KoACD：首个用于认知扭曲分析的韩国青少年数据集

JunSeo Kim
Department of Computer Engineering
Gachon University College of IT Convergence
Seongnam, 13120
kma80kjs@gachon.ac.kr

HyeHyeon Kim
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
Seoul, 03722
hye_hyeon@yonsei.ac.kr

2025 年 6 月 1 日

ABSTRACT

认知扭曲是指可能导致青少年出现抑郁症和焦虑等心理健康问题的消极思维模式。以往使用自然语言处理（NLP）的研究主要集中在小规模成人数据集上，对青少年的研究有限。本研究介绍了 KoACD，这是第一个大规模的韩国青少年认知扭曲数据集，包含 108,717 个实例。我们应用了多大型语言模型（LLM）谈判方法来细化扭曲分类，并采用两种方式生成合成数据：通过认知澄清实现文本清晰度和通过认知平衡实现多样化扭曲表示。通过 LLM 和专家评估验证表明，虽然 LLM 能够对带有显式标记的扭曲进行分类，但在依赖上下文的推理方面表现不佳，而人类评估者展示了更高的准确性。KoACD 旨在提高未来关于认知扭曲检测的研究水平。

1 介绍

消极想法 [1] 是人类认知的自然组成部分，通常有助于个体识别潜在危险、准备应对挑战或进行自我反思。然而，当这些模式变得僵化和过度时，它们会导致情绪困扰，并导致心理健康问题，如抑郁症和焦虑症。特别是青少年，由于他们正在进行的认知和情感发展，可能更容易受到这些适应不良思维模式的影响。

全球每七个儿童和青少年（约 1.66 亿）患有精神疾病，其中 42.9%经历焦虑和抑郁 [2]。这些状况在青春期的患病率上升已成为一个严重的全球性问题。由于这一阶段对自我认同形成和情绪调节至关重要 [3]，理解负面思维模式如何出现并持续对于早期干预和预防是必不可少的。

尤其，抑郁症通常与习惯性的负面思维有关，这种现象被称为认知扭曲 [4]。经历这些扭曲的青少年在事情出错时可能会本能地责怪自己，心想，“我又搞砸了”或“我完全失败了。”这些持续的消极思想会触发情绪困扰，强化抑郁循环 [5]。识别和分析这些模式对于开发有效的应对策略至关重要。为此，构建一个全面的青少年认知扭曲数据集是必要的，以实现更针对性且影响深远的心理健康干预措施。KoACD 数据集可以在 https://github.com/cocoboldongle/KoACD 公开获取。

2 相关工作

2.1 认知扭曲检测

认知扭曲与消极思维模式密切相关，被定义为强化消极情绪的扭曲思维方式 [6]。随着自然语言处理（NLP）的最近发展，研究者们积极开展了利用各种数据集自动分类认知扭曲的研究。

之前对认知扭曲分类的研究主要依赖于小规模的、以成人为主的和英语数据集。早期研究利用了基于 LIWC 的回归模型来分析社交媒体帖子 [7]，而后来的研究则采用了深度学习模型，包括 RNNs、CNNs 和 BERT，使用来自咨询平台和治疗师-患者对话的数据集 [8, 9]。最近的方法利用大型语言模型（LLMs）进行认知扭曲分类，进一步提升了在不同数据集上的性能和适应性 [10, 11]。

表 1 总结了处理认知扭曲的现有数据集。

尽管取得了这些进展，现有的数据集在规模上仍然有限，并且主要集中在说英语的成人身上。为了解决这些问题，我们提出一个专门针对青少年的韩语数据集，填补了年轻人群中认知扭曲研究中的一个重要空白。

表 1: 认知扭曲检测的数据集总结。“样本”列表示实例的数量，“分类”指定类型（二元、多类或多标签）。* 表示非官方命名的数据集。

数据集	语言	样本	目标	数据源	分类
Tumblr					
Cognitive Distortion* [7]	English	459	nonspecific	Tumblr blogs	Binary (2)
MH-C [8]	English	1,164	Adult	TAO Connect	Multi-class (15)
MH-D [8]	English	1,799	Adult	TAO Connect	Binary (2)
CrowdDist [8]	English	7,666	Adult	Mechanical Turk	Multi-class (15)
Clinician-Client					
SMS* [9]	English	7,354	Adult	Clinician-Client SMS	Multi-class (5)
SocialCD-3K [11]	Chinese	3,407	nonspecific	Weibo	Multi-label (12)

2.2 基于 LLM 的谈判

已经尝试探索通过大型语言模型之间的互动和协商超越独立判断的模型的可能性。

自我对弈和上下文学习技术利用 AI 反馈被应用于提高大型语言模型 [12] 的谈判能力，并提出了一种通过开发基于 LLM 的教练助手（ACE）使用 MBA 学生谈判数据提供反馈的方法 [13]。此外，还进行了将谈判方法应用于情感分析的研究。已经证明，使用 LLM 谈判方法在模型之间进行交互可以超越现有的单次决策 [14]。

先前的研究表明，大型语言模型谈判可以产生复杂的结果，但在平衡谈判结果方面仍面临挑战，原因是固定的角色和有限的结构。为克服这一问题，我们采用角色互换和多种类型的大型语言模型来平衡谈判过程。我们还引入多轮谈判以平等考虑 10 种认知扭曲，最终得出最优结论。因此，本研究旨在使用大型语言模型谈判技术生成和验证数据。

3 构建 KoACD

3.1 数据源和预处理

我们爬取了 NAVER 知识 iN 上的帖子，这是一个用户可以提问并获得答案的问答平台，以分析韩国青少年的认知扭曲。NAVER 知识 iN 是韩国最大的开放问答平台，自 [15] 推出以来已有超过 3200 万用户和 8 亿条问

答条目。由于 NAVER 知识 iN 涵盖了广泛的年龄组，我们仅使用了五大青少年咨询组织和服务的数据，涵盖 2011 年至 2024 年的时间段，以专注于青少年的问题。

一个预处理步骤对数据进行了精炼，使其符合研究目的，并排除了无关的问题。这包括删除来自小学生或成人的条目，过滤不适当的内容，删除模糊的问题，并消除重复项。应用这些标准后，剩下 37,124 个问题用于分析。

3.2 认知扭曲的定义

Aaron Beck，认知疗法的先驱，识别了抑郁患者中的 10 种认知扭曲，并将它们纳入心理治疗 [6]。他强调减少这些扭曲可以减轻压力和焦虑 [16]。我们使用表 2 中列出的这些扭曲来分类反映青少年常报告的情感斗争的问题。

表 2: 认知扭曲的分类及其定义和示例。

认知扭曲类型	定义	示例
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 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."

3.3 多语言模型谈判以识别认知扭曲

为了有效识别认知扭曲，本研究设计了一个过程，使用多语言模型谈判方法来推导最优扭曲，其中相关扭曲通过语言模型的交互 [12] 逐渐被推导出来。

本研究采用基于谷歌的 Gemini 1.5 Flash[17] 和 OpenAI 的 GPT-4o mini[18] 之间互动的多 LLM 谈判方法。两个模型协同工作以识别最准确的认知扭曲。一个模型充当分析器，另一个则作为评估者。通过它们的合作，认

知扭曲逐渐得到细化。谈判进行多达五轮，系统地探索所有十个预定义的认知扭曲，因为每轮包含两回合，每一回合评估一种扭曲。这种结构允许同一句子在最终分类之前可能被解释为多种认知扭曲。

以下是角色的工作原理：

- **分析器**： 识别句子中最相关的认知扭曲，并建议与其匹配的句子。
- **评估器**：审查分析器提出的建议并对其准确性提供反馈。

在每一轮谈判中，模型轮流扮演分析者和评估者的角色。

一轮由两个回合组成，其结构如下：

- **T1（初始分析）**：识别句子中最具相关性的认知扭曲（选项：十种认知扭曲之一）。
- **T1（评估）**：评估从 T1（初步分析）提出的认知扭曲是否准确反映了句子中存在的扭曲（选项："是"或"否"；评价者提供理由）。
- **T2（再分析）**：如果 T1（评估）导致拒绝，选择下一个最相关的认知扭曲，排除之前被拒绝的选项（选项：剩余的认知扭曲之一）。
- **T2（评估）**：确定来自 T2（重新分析）的认知扭曲是否适当（选项："是" 或 "否"；评估者提供理由）。

每个步骤依次进行，结合上一步评估的反馈。T1（评估）评估 T1（初始分析）中提出的失真，并且 T2（重新分析）根据该反馈细化选择。类似地，T2（评估）验证在 T2（重新分析）中选定的失真的适用性。

在整个谈判过程中，被认为不适当的认知扭曲会被系统地排除，以确保选择最合适的认知扭曲。为了保持公平，在 T2（再分析）阶段模型会交替角色，以便双方都能平等贡献于谈判。

如果在五个回合后仍未达成共识，则该问题被归类为未知。这表明在谈判过程中提出的所有认知扭曲都被认为本质上是不适当的。

识别认知扭曲所需的问题轮次或将其分类为未知的情况在不同数据集之间有所不同。有些句子可以很快得出结论，而另一些则需要多轮才能最终分类。此谈判过程的整体结构如图 1 所示。

3.4 独立评估

谈判过程完成后，Anthropic 的 Claude 3 Haiku[19] 用于对最终的认知扭曲及其相应的句子进行独立验证。Claude 3 Haiku 不参与谈判过程；相反，它评估所选认知扭曲是否与给定的句子正确一致。

在谈判过程中，模型会在整个原始文本的背景下评估认知扭曲，而 Claude 3 Haiku 仅根据选定的句子来判断其适当性。这一额外的验证步骤有助于识别潜在的误分类，并确保认知扭曲与句子正确关联。

Claude 3 Haiku 赋予了一个从 1 到 3 的相关性评分，只有获得 3 分的认知扭曲-句子配对被用作生成合成数据的最终阶段数据。验证分数分布的总结见表 3。

表 3: 验证分数的分布。

得分	计数	比例（%）
1	11	0.06
2	874	4.41
3	18,897	95.53
Total	19,782	100.00

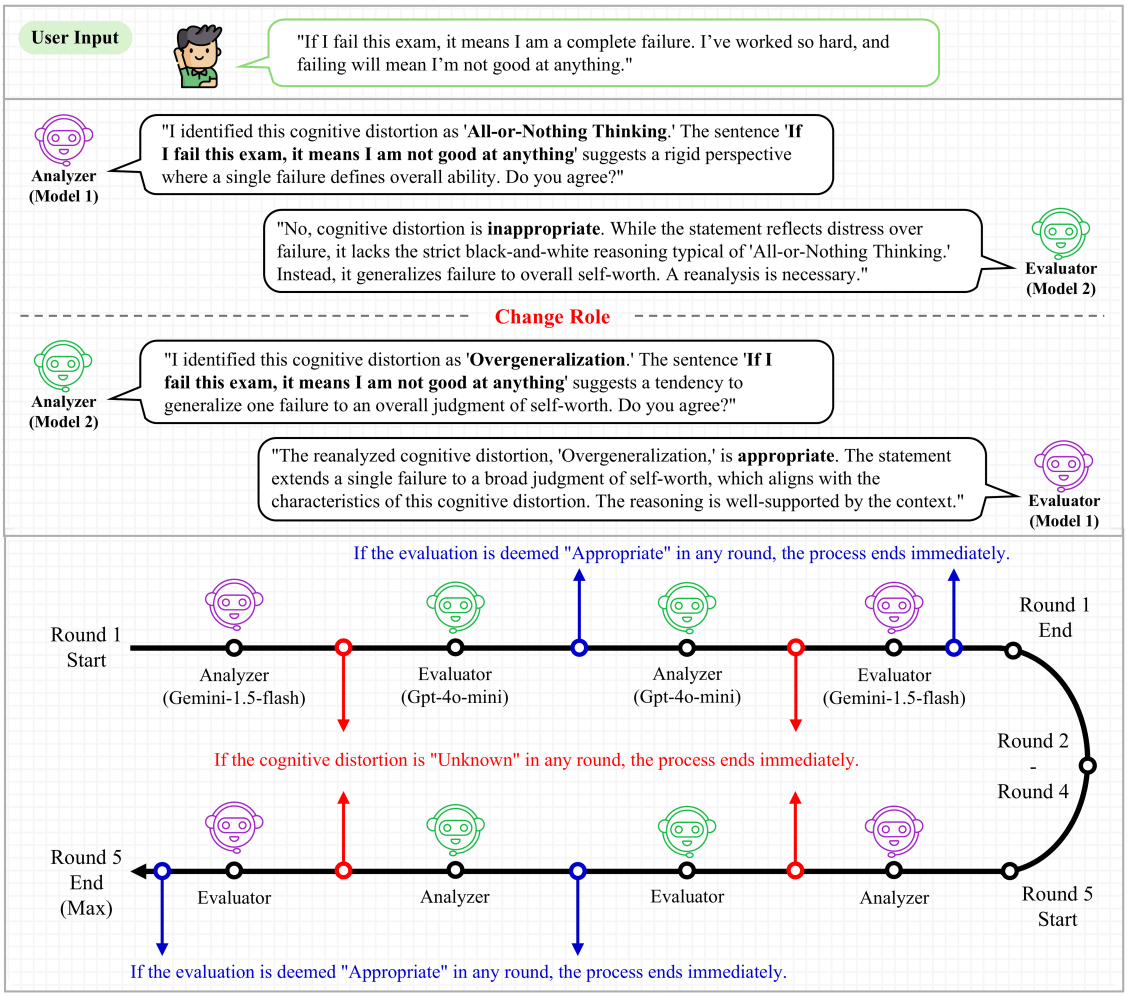

图 1: 通过谈判识别和评估认知扭曲的过程。

3.5 合成数据生成

原始数据由青少年撰写的自由文本组成，通常包含拼写错误、过度使用表情符号或措辞不清，使其难以解释。此外，某些文本缺乏上下文连贯性，叙述不连贯或背景信息不足，无法准确评估认知扭曲。因此，这些数据可能很难直接使用。

此外，我们提出的十个认知扭曲类别分布不平衡，导致数据集可能存在偏见。为了解决这些问题，我们采用两种方法生成合成数据。

3.5.1 认知扭曲的认知澄清

生成合成数据的第一种方法是识别认知扭曲，并以更清晰、更有结构的形式重新表述文本，同时保留原始文本的意义，保持情感基调和上下文。

我们使用了三个大型语言模型——Gemini 1.5 Flash、Claude 3 Haiku 和 GPT-4o mini——独立生成了来自 18,897 个样本的各种表达，确保生成内容具有更大的多样性。

3.5.2 平衡认知扭曲与保持上下文的数据

第二种方法旨在通过利用分类为“未知”或无法识别合适认知扭曲的数据来解决认知扭曲的不平衡问题。首先，我们分析了认知扭曲的分布情况，以检测哪些类型的比例偏低。然后，通过对标记为“未知”的 17,342 个样本进行重构和重组生成合成数据，确保整体上下文得以保留。

表 4 总结了通过认知澄清和认知平衡方法产生的认知扭曲的分布情况，以及结合两种方法后的总体总数。

表 4: 认知扭曲类型在合成数据生成方法中的分布。

认知扭曲类型	认知澄清（%）	认知平衡（%）	总计 (%)
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%)

4 使用聚类验证合成数据

为了验证我们创建的合成数据的有效性，我们基于两个标准进行了聚类：(1) 触发青少年负面情绪的话题和 (2) 在广泛用于评估和诊断精神障碍的 DSM-5 框架中概述的负面情绪和症状 [20]。

4.1 基于主题的青少年消极思维分类

韩国国家青年政策研究院（NYPI），隶属于性别平等和家庭部，将青少年的关注点分为五个领域：(1) 学术和职业问题，(2) 关系（友谊、恋爱、欺凌），(3) 身体和心理健康，(4) 家庭问题，以及 (5) 外貌和自我形象。

为了评估我们的合成数据在青少年负面思维方面与这些类别的对齐情况，我们应用了 K-均值聚类算法，这是一种无监督机器学习算法，它将数据划分为不同的组群，以从 69,925 名青少年的问题中提取的关键词进行了分析。此过程将数据分成了五个预定义的主题集群，每个集群都根据相关性分配了子关键词。结果创建了一个包含五个主题和 139 个关键词的词典。

我们确定了每个主题中最频繁的关键词。最常见的主题是学术表现和职业担忧，共有 99,076 次实例（36.9%），其次是人际关系（73,586 次，27.4%）、身心健康（71,249 次，26.5%）、家庭问题（20,532 次，7.6%）以及外貌和自我形象（4,007 次，1.5%）。

4.2 基于 DSM-5 的青少年消极思维分类

认知扭曲可能导致抑郁，因此我们检查了 DSM-5 的九个类别以确定是否存在显著关系。为了探讨这一点，我们分析了 69,925 名青少年的问题，并使用 NLTK 文本挖掘技术识别与 DSM-5 相关的词汇分布。这些分布随后被用于创建 DSM 分类词典，结果产生了九个类别和 143 个关键词。

对于关键词映射，我们使用了包含 108,717 个合成数据点的数据集，允许每个数据点有多个关键词。对于基于 DSM 的关键词映射，69,290 个数据点（63.7%）成功映射，分配了 115 个独特的关键词共 1,335,337 次。对于基于触发负面情绪主题的映射，103,183 个数据点（94.9%）成功映射，分配了 129 个独特的关键词共 268,450 次。

在 DSM-5 症状类别中，九个类别中有五个出现了超过 15,000 次。最常见的关键词是 B。兴趣或愉悦感丧失（出现次数为 321,157 次，占比 23.8%），其次是 H。注意力下降（出现次数为 25,580 次，占比 18.9%），A。抑郁情绪（出现次数为 25,258 次，占比 18.7%），D。失眠或过度睡眠（出现次数为 24,864 次，占比 18.4%），以及 E。心理运动性激越或迟滞（出现次数为 15,235 次，占比 11.3%）。

我们发现了 34 个触发认知扭曲的主题关键词和 20 个 DSM-5 类别关键词，每个关键词的频率都在 1000 次或以上，如图 2 所示。

生成的合成数据主要强调了学术和职业压力，以及友谊和恋爱关系等社会冲突，而对外表和自我形象问题的描述较少。此外，其认知扭曲与 DSM-5 抑郁症状关键词中的五个紧密相关。

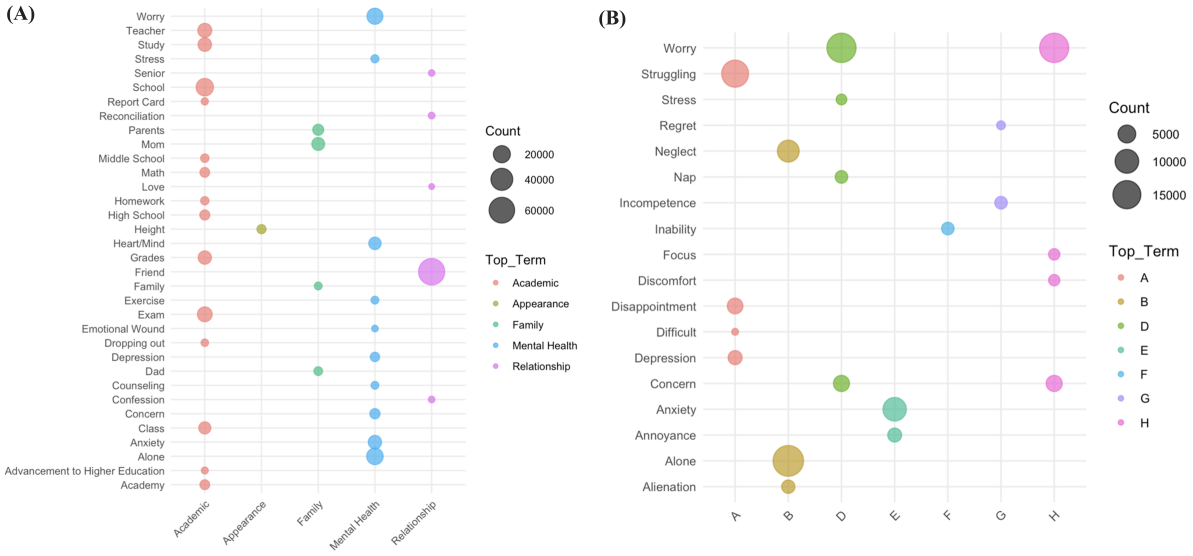

图 2: 高频关键词的聚类分布：(A) 引发负面情绪的主题和 (B) DSM-5 症状关键词（≥ 出现了 1,000 次）。

5 评估

我们评估了两种类型的合成数据的质量：一种是从明确识别的认知扭曲生成的数据，另一种是为了应对认知扭曲不平衡而生成的数据。评估独立使用了三个评价标准，并且包括了大型语言模型和人类的评估。

5.1 评估标准

我们使用三个标准评估生成的合成数据：(1) 一致性，(2) 准确性，和 (3) 流畅度。评分范围从 1 到 3，其中 1 表示 “不适当”，3 表示 “非常适当”。一致性检查原始数据与合成数据之间认知扭曲是否逻辑上保持一致。准确性评估标注的认知扭曲是否符合正确的分类。流畅度评价句子的自然程度、语法正确性和易读性。

5.2 比较准则下的大语言模型与人类评估

为了保证客观性，生成合成数据的模型被排除在评估之外。另外两个模型独立对数据进行了评分，并将它们的分数平均作为最终结果。

对于人类评估，每个失真随机选择了 50 或 100 个合成样本，总计 900 个样本。两位心理学专家使用与 LLM 评估相同的准则独立进行了评估，Cohen 的卡帕值为 0.78，表明有显著的一致性。

表 5 概括了来自大语言模型和人类的评估结果，突出了两种合成数据——认知澄清和认知平衡——在三个标准上的差异。

人类评估分数在所有标准中都较低，除了流畅性，准确性的差距最大。这种差异源于大语言模型在检测显式文本模式方面的优势，但在进行认知扭曲评估所需的关键隐含推理方面却显得吃力，突显了它们的局限性。表 6 提供了详细的专家反馈。

关于两种合成数据生成方法，在大语言模型评估中，认知澄清法在所有标准上的得分比认知平衡法高 0.1 到 0.3 分。然而，在人类评估中，只有认知平衡法显示出更高的准确性。

表 5: 合成数据的 LLM 评估和人类评估结果。LLM 评估（左侧）报告了三个模型分配的平均分 ś 标准差得分，其中标准差表示模型之间的变异程度。人类评估（右侧）展示了两位专家交叉验证后给出的平均分数。

标准	**大语言模型评估**		**人类评估**	
	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

表 6: 专家对一个合成数据准确率为 2 的案例进行分析：解释大型语言模型误分类的原因。

原始 distress 问题	每当我见到我的堂兄弟姐妹时,我妈妈都会问我为什么我不高。即使比我矮的朋友也会熬夜到凌晨两点以后，当我看到他们长高的时候，我只能想着为什么我不高。
[认知扭曲的类型] 合成数据	**["应当" 陈述]** 我妈妈经常把我跟她表亲比较，说：“你为什么这么矮？’我不明白为什么只有我那么矮，而我的朋友们都在长高。我爸爸妈妈都很高，但我觉得个子矮有什么不对。
来自专家的指令	The expert provided two points in the accuracy evaluation of the cognitive distortion type, and chose mental filtering rather than should statement. 认为应该长得高（应该陈述）通常来自父母。在这篇文章中，我们确认母亲不高这一事实引发了焦虑。然而，这假定个人有消极的想法（心理过滤），因为她相信自己不会再长高。虽然从字面上看，“应该”陈述似乎是主要问题，但心理过滤——一种自我判断的错误——被认为是主要的认知扭曲。

5.3 认知扭曲分类中 LLM 和人类表现的比较

为了进一步分析基于 LLM 的评估和人类评估之间的差异，我们比较了每种认知扭曲的得分。表 7 展示了对比结果，突出了两种评估方法的关键差异。分数在 LLM 与人类评估之间进行了比较，并将较高的值用粗体表示。“差异”列显示了分数差距，其中差异为 0.4 或更大的也用粗体表示。

LLMs 的平均评估分数为一致性 2.287，准确性 2.582，流畅性 2.598，而人类评估的平均分数为一致性 2.278，准确性 2.558，流畅性 2.815。在人类评估中，流畅性得分较高，然而一致性与准确性的得分没有显著差异，尽

管总体上人类的得分略低一些。在人类评估中流畅性得分较高的原因可能是 LLMs 评估了合成生成的句子，这些句子结构自然且没有停顿。

在评估不同类型的认知扭曲时，人类的得分在某些情况下低于 LLMs，尤其是在准确性方面。例如，“情感推理”（2.887 对 2.200）和“夸大与缩小”（2.531 对 2.100）的分数显示出显著差异。这种差异可能是因为 LLMs 擅长识别清晰的语言模式，如“应该陈述”、“标签化”和“忽视积极面”。然而，对于需要推理论证的认知扭曲，例如“心理过滤”和“夸大与缩小”，人类评估更为可靠，因为这些依赖于更深层次的语境理解。

这些发现强调了大型语言模型更多依赖于显式的语言模式，而人类评估者则考虑更深层次的语境推理，这可能会影响他们识别需要隐含推断的扭曲的能力。

表 7: LLMs 和人类对认知扭曲的比较评估：一致性 (Cos)，准确性 (Acc)，流畅性 (Flu)。*容易被 LLMs 检测到的认知扭曲类型。

认知扭曲类型	大语言模型评估			人工评估			差异		
	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 Filtering	**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

6 结论与未来工作

我们开发了 KoACD，这是一个关于韩国青少年认知扭曲的数据集，克服了以英语为母语的成年人的小规模数据集的局限性。KoACD 通过创建合成数据提供了对认知扭曲的平衡表示。据我们所知，这是首个专门为韩国青少年设计的数据集。

我们引入了一种多 LLM 谈判方法，以提高合成数据的客观性和准确性。通过使用多个 LLM 进行谈判和细化认知扭曲标签，我们最小化了偏差并提升了数据质量。专家和 LLM 评估确认，在存在清晰语言线索的情况下，LLMs 表现良好，而人类评估者在依赖上下文的情况中表现出更高的准确性。LLMs 与人类评估之间的差异突显了 LLMs 对表面语言模式的依赖。

未来的工作将集中在使用青少年特定的数据对模型进行微调，以增强对认知扭曲的语境理解。此外，我们旨在通过开发更好的区分认知扭曲算法来提高大语言模型的性能，减少对特定类型的认知偏差，并在检测中提升平衡性和准确性。

7 限制

我们认识到在检测认知扭曲的方法以及 KoACD 数据集方面存在一些局限性：

认知扭曲分类 我们将最适当的认知扭曲分配给每个问题，但某些问题可能同时涉及多种扭曲。一些类型的扭曲之间的界限模糊不清，使得分类变得困难，并可能导致模型与人类评分者之间出现差异。为了解决这些问题，需要采用多标签分类方法和更精细的标准。

多语言模型谈判方法 我们设计了 LLMs 在分析员和评估员角色之间交替，但结果会根据使用的模型而有所不同。因此，使用不同 LLMs 的谈判结果也应该被考虑。此外，由于无法将数据归类到十个认知扭曲类别中，分析师和评估员之间的差异有时会导致数据被分类为 “未知”，即使经过五轮谈判也是如此。对这类数据的解释至关重要，并且需要进一步的研究来开发更准确的检测方法。

大语言模型和人类评估 虽然 KoACD 是一个大型数据集，但人类评分员审查的数据量相对较小。尽管人类评分员在考虑上下文以做出准确判断方面表现出色，但在评估过程中可能会出现主观性和由于评分员标准不同而导致的不一致性。未来的研究应侧重于获取更多的人类评估数据，并制定更精确的评估标准以提高可靠性。

8 伦理考虑

在本研究中，我们收集了来自 NAVER 知识 iN 的公开可访问数据，用户匿名参与该平台。我们在研究过程中仅使用了公开可用的数据，并未直接与 NAVER 知识 iN 用户互动。

我们已经确定，数据收集过程可能包含各种不适当的主题，如仇恨言论、暴力、性内容和污言秽语。因此，我们试图通过应用严格的过滤标准尽可能排除此类数据。然而，我们无法完全排除数据中可能存在一些不当内容的可能性。

我们意识到 AI 模型可能在不适当的数据上进行训练，从而产生有偏见或不符合伦理的结果的风险。因此，持续监控 AI 模型的道德使用并改进过滤技术以应对这一风险非常重要。

参考文献

[1] Paula R. Pietromonaco and Hazel Markus. The nature of negative thoughts in depression. *Journal of Personality and Social Psychology*, 48(3):799–807, 1985.

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

[3] Jennifer H. Pfeifer and Elliot T. Berkman. 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, 2018.

[4] Katerina Rnic, David J. A. Dozois, and Rod A. Martin. Cognitive distortions, humor styles, and depression. *Europe's Journal of Psychology*, 12(3):348–362, 2016.

[5] Saghar Chahar Mahali, Shadi Beshai, Justin R. Feeney, and Sandeep Mishra. 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, 2020.

[6] Aaron T. Beck. *Cognitive therapy and the emotional disorders*. Penguin, 1979.

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

[8] Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 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, 2020.

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

[10] Zhiyu Chen, Yujie Lu, and William Yang Wang. 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, 2023. Association for Computational Linguistics.

[11] 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. 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, 2023.

[12] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. arXiv preprint, 2023.

[13] Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. 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, 2024. Association for Computational Linguistics.

[14] Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. Sentiment analysis through llm negotiations. arXiv preprint, 2023.

[15] Moonkyoung Jang and Seongcheol Kim. 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, 2024.

[16] Aaron T. Beck. Cognitive therapy: A 30-year retrospective. *American Psychologist*, 46(4):368, 1991.

[17] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint, 2024.

[18] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence`.

[19] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024.

[20] Geonju Lee, Dabin Park, and Hayoung Oh. Methodology of labeling according to 9 criteria of dsm-5. *Applied Sciences*, 13:10481, 2023.