says, 'Why are you so short?' I don't understand why I'm the only one who's so short when all my friends are growing taller. Both my mom and dad are tall, but I feel like something is wrong with being short."*
Commands from Expert	The expert provided two points in the accuracy evaluation of the cognitive distortion type, and chose mental filtering rather than 'should' statement. *The belief that one should be tall ('should' statement) usually comes from parents. In this article, we confirmed that the fact that the mother is not tall triggers anxiety. However, this presupposes that the individual has negative thoughts (mental filtering), as she believes she will not grow taller. While the 'should' statement seems to be the main issue in literal terms, mental filtering—an error in self-judgment—is considered the primary cognitive distortion."*

Table 6: Expert analysis of case with synthetic data accuracy score of 2: Explanations for LLM misclassification

5 Evaluation

We evaluate the quality of two types of synthetic data: data generated from clearly identified cognitive distortions (Section 3.5.1) and data generated to address cognitive distortion imbalances (Section 3.5.2). The evaluation was conducted independently using three evaluation criteria, with both LLMs and human assessments.

5.1 Evaluation criteria

We evaluated the generated synthetic data using three criteria: (1) Consistency, (2) Accuracy, and (3) Fluency. Scores ranged from 1 to 3, with 1 indicating 'inappropriate' and 3 indicating 'highly appropriate'. Consistency checked if the cognitive distortion was logically maintained between the original and synthetic data. Accuracy assessed whether the labeled cognitive distortion matched the correct classification. Fluency evaluated how natural, grammatically correct, and easy to read the sentences were. The prompts used for these criteria are provided in Appendix H.

5.2 Comparison of LLMs and Human Evaluations Across Criteria

To ensure objectivity, the model generating synthetic data was excluded from evaluation. Two other models independently scored the data, averaging their scores for the final result. Evaluation parameters for the three LLMs are in Table 11, Appendix E.

For human evaluation, 50 or 100 synthetic samples per distortion were randomly selected, totaling 900. Two psychology experts independently assessed them using the same criteria as LLM evaluation, with a Cohen's kappa of 0.78 indicating substantial agreement.

Table 5 summarizes the evaluation results from both LLMs and humans, highlighting the differences between the two types of synthetic data (Section 3.5)—cognitive clarification and cognitive balancing—across the three criteria. Detailed results for each model are in Table 10 in Appendix D.

Human evaluation scores were lower across all criteria except fluency, with accuracy showing the largest gap. This difference stems from LLMs' strength in detecting explicit text patterns while struggling with the implicit reasoning essential for cognitive distortion evaluation, highlighting their limitations. Table 6 provides detailed expert feedback.

Regarding the two synthetic data generation methods, in the LLM evaluation, the cognitive clarification method scored 0.1 to 0.3 points higher on all criteria than the cognitive balancing method. However, in the human evaluation, only the cognitive balancing method showed higher accuracy.

Cognitive Distortion Type	LLMs Evaluation			Human Evaluation			Difference		
	Cos	**Acc**	**Flu**	**Cos**	**Acc**	**Flu**	**Cos**	**Acc**	**Flu**
All-or-Nothing Thinking	2.203	**2.607**	2.470	**2.610**	2.590	**2.730**	**0.407**	0.017*	0.260
Overgeneralization	**2.287**	**2.767**	2.609	2.280	2.520	**2.860**	0.007	0.247*	0.251
Mental Filter	2.247	**2.677**	2.578	**2.480**	2.460	**2.830**	0.233	0.217*	0.252
Discounting the Positive	**2.153**	2.240	2.640	2.120	**2.710**	**2.880**	0.033	**0.470**	0.240
Jumping to Conclusions	2.279	2.361	2.550	**2.560**	**2.890**	**2.840**	0.281	**0.529**	0.290
Magnification and Minimization	2.212	**2.531**	2.625	**2.330**	2.100	**2.730**	0.118	**0.431***	0.105
Emotional Reasoning	**2.624**	**2.887**	2.713	2.020	2.200	**2.880**	**0.604**	**0.687***	0.167
Should Statements	**2.315**	2.562	2.654	2.110	**2.600**	**2.770**	0.205	0.038	0.116
Labeling	2.309	2.563	2.499	**2.380**	**2.700**	**2.770**	0.071	0.137	0.271
Personalization	**2.250**	2.632	2.648	1.890	**2.810**	**2.860**	0.360	0.178	0.212
Total mean	2.287	2.582	2.598	2.278	2.558	2.815	0.231	0.295	0.216

Table 7: Comparative evaluation of cognitive distortions by LLMs and humans: Cos (Consistency), Acc (Accuracy), and Flu (Fluency). *Types of cognitive distortions easily detected by LLMs.

5.3 Comparison of LLM and Human Performance in Cognitive Distortion Classification

To further analyze the differences between LLM-based and human evaluations, we compared the scores for each cognitive distortion. Table 7 presents comparative results, highlighting key discrepancies between the two evaluation methods. Scores were compared between LLM and human evaluations, with the higher values in bold. The 'Difference' column shows score gaps, with differences of 0.4 or greater also bolded.

The average LLMs evaluation scores were 2.287 for consistency, 2.582 for accuracy, and 2.598 for fluency, while the average human evaluation scores were 2.278 for consistency, 2.558 for accuracy, and 2.815 for fluency. Fluency was higher in human evaluation, whereas consistency and accuracy showed no significant difference, though human scores were slightly lower overall. The higher fluency score in human evaluation is likely because LLMs assessed synthetically generated sentences, which were naturally structured and free of pauses.

In the evaluation of cognitive distortions by type, human scores were lower than those of LLMs in some cases, particularly in accuracy. For example, scores for "Emotional Reasoning" (2.887 vs. 2.200) and "Magnification and Minimization" (2.531 vs. 2.100) showed notable differences. This discrepancy may be because LLMs excel at detecting clear linguistic patterns, such as "Should Statements," "Labeling," and "Discounting the Positive." However, human evaluation tends to be more reliable for distortions requiring inferential reasoning, such as "Mental Filtering" and "Magnification and Minimization," since these rely on deeper contextual understanding.

These findings highlight that LLMs rely more on explicit linguistic patterns, whereas human evaluators consider deeper contextual reasoning, which may impact their ability to identify distortions that require implicit inference.

6 Conclusion and Future Work

We developed KoACD, a dataset of cognitive distortions in Korean adolescents, overcoming the limitations of small-scale datasets focused on English-speaking adults. KoACD offers a balanced representation of cognitive distortions through the creation of synthetic data. To our knowledge, it is the first dataset specifically designed for Korean adolescents.

We introduced a multi-LLM negotiation method to improve the objectivity and accuracy of the synthetic data. By using multiple LLMs to negotiate and refine cognitive distortion labels, we minimized biases and enhanced data quality. Expert and LLM evaluations confirmed that LLMs performed well when clear linguistic cues were present, while human evaluators showed higher accuracy in context-dependent situations. Discrepancies between LLMs and human evaluations highlighted the LLMs' reliance on superficial linguistic patterns.

Future work will focus on fine-tuning models with adolescent-specific data to enhance contextual understanding of cognitive distortions. Additionally, we aim to improve LLM performance by developing algorithms that better distinguish cognitive distortions, mitigating biases toward specific types and enhancing both balance and accuracy in detection.

7 Limitations

We recognize that there are some limitations to the methods for detecting cognitive distortions and to the KoACD dataset:

Cognitive Distortion Classification We assigned the most appropriate cognitive distortion to each question, but some questions may involve multiple distortions simultaneously. The boundaries between some types of distortions are blurred, making classification challenging and leading to potential discrepancies between the model and human raters. To address these issues, a multi-label classification method and more refined criteria are needed.

Multi-LLMs Negotiation Methods We designed the LLMs to alternate between Analyzer and Evaluator roles, but the results can vary depending on the model used. Therefore, negotiation results with different LLMs should also be considered. Additionally, discrepancies between analysts and evaluators sometimes result in data being classified as "Unknown," even after five rounds of negotiation, due to the inability to fit the data within the ten cognitive distortion categories. Interpretation of such data is essential, and further research is needed to develop more accurate detection methodologies.

LLMs and Human Evaluation While the KoACD is a large dataset, the amount of data reviewed by human raters is relatively small. Although human raters excel at considering context for accurate judgments, subjectivity in the evaluation process and inconsistency due to differing standards among raters may arise. Future research should focus on securing more human evaluation data and developing more precise evaluation standards to increase reliability.

8 Ethical Considerations

In this study, we collected publicly accessible data from NAVER Knowledge iN, and users participate anonymously on the platform. We only used publicly available data in the course of our research and did not interact directly with NAVER Knowledge iN users.

We have identified that the data collection process may include various inappropriate topics, such as hate speech, violence, sexual content, and profanity. Accordingly, we have attempted to exclude such data as much as possible by applying strict filtering criteria. However, we cannot completely rule out the possibility that some inappropriate content may be included in the data.

We are aware of the risk that AI models may be trained on inappropriate data and produce biased or unethical results. Therefore, it is important to continuously monitor the ethical use of AI models and improve filtering techniques to address this risk.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Aaron T. Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Aaron T. Beck. 1991. Cognitive therapy: A 30-year retrospective. *American Psychologist*, 46(4):368. https://doi.org/10.1037/0003-066X.46.4.368.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.284.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from AI feedback. *arXiv preprint*, arXiv:2305.10142. https://doi.org/10.48550/arXiv.2305.10142.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, HengTze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan,

George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe BarthMaron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence SafranekShrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakicevi ´ c, Dominik Grewe, ´ Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu 11 Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph HughesFitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani 12 Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin,

Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint, arXiv:2403.05530. https://doi.org/10.48550/arXiv.2403.05530.

Moonkyoung Jang and Seongcheol Kim. 2024. Key traits of top answers on Korean social Q&A platforms: Insights into user performance and entrepreneurial potential. *Humanities and Social Sciences Communications*, 11(1):1–13. https://doi.org/10.1057/s41599-024-03239-3.

Lee, Geonju, Dabin Park, and Hayoung Oh. 2023. Methodology of labeling according to 9 criteria of DSM-5. *Applied Sciences*, 13(10481). https://doi.org/10.3390/app131810481.

Saghar Chahar Mahali, Shadi Beshai, Justin R. Feeney, and Sandeep Mishra. 2020. Associations of negative cognitions, emotional regulation, and depression symptoms across four continents: International support for the cognitive model of depression. *BMC Psychiatry*, 20(18):1–12. https://doi.org/10.1186/s12888-019-2423-x.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Jennifer H. Pfeifer and Elliot T. Berkman. 2018. The development of self and identity in adolescence: Neural evidence and implications for a value-based choice perspective on motivated behavior. *Child Development Perspectives*, 12(3):158–164. https://doi.org/10.1111/cdep.12279.

Paula R. Pietromonaco and Hazel Markus. 1985. The nature of negative thoughts in depression. *Journal of Personality and Social Psychology*, 48(3):799–807. https://doi.org/10.1037/0022-3514.48.3.799.

Hongzhi Qi, Qing Zhao, Jianqiang Li, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yi Jing Yu, Fan Wang, Huijing Zou, Bing Xiang Yang, and Guanghui Fu. 2023. Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in Chinese social media. *Preprint, version 1*. Available at Research Square. https://doi.org/10.21203/rs.3.rs-3523508/v1.

Katerina Rnic, David J. A. Dozois, and Rod A. Martin. 2016. Cognitive distortions, humor styles, and depression. *Europe's Journal of Psychology*, 12(3):348–362. https://doi.org/10.5964/ejop.v12i3.1118.

Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. 2024. ACE: A LLM-based Negotiation Coaching System. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), pages 12720–12749, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.709.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *Proceedings of the 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE. https://doi.org/10.1109/BIBE50027.2020.00052.

Taetem Simms, Christopher Ramstedt, Marc Rich, Matthew Richards, Thomas Martinez, and Christophe Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 508–512. IEEE. https://doi.org/10.1109/ICHI.2017.39.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through LLM negotiations. *arXiv preprint*, arXiv:2311.01876. https://doi.org/10.48550/arXiv.2311.01876.

Justin S. Tauscher, Kevin Lybarger, Xiruo Ding, Ayesha Chander, William J. Hudenko, Trevor Cohen, and Dror Ben-Zeev. 2023. Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric Services*, 74(4):407–410. https://doi.org/10.1176/appi.ps.202100692.

UNICEF. 2018. *Ensuring mental health and well-being in an adolescent's formative years can foster a better transition from childhood to adulthood.* UNICEF.

Institution	2011 - 2015	2016 - 2020	2021 - 2024	Total
여성가족부.한국청소년상담복지개발원 Korea Youth Counseling & Welfare Institute	17,479	19,465	15,699	52,643
인천시 청소년 지원센터 Incheon Youth Support Center	1,357	1,355	-	2,712
울산시 청소년 지원센터 Ulsan Youth Support Center	1,655	2,360	1,518	5,533
경기도 청소년 지원센터 Gyeonggi Youth Support Center	6,862	-	53	6,915
청소년 모바일 상담센터 Youth Mobile Counseling Center	-	407	1,715	2,122
Overall Total	27,353	23,587	18,985	69,925

Table 8: Distribution of Q&A and worry Q&A data by institution and year

A Distribution of Data Sources

We collected 69,925 questions and answers from five major organizations and services specializing in adolescent counseling on NAVER Knowledge iN. Table 8 shows the data collection status by organization and year, along with the distribution of questions collected from 2011 to 2024.

B Data Preprocessing Details

The collected data was refined and pre-processed to ensure relevance. The following criteria were applied to remove misaligned data:

1. Non-adolescent questions: To exclude questions written by elementary school students or adults, we applied keyword-based filtering, resulting in the removal of 14,075 questions.
2. Inappropriate content: A total of 7,397 questions containing inappropriate sexual content were removed to maintain alignment with the research scope.
3. Lack of specificity: To eliminate vague questions that hinder meaningful analysis, 9,240 questions with 15 words or fewer in the detailed worry column were deleted.
4. Duplicate entries: To ensure data uniqueness and prevent redundancy, 2,089 duplicate questions were removed.

After applying the above criteria for pre-processing, 37,124 data points were selected and used in the study.

Round	Turn 1	Turn 2	Total (Cumulative %)
Round 1	17,694	1,841	17,694 (51%)
Round 2	6,132	361	6,493 (75%)
Round 3	4,240	57	4,297 (87%)
Round 4	1,243	37	1,280 (91%)
Round 5	784	4,735	5,519 (100%)

Table 9: Turn counts across negotiation rounds

C Changes in Cognitive Distortion Classification Through Negotiation

We analyze the distribution of data based on the number of negotiation rounds required to determine cognitive distortions. Table 9 presents the count of instances finalized at each round, illustrating how much data was classified early versus how much required additional rounds. The cumulative percentage represents the proportion of data for which cognitive distortion classification was completed at each round.

D Detailed Evaluation Results of LLM-Based Assessment

This appendix presents the detailed evaluation results of the LLM-based assessment for the two synthetic data generation methods: Cognitive Clarification and Cognitive Balancing. Each model's performance was assessed based on three criteria—Consistency, Accuracy, and Fluency—using independent evaluations by Gemini 1.5 Flash, GPT-4o mini, and Claude 3 Haiku, as shown in Table 10.

Generation Method	Generation Model	Evaluation model								
		Gemini-1.5-flash			GPT-4o mini			Claude-3-haiku		
		Cos	Acc	Flu	Cos	Acc	Flu	Cos	Acc	Flu
Cognitive Clarification	Gemini-1.5-flash	-	-	-	2.596	2.929	2.948	2.400	2.779	2.416
	GPT-4o mini	2.142	2.498	2.638	-	-	-	2.508	2.774	2.472
	Claude-3-haiku	2.150	2.519	2.643	2.606	2.754	2.814	-	-	-
Cognitive Balancing	Gemini-1.5-flash	-	-	-	2.253	2.718	2.740	2.134	2.589	2.298
	GPT-4o mini	1.882	2.090	2.515	-	-	-	2.111	2.333	2.312
	Claude-3-haiku	1.966	2.162	2.547	2.284	2.604	2.760	-	-	-

Table 10: Detailed evaluation results of synthetic data: Cos (Consistency), Acc (Accuracy), and Flu (Fluency)

Methodology	Model	Temperature	Max Tokens	Top-p
(A) Negotiation process	Gemini 1.5 Flash	0.5	1,024	0.9
	GPT-4o mini	0.5	1,024	0.9
	Claude-3 Haiku	0.5	1,024	0.9
(B) Synthetic data generation	Gemini 1.5 Flash	1.0	1,024	0.9
	GPT-4o mini	1.0	1,024	0.9
	Claude-3 Haiku	1.0	1,024	0.9
(C) Evaluation	Gemini 1.5 Flash	0.5	512	0.9
	GPT-4o mini	0.5	512	0.9
	Claude-3 Haiku	0.5	512	0.9

Table 11: Hyperparameters for negotiation process, synthetic data generation, and evaluation

E Hyperparameters of LLMs Models

We utilized Claude-3 Haiku, Gemini 1.5 Flash, and GPT-4o Mini at different stages of this study, summarizing the hyperparameters used at each step. Table 11(A) presents the hyperparameters for the negotiation process and independent evaluation, with Gemini 1.5 Flash and GPT-4o Mini used during the negotiation, and Claude-3 Haiku employed for independent evaluation. Table 11(B) summarizes the hyperparameters for synthetic data generation, while Table 11(C) outlines the hyperparameters used for evaluating synthetic data.

F Validating Synthetic Data with Clustering

To validate the synthetic data, we conducted clustering based on two criteria: (1) topics that elicit cognitive distortion in adolescents and (2) negative emotions and symptoms from the DSM-5.

To perform clustering, we first created mapping dictionaries for each criterion. Table 12 lists keywords for cognitive distortion topics in adolescents, and Table 13 shows DSM-5 depression symptom categories with related keywords.

We discovered keywords mapped to each category based on the topic-based mapping (Table 14) and DSM-5 symptom-based mapping (Table 15) of the synthetic data in the mapping dictionary. Only keywords with a mapping frequency of over 1,000 were selected, and the results were checked with the keyword in English, Korean, and the mapping frequency.

Topic (Korean, n)	List of Keywords (Korean)
Academic performance and career concerns (학업 성취도 및 진로, n=37)	Academics (학업), Academy (학원), Advancement to Higher Education (진학), Class (수업), Club (동아리), College Entrance Exam (입시), College Entrance Exam (수능), Discipline (규율, 생활지도), Dropping out (자퇴), English (영어), English Academy (영어학원), Enrollment (재학), Exam (시험), Extracurricular Activities (과외활동), Final Exam (기말고사), GED (검정고시), Grades (성적), Harass/Bully (괴롭히다), High School (고등학교), Homework (숙제), Interpersonal Relationships (인간관계), Math (수학), Middle School (중학교), Midterm Exam (중간고사), Mock Exam (모의고사), Private Tutoring (과외), Rank (등급), Report Card (성적표), Retaking the College Entrance Exam (재수), Scholarship (장학금), School (학교), School Life (학교생활), School Record (내신), Specialized High School (특성화고), Study (공부), Teacher (선생님), Timetable (시간표), Vocational School (실업계)
Friendships, romantic relationships, and interpersonal relationships (우정, 연애, 대인관계, n=24)	Acquaintance (지인), Best Friend (단짝), Boyfriend (남자친구), Boyfriend (남친), Break up (헤어지다), Bullying (왕따), Close Friend (친한친구), Confession (고백), Crush (짝사랑), Dating (사귀다), Exclusion/Ostracism (따돌림), Friend (친구), Friendship (친구사이), Girlfriend (여자친구), Heartbreak (실연, 마음의 상처), Jealousy (질투), Love (사랑), Loyalty (우정, 의리), Reconciliation (화해), Rumors (소문), Senior (선배), Trust (신뢰)
Physical and mental health (신체적, 정신적 건강, n=46)	Alone (혼자), Anxiety (불안), Appetite Loss (식욕 감퇴), Binge Eating Disorder (폭식증), Comfort (위로), Confidence (자신감), Concern (고민), Counseling (상담), Counselor (상담사), Depression (우울), Depression (우울증), Domestic Violence (가정폭력), Eating Disorder (섭식 장애), Emotional Wound (상처), Exercise (운동), Fatigue (피로), Guilt (죄책감), Headache (두통), Heart/Mind (마음), Inferiority Complex (열등감), Inner Self (내면), Insomnia (불면증), Loneliness (외로움), Mental Illness (정신병), Mental Strength (멘탈), Obesity (비만), Panic Disorder (공황장애), Psychiatry (정신과), Psychological Counseling (심리상담), Psychology (심리학), Psychotherapy (심리치료), Running Away (가출), School Bullying (학폭), School Violence (학교폭력), Self-esteem (자존감), Sleep Disorder (수면 장애), Stress (스트레스), Therapy (치료), Trust (신뢰), Unconscious Mind (무의식), Violence (폭력), Worry (걱정)
Family issues (가족 문제, n=23)	Dad (아빠), Divorce (이혼), Domestic Conflict (가정 불화), Domestic Violence (가정폭력), Estrangement (소원함), Family (가족), Family Breakdown (가족 해체), Family Conflict (가족 갈등), Family History (가족사), Father (아버지), Financial Issues (경제적 문제), Home/Family Environment (가정), Lack of Parental Support (부모의 무관심), Mom (엄마), Mother (어머니), Neglect (방임), Older Sister (누나), Older Sister (언니), Parents (부모님), Single-parent Family (한부모 가정), Younger Brother (남동생), Younger Sibling (동생), Younger Sister (여동생)
Appearance and self-image (외모 및 이미지, n=14)	Acne (여드름), Appearance (외모), Beauty Standards (외모 기준), Body Image (신체 이미지), Body Proportions (신체 비율), Body Shape (몸매), Bulking Up (벌크업), Diet (다이어트), Facial Features (얼굴 생김새), Height (키), Makeup (메이크업), Muscle (근육), Plastic Surgery (성형), Skin (피부)

Table 12: List of Keywords of Negative Emotion-Triggering Topics in Adolescents

DSM-5 Depression Symptom Class (Korean, n)	**List of Symptom or Emotion (Korean)**
A. Depressed mood (우울한 기분, n=24)	Crying (울다), Depression (우울, 우울증), Despair (절망), Disappointment (실망), Emptiness (공허, 허탈), Frustration (좌절), Guilt (죄책감), Hard (힘들다), Heartache (상심), Hopelessness (무기력, 희망 없음), Loss (상실), Pain (고통), Sad (슬프다), Scared (무서운, 겁나는), Suffering (괴로움), Tough (힘들), Unhappiness (불행), Upset (화나다, 속상함), Worthlessness (무가치함)
B. Loss of interest/pleasure (흥미 또는 즐거움의 상실, n=14)	Alienated (소외), Alone (혼자, 홀로), Apathy (냉담), Bore (지루함), Bullying (따돌림), Unpleasant (불쾌함), Ignored (무시), Indifference (무관심), Isolated (고립), Loneliness (외로움), Lonely (외로워), Meaningless (무의미함), Disinterest (흥미 없음)
C. Weight loss or gain (체중 감소 또는 증가, n=11)	Appetite (식욕), Binge Eating (폭식), Body (몸매), Diet (다이어트, 식단), Fat (살찌다), Loss of Appetite (식욕 감퇴), Nausea (매스꺼움), Overweight (과체중), Underweight (저체중), Weight (체중)
D. Insomnia or hypersomnia (불면증 또는 과다수면, n=15)	Daytime Fatigue (주간 피로), Hypersomnia (과다수면), Insomnia (불면, 불면증), Restless Sleep (뒤척임), Sleep (수면, 잠), Sleep Deprivation (수면 부족), Sleep Disorder (수면장애), Sleep Patterns (수면 패턴), Sleepiness (졸음), Sleeping Pills (수면제), Stress (스트레스), Worry (고민, 걱정)
E. Psychomotor agitation or retardation (정신운동 초조 또는 지연, n=26)	Anger (분노), Anger management (분노 조절, 분노 관리), Anxiety (불안, 불안감), Anxiety disorder (불안 장애), Irritability (과민, 과민성, 짜증), Nervousness (초조, 신경질), Obsessive (강박증), Obsessive-compulsive disorder (강박장애), Sensitive (예민, 예민한), Tension (긴장, 긴장감), Restlessness (안절부절), Fidgeting (꼼지락거림, 안절부절못함), Agitation (초조, 불안, 동요), Impulsivity (충동성, 충동적 행동), Hyperactivity (과잉행동), Slow movement (느린 동작, 둔한 행동)
F. Fatigue (피로감, n=19)	Dejected (낙담, 허탈), Empty (공허), Exhausted (지치다, 탈진), Fatigued (피로), Helpless (무력감), Incompetence (무능, 능력 부족), Inferiority (열등감, 자신감 부족), Lethargy (무기력, 무기력증), Powerless (무기력한, 힘이 없는), Sleepiness(졸음), Sleepy (졸린), Tired (피곤, 피곤함)
G. Inappropriate guilt (부적절한 죄책감, n=11)	Guilt (죄책감), Helplessness (무력감), Incompetence (무능, 무능함, 무능력), Inferiority (열등감), Regret (후회), Self-blame (자책), Shame (수치, 창피, 수치심)
H. Decreased concentration (집중력 저하, n=13)	Concentration (집중, 집중력), Concern (염려, 우려, 고민), Confusion (혼란), Distracted (산만함, 주의 산만), Discomfort/Inconvenience (불편, 불편함), Forgetfulness (건망증), Judgment (판단), Worry (걱정)
I. Thoughts of suicide (자살 사고, n=10)	Death (죽음), Desperation (절박함, 절망), Die (죽다), Fear (두려움), Panic Disorder (공황장애), Self-harm (자해), Suicide (자살), Suicidal Ideation (자살 충동, 자살 사고)

Table 13: DSM-5 Depression Symptom related Classes and Keywords

Topic	Keyword_KR	Keyword_ENG	Count
Relationship	친구	Friend	64,041
Academic	학교	School	21,298
Mental Health	혼자	Alone	18,482
Mental Health	걱정	Worry	16,636
Academic	시험	Exam	13,168
Academic	선생님	Teacher	11,163
Academic	공부	Study	9,987
Mental Health	불안	Anxiety	9,716
Academic	성적	Grades	9,407
Family	엄마	Mom	8,458
Mental Health	마음	Heart/Mind	7,543
Academic	수업	Class	6,987
Family	부모님	Parents	5,121
Mental Health	고민	Concern	3,966
Academic	고등학교	High School	3,725
Academic	학원	Academy	3,409
Academic	수학	Math	3,347
Mental Health	우울	Depression	2,927
Appearance	키	Height	2,690
Family	아빠	Dad	2,312
Academic	중학교	Middle School	2,065
Academic	숙제	Homework	1,992
Mental Health	스트레스	Stress	1,798
Mental Health	운동	Exercise	1,639
Family	가족	Family	1,636
Mental Health	상담	Counseling	1,619
Academic	자퇴	Dropping out	1,458
Academic	성적표	Report Card	1,343
Academic	진학	Advancement to Higher Education	1,241
Relationship	화해	Reconciliation	1,148
Mental Health	상처	Emotional Wound	1,125
Relationship	고백	Confession	1,113
Relationship	선배	Senior	1,086
Relationship	사랑	Love	1,025

Table 14: Frequency Distribution of 34 Keywords Across Topics

Topic	Keyword_KR	Keyword_ENG	Count
A. Depressed mood	실망	Disappointment	3,722
A. Depressed mood	우울	Depression	2,927
A. Depressed mood	힘들	Struggling	13,694
A. Depressed mood	힘들다	Difficult	1,332
B. Loss of interest/pleasure	무시	Neglect	8,226
B. Loss of interest/pleasure	소외	Alienation	2,609
B. Loss of interest/pleasure	혼자	Alone	18,482
D. Insomnia or hypersomnia	걱정	Worry	16,636
D. Insomnia or hypersomnia	고민	Concern	3,966
D. Insomnia or hypersomnia	스트레스	Stress	1,798
D. Insomnia or hypersomnia	잠	Nap	2,357
E. Psychomotor agitation or retardation	불안	Anxiety	9,716
E. Psychomotor agitation or retardation	짜증	Annoyance	2,868
F. Fatigue	무능	Inability	2,332
G. Inappropriate guilt	무능	Incompetence	2,332
G. Inappropriate guilt	후회	Regret	1,451
H. Decreased concentration	걱정	Worry	16,636
H. Decreased concentration	고민	Concern	3,966
H. Decreased concentration	불편	Discomfort	1,959
H. Decreased concentration	집중	Focus	1,971

Table 15: Frequencies of 20 keywords across DSM-5 symptom categories

G Examples of Synthetic Data in KoACD

In Table 16, we provide one synthetic example per cognitive distortion, totaling 10.

DSM-5 Depression Symptom Class (Korean, n)	List of Symptom or Emotion (Korean)
All-or-Nothing Thinking (흑백사고)	*"I'm really bad at studying, and my grades are at the bottom. I can't even think about college, and getting a job seems impossible too. It feels like I have no future, like I've completely failed."* *(공부를 전혀 못해서 성적이 바닥이에요. 대학은 엄두도 못 내겠지만, 취업도 어렵겠죠. 결국 제대로 된 미래가 없을 것 같아 완전히 실패한 것 같아요.)*
Overgeneralization (과잉일반화)	*"I feel anxious because it seems like my classmates avoid talking to me. One day, I felt so left out that I cried. There have been so many times when everyone gathered and left me out. Now, I'm scared of being alone."* *(수업 시 친구들이 나와의 이야기를 피하는 것 같아 불안해. 하루는 소외된 기분이 들어 울었어. 모두가 모여서 나를 제외하고 나선 적이 많아, 이제 혼자가 될까 두려워.)*
Mental Filtering (부정적 편향)	*"I got my math test results—80 out of 100. It's over. Not even an A, and the top spot in the school is out of reach. Everything's ruined. My future's looking dark. I won't get into college. I won't be able to do anything. Maybe I should just give up."* *(수학 시험 성적표를 받았다. 80 점. 망했다. A 등급은 커녕, 전교 1 등은 물 건너갔다. 모든 게 끝장났다. 내 미래는 어둡다. 대학도 못 갈 거야. 아무것도 안 될 거야. 그냥 포기해야겠다.)*
Discounting the Positive (긍정 축소화)	*"I wasn't good at studying in middle school, but this time I finally got a score in the 60s. Instead of being happy for me, my parents got mad and said, 'Is that something to brag about?' It really hurt because it felt like all my effort didn't matter."* *(내가 중학교 때는 공부를 잘 못했었는데, 이번에 겨우 60 점대 맞았다고 자랑이냐며 부모님께서 화내셨어요. 노력한 게 인정받지 못하는 것 같아 너무 속상했습니다.)*
Jumping to Conclusions (성급한 판단)	*"There's a girl I like at my academy. I want to talk to her, but I'm scared I might get rejected or even end up being an outcast. If I confess and she's not interested, it'll hurt even more, so I keep hesitating. In the end, I'm just doing nothing."* *(학원에서 좋아하는 여자애가 있어요. 말을 걸고 싶지만, 왕따 당할까봐 두려워요. 고백하면 만약 저에게 관심이 없다면 더 힘들 것 같아서 늘 망설여져요. 결국 아무것도 못하고 있네요.)*
Magnification and Minimization (확대와 축소)	*"I'm having a hard time adjusting to my new school after transferring. I don't have any friends, and keeping up with the lessons is tough, so I'm getting more and more depressed. I'm worried that if I stay isolated like this, I'll never be happy."* *(전학 와서 새로운 학교에 잘 적응하지 못하고 있어요. 친구도 없고 공부도 따라가기 힘들어서 점점 우울해지고 있어요. 이렇게 고립되면 평생 행복할 수 없을 것 같아 걱정돼요.)*
Emotional Reasoning (감정적 추론)	*"I had an argument with my mom. I don't think I can take this conflict anymore. If I go to school, I feel like I'll just keep crying, but if I don't go, there'll be an even bigger fight. I honestly don't know what to do."* *(엄마와 싸웠어요. 더 이상 갈등을 견딜 수 없을 것 같아요. 학교에 가면 계속 울고 있을 것 같고, 학교에 가지 않으면 더 큰 싸움이 벌어질 거예요. 과연 어떻게 해야 할지 모르겠어요.)*
"Should" Statements (“해야한다” 진술)	*"I know I should study hard for this exam, but it feels so tough every time I try, and I just want to give up. But I know I can't, I have to keep going and work hard to get good grades."* *(이번 시험 준비를 잘해야 할 텐데, 공부할 때마다 너무 힘들어서 포기하고 싶어진다. 하지만 이렇게 해서는 안 되고, 반드시 열심히 공부해서 좋은 성적을 받아야 한다.)*
Labeling (낙인찍기)	*"I'm probably a loser because my test scores are bad. My friends will avoid me, and I'll end up being a loner in high school too. I'm so clueless that I won't be able to make any friends. I have no idea how I'm supposed to keep going."* *(시험 성적이 좋지 않은 내가 찐따일 거야. 친구들도 나를 피할 거고, 고등학교에서도 외톨이가 될 것 같아. 눈치 없는 나는 친구를 사귈 수 없을 거야. 앞으로 어떻게 살아갈지 막막하다.)*
Personalization (개인화)	*"I feel like my friends don't like me. I joined a new club, but they're leaving me out. I don't even know what I did wrong. Even if I made a mistake, they shouldn't treat me like this."* *(나는 친구들이 나를 싫어하는 것 같아. 새로운 동아리에 들어갔는데, 친구들이 나를 배제하고 있어. 내가 뭘 잘못했는지 모르겠어. 설령 내가 실수했더라도 이렇게 대할 순 없잖아.)*

Table 16: Examples of Synthetic Data for Each Cognitive Distortion in KoACD

H Prompt Templates

We present the prompt templates used throughout the study for various stages of cognitive distortion identification, synthetic data generation, and evaluation. These prompts were designed to ensure consistency and accuracy across different processes.

To maintain conciseness, we replaced detailed descriptions and examples of cognitive distortions with the phrase 'Refer to Table 2 for a detailed explanation of each cognitive distortion.' This appendix includes Tables 17, 18, 19, 20, 21, 22, 23, and 24, which provide the full prompt templates for each stage.

Analyzer

You are a psychology expert.
Analyze the text below and, if a relevant cognitive distortion is present, select the most appropriate one.
Choose from the following ten cognitive distortions: All-or-Nothing Thinking, Overgeneralization, Mental Filter, Discounting the Positive, Jumping to Conclusions, Magnification and Minimization, Emotional Reasoning, Should Statements, Labeling, and Personalization.
{previous_cognitive_distortions} were deemed inappropriate in the previous analysis. Do not select them again under any circumstances.

Previously rejected cognitive distortions: {previous_cognitive_distortions}
Reason for rejection: {previous_reasons}

Since {previous_cognitive_distortions} were already deemed inappropriate:

1. Do not select any of the above cognitive distortions again.
2. You must choose only from the remaining cognitive distortions.
3. If none of the remaining cognitive distortions are appropriate, respond with "Unknown."

When identifying cognitive distortions, carefully refer to the definitions and examples of the ten distortions to consider a variety of cognitive distortions.
When deciding on a cognitive distortion, analyze the overall context of the text rather than focusing on a single sentence.

Criteria for Responding with "Unknown":
- The response requires speculation or subjective interpretation.
- The intent of the sentence is unclear.
- The speaker is not explicitly identified.
- The text consists only of simple emotional expressions.
- The text is merely a description of a situation or a question.
- Context from prior conversations is necessary for understanding.
- The text lacks value judgments or personal interpretation.
- The meaning is unclear without external context.
- The experience is described from another person's perspective.
- Negative emotions are present, but no specific cognitive distortion is identifiable.
- The text is a request for information, advice, or help.

Important: If you determine "Unknown," this is a final decision, and no further analysis or reconsideration is needed. If any of the above criteria apply, immediately respond with "Unknown" without considering alternative interpretations.

Text to Analyze:
{input_text}

List of Cognitive Distortions:
Refer to Table 2 for a detailed explanation of each cognitive distortion.

Analysis Request:

1. When determining cognitive distortions, consider the overall context.
2. Copy and paste all relevant sentences or paragraphs that support the selected cognitive distortion. Include at least two complete sentences.
3. Provide a clear explanation for selecting the sentences, ensuring a logical cause-and-effect relationship in your reasoning.
4. If no cognitive distortion applies, respond with "Unknown."

Output Format:
- Cognitive Distortion: [Selected Cognitive Distortion]
- Relevant Sentences/Paragraphs: [Text]
- Reason for Selection: [Explanation]

Additional Output Rules:
- All responses must be grammatically complete sentences.
- Sentences should not be cut off mid-thought.
- The final sentence of the response must be fully structured and complete.
- Do not use Markdown formatting.
- When outputting [Selected Cognitive Distortion], do not select any distortions from {previous_cognitive_distortions}.

Table 17: Prompt for the analyzer role in the negotiation process

Evaluator

You are a psychology expert.
Strictly evaluate the following cognitive distortion analysis provided by the analyzer.
Refer to the cognitive distortions list for definitions and examples.

Original Text:
{input_text}

Analyzer's Assessment:
Cognitive Distortion: {cognitive_distortions}
Relevant Sentences/Paragraphs: {related_text}
Reason for Selection: {reason_text}

List of Cognitive Distortions:
Refer to Table 2 for a detailed explanation of each cognitive distortion.

Evaluation Rules:
1. Is the selected cognitive distortion present in the text?
- Assess whether the identified cognitive distortion can be reasonably inferred from the original text.
- Do not rely on isolated sentences; patterns must be found within the overall flow of the text.
2. Do the selected relevant sentences and reasoning properly support the cognitive distortion?
- Check whether the selected sentences accurately align with the definition and examples of the cognitive distortion.
- Evaluate whether the explanation logically connects the chosen sentences to the cognitive distortion.
- Ensure that the justification is not overly interpretative or speculative.

Judgment Criteria:
- If any of the evaluation rules are violated, classify the analysis as "Inappropriate."
- If deemed inappropriate, clearly specify which rule was violated.
- If the response is "Unknown," accept it immediately.

Output Format:
Evaluation Result: [Appropriate / Inappropriate]
Evaluation Reason: [Detailed explanation for each rule]
Conclusion:
[If appropriate] "The current analysis is valid."
[If inappropriate] "The cognitive distortion should be reassessed."

Additional Output Rules:
- The evaluation reason must be fully structured in grammatically complete sentences.
- Sentences should not be cut off mid-thought.
- The final sentence must be a fully completed statement.
- Do not use Markdown formatting.

Table 18: Prompt for the evaluator role in the negotiation process

Independent Evaluator
You are a psychology expert. Thoroughly evaluate the appropriateness of the extracted cognitive distortion and its associated sentences/paragraphs. **Content to Evaluate:** Selected Cognitive Distortion: {selected_cognitive_distortion} Relevant Sentences: {related_sentences} **Evaluation Criteria (1-3 points):** 1 Point: Inappropriate - The relevant sentences do not contain the identified cognitive distortion. - OR the sentences are incomplete or lack clear context. 2 Points: Partially Appropriate - The relevant sentences contain a cognitive distortion, but it does not match the selected one. - OR another cognitive distortion would be a better fit. 3 Points: Appropriate - The relevant sentences clearly demonstrate the selected cognitive distortion. - The content aligns well with the definition and examples of the cognitive distortion. **Output Format:** Score: [1-3 points] **Important Notes**: - Only scores of 1, 2, or 3 may be used. - Intermediate scores (e.g., 1.5 or 2.5) are not allowed. - The evaluation rationale must be consistent with the assigned score.

Table 19: Prompt for independently verifying cognitive distortion

Cognitive Clarification Method

Generate a realistic fictional adolescent story based on the given cognitive distortion and reference case.
When writing the fictional story, ensure that the age and content remain within the adolescent range.
Consider a variety of situations that may occur both inside and outside of school.
Strictly follow the output format specified below.

Input Information:
Cognitive Distortion: {cognitive_distortions}
Relevant Real-Life Sentence/Paragraph: {example_text}
Original Text: {input_text}

Story Writing Requirements:
1. Length: Must be 40 words or fewer (Exceeding 40 words is strictly prohibited).
2. Format: [Gender/Age] --- [Story Content]
3. Age: Must be between 13 and 19 years old.
If gender, age, or school grade is mentioned in the original text, use that information to generate [Gender/Age].
(Gender: Male or Female, Middle School: 14-16 years old, High School: 17-19 years old)
4. Theme: Events that occur in school, home, friendships, or daily adolescent life.
5. Perspective: Write from a first-person point of view.
6. Content:
Clearly establish the situation (when, where, what, how).
Maintain a logical cause-and-effect relationship within the story.
The narrator (first-person) should naturally exhibit cognitive distortion.

Constraints:
1. The story must be inspired by the given real-life sentence, adapting it to a similar but new context.
2. Utilize grammatical transformations, such as active/passive voice changes and word order modifications.
3. Avoid starting the story with any of the following words:{used_words}
4. The word "today" must not be used.
5. Do not explicitly mention cognitive distortion terms in the story.
(e.g., Do NOT use terms like "overgeneralization" or "all-or-nothing thinking.")

Output Format:
[Gender/Age] --- [Generated Story]

Important Notes:
Ensure that the cognitive distortion characteristics reflected in the reference sentence are incorporated into the new story in a different yet relevant context.
Do NOT exceed 40 words in the generated story (Strict limit: 40 words maximum).

Table 20: Prompt for cognitive clarification-based synthetic story generation

Cognitive Balancing Method

Generate a realistic fictional adolescent story that reflects the characteristics of {cognitive_distortions}, based on the provided real-life example.
Strictly follow the output format specified below.

Input Information:
Original Text: {input_text}

Story Writing Requirements:
1. Length: Must be 40 words or fewer (Exceeding 40 words is strictly prohibited).
2. Format: [Gender/Age] --- [Story Content]
3. Age: Must be between 13 and 19 years old.
If gender, age, or school grade is mentioned in the original text, use that information to generate [Gender/Age].
(Gender: Male or Female, Middle School: 14-16 years old, High School: 17-19 years old)
4. Theme: Events that occur in school, home, friendships, or daily adolescent life.
5. Perspective: Write from a first-person point of view.
6. Content:
Clearly establish the situation (when, where, what, how).
Maintain a logical cause-and-effect relationship within the story.
The narrator (first-person) should naturally exhibit cognitive distortion.

List of Cognitive Distortions:
Refer to Table 2 for a detailed explanation of each cognitive distortion.

Current Cognitive Distortion for Story Generation:
{cognitive_distortions}

Constraints:
1. The story must be inspired by the given real-life sentence, adapting it to a similar but new context.
2. Utilize grammatical transformations, such as active/passive voice changes and word order modifications.
3. Avoid starting the story with any of the following words:{used_words}
4. The word "today" must not be used.
5. Do not explicitly mention cognitive distortion terms in the story.
(e.g., Do NOT use terms like "overgeneralization" or "all-or-nothing thinking.")

Output Format:
[Gender/Age] --- [{cognitive_distortions} Reflected Story]

Important Notes:
Do NOT exceed 40 words in the generated story (Strict limit: 40 words maximum).

Table 21: Prompt for cognitive balancing-based synthetic story generation

Consistency Evaluator

Please evaluate the consistency of the given text and assign a single score between 1 and 3.

Input Information:
Selected Cognitive Distortion: {selected_cognitive_distortion}
Relevant Sentences: {related_sentences}
Generated Story: {generated_story}

Evaluation Criteria (1-3 points):
Assess whether the selected cognitive distortion is neither exaggerated nor minimized and whether the original meaning of the relevant sentences is preserved while being appropriately expressed in the generated story.
- Is the selected cognitive distortion accurately maintained without distortion from the relevant sentences?
- Has the meaning of the relevant sentences been appropriately conveyed in the generated story without excessive modification?
- Does the generated story logically align with the selected cognitive distortion and its context?

Scoring Guidelines:
1 Point: The selected cognitive distortion or the context of the relevant sentences is significantly distorted or altered in the generated story.
2 Points: The selected cognitive distortion and the context of the relevant sentences are partially retained, but there are some inconsistencies or unnatural expressions.
3 Points: The selected cognitive distortion and the context of the relevant sentences are naturally maintained, forming a logically coherent story.

Output Format:
Score: [1-3 points]

Important Notes:
- Only scores of 1, 2, or 3 may be used.
- Intermediate scores (e.g., 1.5 or 2.5) are not allowed.
- The evaluation rationale must be consistent with the assigned score.

Table 22: Prompt for consistency evaluation of synthetic data

Accuracy Evaluator
Please evaluate the accuracy of the given text and assign a single score between 1 and 3. **Input Information:** Selected Cognitive Distortion: {selected_cognitive_distortion} Relevant Sentences: {related_sentences} Generated Story: {generated_story} **Evaluation Criteria (1-3 points):** Evaluate whether the generated story is correctly classified under the most relevant cognitive distortion among the ten defined categories. - Does the selected cognitive distortion correctly classify the cognitive distortion present in both the relevant sentences and the generated story? - When compared to other cognitive distortions, is the selected cognitive distortion the most appropriate choice? - Is there a logical consistency between the selected cognitive distortion and the way it is expressed in the generated story? **Scoring Guidelines:** 1 Point: The selected cognitive distortion significantly mismatches the cognitive distortion found in the relevant sentences and the generated story. 2 Points: The selected cognitive distortion is partially appropriate, but another cognitive distortion might be a better fit. 3 Points: The selected cognitive distortion is the most accurate classification of the cognitive distortion found in the relevant sentences and the generated story. **Output Format:** Score: [1-3 points] **Important Notes**: - Only scores of 1, 2, or 3 may be used. - Intermediate scores (e.g., 1.5 or 2.5) are not allowed. - The evaluation rationale must be consistent with the assigned score.

Table 23: Prompt for accuracy evaluation of synthetic data

Fluency Evaluator

Please evaluate the fluency of the given text and assign a single score between 1 and 3.

Input Information:
Generated Story: {generated_story}

Evaluation Criteria (1-3 points):
Evaluate whether the generated story is grammatically sound and maintains human-like fluency in its structure and readability.
- Is the sentence structure natural and fluent?
- Are there any grammatical errors?
- Is the flow between sentences smooth, making the overall story cohesive?

Scoring Guidelines:
1 Point: The text contains many grammatical errors or is highly unnatural.
2 Points: The text has minor grammatical issues or slightly awkward expressions but is still generally understandable.
3 Points: The text is grammatically correct and reads naturally with a smooth sentence structure.

Output Format:
Score: [1-3 points]

Important Notes:
- Only scores of 1, 2, or 3 may be used.
- Intermediate scores (e.g., 1.5 or 2.5) are not allowed.
- The evaluation rationale must be consistent with the assigned score.

Table 24: Prompt for fluency evaluation of synthetic data