基于逻辑推理和多任务融合的认知刺激对话生成方法

蒋玉茹[1]，李梦媛[1]*，陶宇阳[1]，区可明[2]，佘泽鹏[2]，施水才[1,2]
1. 北京信息科技大学计算机学院，北京
2. 北京拓尔思信息技术有限公司，北京
{jiangyuru,limengyuan}@bistu.edu.cn
{ou.keming,she.zepeng,shi.shuicai}@trs.com.cn

摘要

在全球老龄化背景下，带有认知刺激的对话系统是保持老年人认知健康的重要手段。中文认知刺激对话数据集(Chinese Cognitive Stimulation Conversation Dataset, CSConv)和模型构建的研究工作刚刚开始。本文将认知刺激对话生成视为一个多任务融合的逻辑思维推理过程，将情感分类任务、决策任务和对话回复生成任务间的逻辑关系，建模为一个推理过程，来引导大语言模型生成。针对决策任务，本文提出分层编码器结构的决策模型。决策实验结果表明，决策模型有效的提高了决策任务的准确率。针对多任务过程，本文提出多任务融合方法，将三个任务对应的模型结合在一起。生成实验结果表明，分类、决策及生成的多任务融合方法，显著提升了对话回复能力，证明了该方法的有效性和先进性。

关键词： 认知刺激对话；逻辑推理；决策任务；多任务融合方法

Cognitive stimulation dialogue generation method based on logical reasoning and multi-task integration

Yuru Jiang[1]，Mengyuan Li[1]*，Yuyang Tao[1]，Keming Ou[2]，Zepeng She[2]，Shuicai Shi[1,2]
1. School of Computer Science,
Beijing Information Science and Technology University
2. TRS Information Technology Co.,Ltd.，Beijing
{jiangyuru,limengyuan}@bistu.edu.cn
{ou.keming,she.zepeng,shi.shuicai}@trs.com.cn

Abstract

In the context of global aging, dialogue systems with cognitive stimulation are an important means to maintain the cognitive health of the elderly. Research work on the Chinese Cognitive Stimulation Conversation dataset (CSConv) and model construction has just begun. This study regards cognitive stimulation dialogue generation as a multi-task integration thinking and reasoning process, and models the logical relationship between emotion classification tasks, decision-making tasks and dialogue response generation tasks as a reasoning process to guide the generation of large language models. For decision-making tasks, this paper proposes a decision-making model with a hierarchical encoder structure. The decision-making experiment results show that the decision-making model effectively improves the accuracy of decision-making tasks. For the multi-task process, this study proposes a multi-task integration method to combine

* 通讯作者

基金项目：北京市自然科学基金项目（4242019），题目：智慧养老服务中对话机器人关键技术研究

the models corresponding to the three tasks. The generation experimental results show that the multi-task integration method of classification, decision-making and generation significantly improves the dialogue response ability, proving the effectiveness and advancement of the method.

Keywords: Cognitively stimulating dialogue , Logical reasoning , Decision-making tasks , Multi-task integration method

1 引言

人机对话系统一直是人工智能研究中的热点。在全球老龄化背景下，利用人工智能实现智慧康养，构建面向老人的对话系统成为对话系统研究的重要方向。研究表明，对话系统对老人认知能力的改善具有积极的作用(De Oliveira et al., 2014; Park et al., 2019)。其中，认知刺激对话是恢复老人认知功能的重要心理治疗手段(Tokunaga et al., 2019; Tokunaga et al., 2021)。认知刺激治疗是一种帮助轻度到重度以下认知障碍症患者的非药物干预治疗手段(Spector et al., 2003)。认知刺激对话是在认知刺激治疗原则[1]指导下，以刺激患者思维为目的生成的对话。结合认知刺激治疗的生成式对话系统在智慧养老领域中有着广阔的应用场景，可以应用于家居生活场景和养老社区服务场景等，通过聊天的方式，帮助老人认知恢复。因此认知刺激对话有着迫切的现实需求、应用价值和多样的应用场景。

之前大部分面向老年人的对话系统主要关注在特定场景中的特定功能研究，例如Tokunaga et al. (2021)等人，提出的一个利用照片故事的机器人对话系统，对老人进行日常的认知训练。Navarro et al. (2018)等人提出了一个认知刺激的治疗系统，用以评估患者的认知障碍和情绪健康，来自动生成新的个性化刺激计划。然而此前，由于认知刺激对话语料的缺失，没有面向关于认知刺激对话的生成式系统研究。

近期，面向老人的生成式认知刺激对话系统研究取得了突破。Jiang et al. (2023)构建了一个中文认知刺激对话数据集CSConv，将情感支持(Liu et al., 2021)和认知刺激同时引入聊天对话中，提出了面向老年人的认知刺激对话新任务。数据集包含2600个对话，每句话语标注了三个标签：认知刺激治疗原则标签、情感标签和情感支持策略标签，其中情感支持策略标签由Liu et al. (2021)提出，认知刺激治疗原则标签由Jiang et al. (2023)从认知刺激治疗手册中总结而来。然后，作者提出了融合多源知识的生成式对话模型CSD，取得了最好的效果。

图 1: 人类思维流程图

认知刺激对话任务不同于普通的闲聊任务，它是带有目的性的对话，任务难度比较高。在进行认知刺激对话的时候，人类思维首先会根据当前话语的内容进行决策，确定适用的治疗原则，随后根据决策的结果进行相应的回应，如图1所示。在情感支持对话(Liu et al., 2021)和认知刺激对话(Jiang et al., 2023)工作中，都提供了一个对话示例，示例中的灰色框描述了在输入和回复之间回复者的思考过程。由此可见，模型在回复时同样会参考人类思维的过程，即先决策后回复。这也表明，认知刺激对话任务是多任务的融合。

因此，本文提出一个多任务融合的方法，如图2所示，把认知刺激对话任务分成三个子任务，即分类任务、决策任务和对话回复生成任务。本文将三个任务间的逻辑关系，视作一个复杂的逻辑推理过程，来引导大语言模型生成。分类任务即情感信息分类，决策任务具体分为认知刺激治疗决策和情感支持策略决策，生成任务为生成模型利用情感信息和决策信息的引导来做出相应的回复。

[1]《认知刺激治疗CST：为认知障碍症设计的循证小组活动（导师手册）》

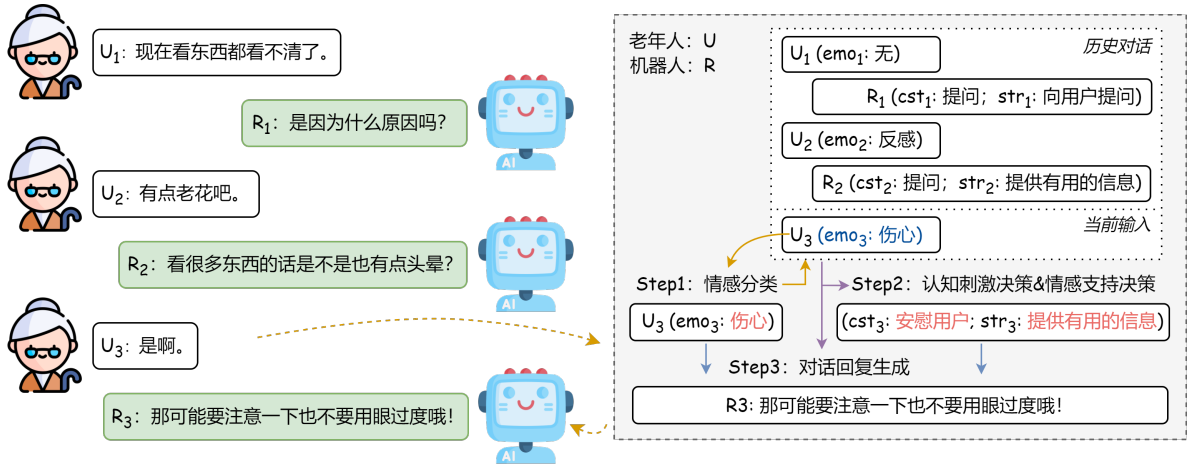

图 2: 多任务融合流程图

针对决策任务，本文提出分层编码器结构的决策模型。在编码阶段，通过融入历史对话，情感信息以及历史决策信息，来增强对话的编码表示，从而提升决策准确率，为生成模型提供更准确的信息。对于生成模型而言，从情感信息和决策信息到回复生成中间还有很大的跨度。生成模型不仅要理解标签含义，还肩负着理解情感信息和决策信息背后蕴藏的指令。大语言模型自出现以来，展现了惊人的推理能力和对指令的服从能力，可以很好的理解指令信息并做出相应的回应，对于不同标签信息背后的含义有着更好的理解。于是本研究使用大语言模型作为生成模型，使用基于提示的方法，探究大语言模型深层次的理解和推理能力。

本研究具体的研究贡献总结如下：

（1）针对认知刺激对话任务，本文参考人类先决策后回复的思维过程，提出多任务流程。将认知刺激对话分为三个子任务，具体为情感分类任务，决策任务，以及大语言模型对话回复生成任务。

（2）针对决策任务，本文提出分层编码器结构的决策模型DM。决策实验结果表明，DM有效的提高了情感支持决策和认知刺激决策的准确率。生成实验表明，随着决策模型准确率的提高，对生成模型生成对话回复也具有正向的促进作用。决策实验和生成实验的结果证明了决策任务和决策模型的有效性。

（3）针对认知刺激对话过程中多任务的特点，本文提出了多任务融合方法CSMI，将情感分类模型、决策模型和大语言模型融合到一起。本研究进行了大量的实验，结果表明CSMI提升了对话回复质量，证明了CSMI的有效性与先进性。

2 相关工作

在面向老人的对话系统研究中，一些研究表明人机对话系统可以帮助老人应对孤独，降低老人的抑郁和焦虑水平，改善生活质量。Ryu et al. (2020)设计和开发了一个心理健康护理对话机器人Yeonheebot，来减少老人的焦虑和抑郁。Yeonheebot会启动关于老年人幸福日常护理的对话，并提供与他们兴趣相关的功能，以吸引用户使用。名为Charlie(Valtolina and Hu, 2021)的面向老人的陪伴型聊天机器人，能提供智力游戏、能主动提醒事务、能聊一些轶事以促进自我共情，意图帮助老人积极应对孤独感，改善生活质量。Lee et al. (2019)也对在对话中融入自我共情(Sharma et al., 2021; Sharma et al., 2020; Rashkin et al., 2019)进行了研究，其实验表明：让用户去关心机器人比让机器人关心用户更能让用户获得自我共情。

还有一些研究表明人机对话系统对老人的认知能力改善具有积极的作用。Tokunaga et al. (2021)研究了一个名字叫Bono 的对话系统，提出一个基于对话的方法对老人进行日常的认知训练，利用对话机器人，向老人展示照片并讲述相关故事，以此作为和老人聊天的上下文，然后请老人提问，机器人进行回答。Kim et al. (2021)利用基于智能语音程序的元记忆训练策略，进行了实验，实验结果表明该训练方法能够帮助到具有主观认知能力下降的老人。

De Oliveira et al. (2014)研究表明多感官认知刺激对长期护理机构中老人的认知能力问题具有有益影响。Park et al. (2019)研究了多组件的认知刺激程序对老年人的认知功能改善具有积

极的作用。Navarro et al. (2018)提出一个认知刺激治疗系统，用以评估患者的认知障碍和情绪健康，来自动生成新的个性化刺激计划。可见，认知刺激作为针对认知障碍老人的非药物心理治疗方法，同样对老人的认知能力改善具有重要作用。

近期，认知刺激和生成式对话系统结合的研究上取得了进展，Jiang et al. (2023)构建了一个中文认知刺激对话数据集CSConv，首次提出面向老年人的认知刺激对话任务。其中包含2600个对话，每句话语标注了三个类别标签，然后提出了融合多源知识的生成式对话模型，取得了当前最好的效果。

3 多任务融合的认知刺激对话方法

本文参考人类在认知刺激对话中，分步进行的特点，提出了多任务过程，并基于此设计了多任务融合的方法，以下简称为CSMI（Cognitively Stimulated dialogue based on Multi-task Integration）。本研究将认知刺激对话任务分成三个子任务，分别是分类任务、决策任务和大语言模型对话回复生成任务。

具体来说，一本文选用BERT(Devlin et al., 2018)作为情感分类模型，针对情感分类任务构建了相应的数据集，并利用该数据集对模型进行微调。二本文构建了决策模型，其包括两个主要阶段：（1）对话编码和（2）标签决策。在标签决策阶段，本研究设计了两种结构，分别是基于LSTM(Shi et al., 2015)构建的决策模型，以下简称DM_{LSTM}（Decision Model-LSTM），和基于Transformer Encoder(Vaswani et al., 2017)构建的决策模型，以下简称$\mathrm{DM}_{Transformer}$（Decision Model-Transformer）。三本文采用提示的方法，进一步提升大语言模型的理解能力和推理能力。最后，结合分类模型和决策模型的结果，将对应三个任务的模型融合在一起，来引导大语言模型生成回复。

3.1 数据集处理

在CSConv数据集中，为每句话语u都标注了三种类型的标签：情感标签u_{emo}、认知刺激治疗原则标签u_{cst}、情感支持策略标签$u_{strategy}$。在Jiang et al. (2023)工作中，作者把每一个标签转换成一个词进行编码，拼接在句子编码之后输入给生成模型，因此没有给出相应的中文标签。于是，对于情感标签和认知刺激治疗原则标签，本研究按照Jiang et al. (2023)工作中的释义，翻译成中文后，再总结为对应的中文标签，如表1所示。对于情感支持策略标签，按照Liu et al. (2021)给出的示例，总结成对应的中文标签，如表2所示。

英文标签	中文标签
None	无
Disgust	反感
Sadness	伤心
Fear	害怕
Surprise	惊喜
Like	喜欢
Happiness	幸福
Anger	愤怒

(a) 情感标签对照表

英文标签	中文标签
None	无
Inquiry	提问
Respect	尊重
Reminiscence	回忆过去
Expression	提高用户语言表达能力
Enjoyment	在谈话中获得乐趣
Comfort	安慰用户

(b) 认知刺激治疗原则标签对照表

表 1: 情感和认知刺激治疗原则中文标签对照表

决策任务应预测出机器回复要采取的情感支持策略或者认知刺激治疗原则。于是根据决策任务的定义，本文重新整理了CSConv数据集。在老人话语句u留下情感信息标签u_{emo}，在机器回复句r留下情感支持策略标签$r_{strategy}$和认知刺激治疗原则标签r_{cst}。

3.2 多任务融合方法

针对决策任务，本文提出具有分层编码器结构的决策模型DM，包含两个阶段：（1）对话编码（2）标签决策。在标签决策阶段，本文通过两种不同的结构设计了决策模

英文标签	中文标签
None	无
Question	向用户提问
Restatement or Paraphrasing	进行更简洁的重新表述
Reflection of Feelings	描述用户感受
Self-disclosure	利用类似经历或情感，表达同理心
Affirmation and Reassurance	肯定优势和能力，提供鼓励
Providing Suggestions	提供改变的建议
Information	提供有用的信息
Others	其他

表 2: 情感支持策略中文标签对照表

型DM_{LSTM}和$\mathrm{DM}_{Transformer}$。决策模型结构如图3所示，图片使用认知刺激决策为例进行展示，情感支持决策与此类似。

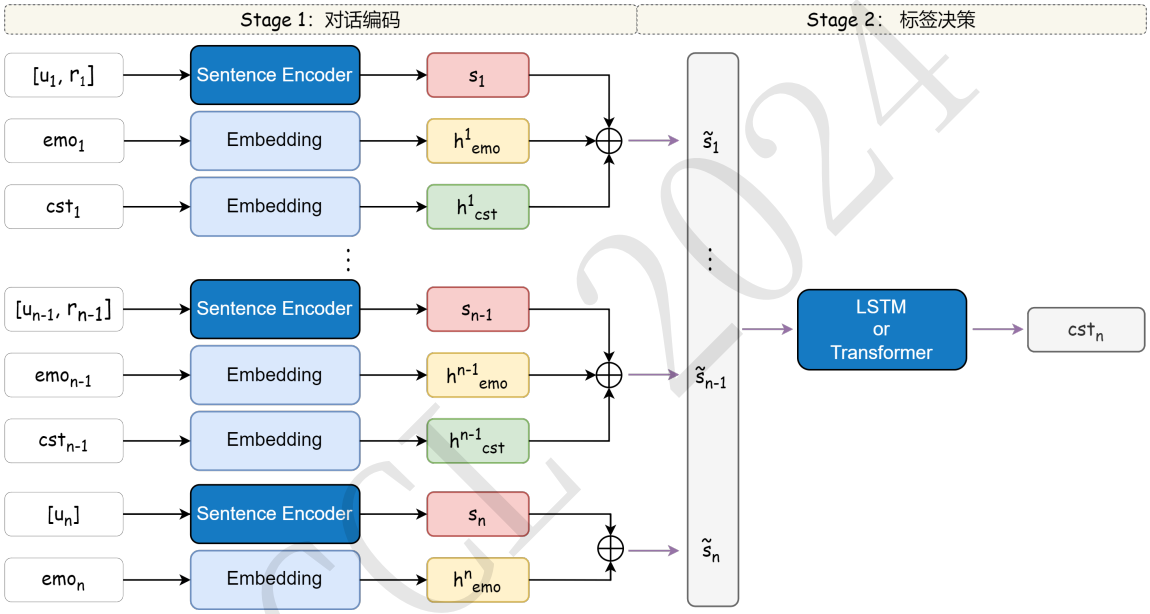

图 3: 决策模型结构图

本文将对话历史中的每一轮设计成一个句对的形式$H = [(u_1, r_1), ..., (u_{n-1}, r_{n-1})]$进行编码，以提高历史对话的结构化表达，其中$u$代表老人的话语，$r$代表机器人的话语。为了更有效地传达句子的语义信息，本文采用句子编码器Sentence BERT(Reimers and Gurevych, 2019)对它们进行编码。再和当前句u_n一起，通过句子编码器做句子表征，如公式（1）所示。若设$s = (u, r)$，则得到对话表征$S = [s_1, s_2, ..., s_n]$，具体表示为：

$$S = SentenceBERT([(u_1, r_1), ..., (u_{n-1}, r_{n-1}), u_n]), S \in \mathbb{R}^{n \times d} \quad (1)$$

其中，n表示对话轮数，d为句子编码器隐含层维度。然后，本文使用两种编码层分别对每一轮的情感信息和历史对话中的决策信息进行编码，即$E_e(emo) \in \mathbb{R}^{n \times d}$、$E_c(cst) \in \mathbb{R}^{n \times d}$或$E_s(str) \in \mathbb{R}^{n \times d}$，其中$n$表示对话轮数，$d$为隐藏层维度。我们将历史情感信息和历史决策信息组成序列，经过编码层后得到输出结果：H_{emo}、H_{cst}和H_{str}，h_p^n表示占位符的编码层结

果。

$$H_{emo} = (h^1_{emo}, h^2_{emo}, ..., h^n_{emo}) \tag{2}$$
$$H_{cst} = (h^1_{cst}, h^2_{cst}, ..., h^{n-1}_{cst}, h^n_p) \tag{3}$$
$$H_{str} = (h^1_{str}, h^2_{str}, ..., h^{n-1}_{str}, h^n_p) \tag{4}$$

将编码层得到的输出结果，与对应的每一轮对话编码结果进行拼接，得到信息增强后的对话编码结果：融合情感和认知刺激决策信息的对话编码$\widetilde{S}_{cst}$以及融合情感和情感支持决策信息的对话编码$\widetilde{S}_{str}$。

$$\widetilde{S}_{cst} = S \oplus H_{emo} \oplus H_{cst} \tag{5}$$
$$\widetilde{S}_{str} = S \oplus H_{emo} \oplus H_{str} \tag{6}$$

对话是一种具有前后顺序关系和上下文联系的序列，因此，在本研究的标签决策阶段中，采用了长短期记忆网络（LSTM）和Transformer Encoder两种不同结构。

LSTM被视为一种具有时间序列特征的递归神经网络模型，其在模型训练中能够有效地捕获句对之间的时序关联，从而有助于深入理解对话中的上下文语境。将对话编码结果传入LSTM得到结果G_{cst}和G_{str}，取其中最后一层的隐含层向量g^n_{cst}和g^n_{str}，经过线性层后得到当前句的决策信息，W_l为决策预测权重矩阵：

$$G_{cst} = LSTM(\widetilde{S}_{cst}) \tag{7}$$
$$l_{cst} = argmax(softmax(W_l g^n_{cst})) \tag{8}$$
$$G_{str} = LSTM(\widetilde{S}_{str}) \tag{9}$$
$$l_{str} = argmax(softmax(W_l g^n_{str})) \tag{10}$$

Transformer Encoder是一种基于自注意力机制的模型，能够通过自注意力有效地建模句子中词与词之间的关系，从而实现对对话信息的深入理解。为了促进Transformer Encoder对句与句关系更为深入的理解，本文采用一个位置编码层$E_p(hp) \in \mathbb{R}^{n \times d}$表示句对之间的上下文关系，其中$n$表示对话轮数，$d$为隐藏层维度与句子编码器隐藏层向量维度保持一致。将对话编码结果和位置编码结果相加，传入Transformer Encoder得到结果M_{cst}和M_{str}，取其中最后一层的隐含层向量m^n_{cst}和m^n_{str}，经过线性层后得到当前句的决策信息，W_l为决策预测权重矩阵：

$$H_p = (h^1_p, h^2_p, ..., h^n_p) \tag{11}$$
$$M_{cst} = TransformerEncoder(\widetilde{S}_{cst} + H_p) \tag{12}$$
$$l_{cst} = argmax(softmax(W_l m^n_{cst})) \tag{13}$$
$$M_{str} = TransformerEncoder(\widetilde{S}_{str} + H_p) \tag{14}$$
$$l_{str} = argmax(softmax(W_l m^n_{str})) \tag{15}$$

针对大语言模型，本文选择中文大语言模型ChatGLM(Du et al., 2022)作为生成模型，并采用基于提示的方法，设计了清晰高效的提示模板，将任务解释和逻辑过程设计成提示模板，嵌入在生成模型的对话模板中。任务提示模板嵌入在ChatGLM对话模板中的“问：”之后，拼接在每次对话的开始，旨在更有效地引导大语言模型了解决策任务，增进其对标签语义信息的理解。任务提示模板设计如下所示：“你是一个具备认知刺激和情感支持功能的聊天机器人，你的聊天对象是一位老人，你要根据认知刺激策略和情感支持策略选择恰当的回复。现在开始聊天：{用户输入}”。对于回复提示模板，本文将三个任务间的逻辑关系，即“根据用户情感和发言决定回复”，视作一个复杂的逻辑思维过程，以此设计提示来引导大语言模型生成。回复模板设计如下：“根据用户发言中的情绪为[{情感标签}]，应该选择的认知刺激疗法是[{原则标签}]，使用的情感支持策略是[{策略标签}]，所以回复为”。将回复提示模板嵌入到ChatGLM对话模板中的“答：”之后。随后微调大语言模型，在微调时，模板中的占位符将被替换为具体的标签信息。

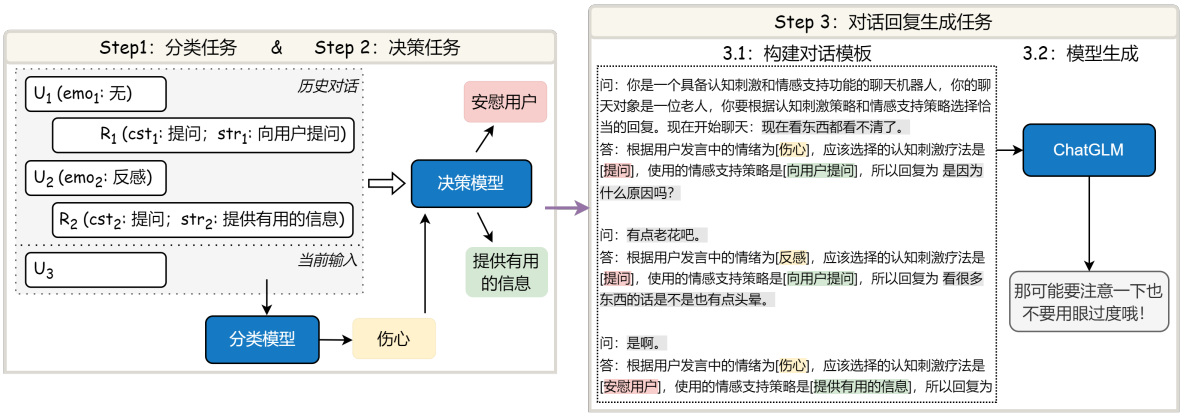

图 4: 多任务融合方法总体模型图

多任务融合方法模型图如图 4所示。首先，将当前话语通过情感分类得到情感信息。然后，结合当前话语及其历史对话和情感信息通过决策模型得到认知刺激治疗原则与情感支持策略两个决策信息。最后，结合情感分类模型和决策模型的结果，将所有需要的信息嵌入在对话模板中，通过生成模型得到对话回复。

4 实验与分析

4.1 实验设置

在情感分类模型中，我们使用的预训练模型是bert-base-Chinese，设置学习率为2e-5。在决策模型中，句子编码器使用的预训练模型是sbert-base-chinese-nli，设置学习率为2e-5，LSTM和Transformer Encoder的学习率为1e-4，批次大小为8，累积步数是4。情感信息和决策信息编码层的向量维度设置为32，输入的最大历史对话长度设置为3轮对话。对于生成模型，我们设计了任务提示模板和回复提示模板，将其添加到ChatGLM的回复模板中，并使用Ptuning方法微调ChatGLM[2]。设置最大序列长度为512，批次大小为16，累积步数为2。

4.2 评价指标

在决策模型的评估中，本研究选用准确率（Accuracy）和F1分数（F1-score）作为主要评价指标。准确率旨在度量模型正确预测标签的数量，而F1分数则综合考虑了精确率和召回率，为模型的综合性能提供了评估标准。

本研究沿用前人的工作，以BLEU(Papineni et al., 2002)作为评估对话生成质量的指标，BLEU着眼于衡量生成文本与参考文本之间的语义相似性。还引入了ROUGE(Lin, 2004)来评估对话生成质量，ROUGE侧重于评估生成文本对信息的涵盖度和完整性。通常情况下，这两种指标结合使用，以全面评估模型的性能。此外，本研究还统计了BERTScore(Zhang et al., 2019)，用以度量生成文本与参考文本之间的语义相似性，进一步丰富了评价的维度。

4.3 对比实验

为了全面评估本文提出的决策模型在决策任务上的表现，本研究设计了一系列对比实验，涉及多种深度学习模型，包含基于LSTM构建的决策模型DM_{LSTM}和基于Transformer Encoder构建的决策模型$\mathrm{DM}_{Transformer}$。此外，为了确保评估的全面性和深度，本研究还纳入了大语言模型作为决策模型的基线方法。由于大语言模型具有推理能力，所以本文尝试让大语言模型自己担任决策模型的角色。思维链（CoT）(Wei et al., 2022)作为一种设计Prompt(Liu et al., 2023)的方法，即Prompt中除了有任务的输入和输出外，还包含推理的中间步骤。由于认知刺激对话任务，从输入到输出也存在着逻辑关系，所以本文选择使用CoT的方法，让大语言模型自己作为决策模型，思维链的设计参考3.2中的回复提示模板。

如表3所示，$\mathrm{DM}_{Transformer}$和$\mathrm{DM}_{LSTM}$在两个决策任务的准确率和F1分数上高于其他基线，由此证明了两种决策模型$\mathrm{DM}_{Transformer}$和$\mathrm{DM}_{LSTM}$具备有效性与先进性。

[2]https://github.com/THUDM/ChatGLM2-6B

模型名称	认知刺激决策		情感支持决策	
	acc	f1	acc	f1
思维链推理	62.81	62.94	65.96	65.29
BERT	63.41	63.38	70.53	70.16
DM$_{Transformer}$	**65.96**	**66.34**	**72.54**	**70.93**
DM$_{LSTM}$	**67.37**	**67.14**	**72.63**	**71.92**

表 3: 决策任务对比实验结果

模型名称	BLEU-2	BLEU-4	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L
CDialGPT$_{base}$	17.55	6.22	57.70	-	-	-
CDialGPT$_{large}$	15.05	5.47	57.81	-	-	-
GPT2-chitchat	34.61	21.04	66.37	-	-	-
Distil-cluecorpussmall	39.94	25.30	69.41	-	-	-
Cluecorpussmall	41.04	26.59	68.65	-	-	-
CSD	45.53	30.90	74.61	-	-	-
ChatGLM	**46.30**	**39.32**	**77.67**	**49.89**	**36.82**	**49.51**
CLC	**43.86**	**36.27**	**77.08**	**46.91**	**33.80**	**46.28**
CSMI-BERT	**44.35**	**36.91**	**77.13**	**48.27**	**34.02**	**47.42**
CSMI-DM$_{Transformer}$	**47.92**	**37.23**	**77.12**	**48.18**	**33.94**	**47.57**
CSMI-DM$_{LSTM}$	**48.83**	**38.85**	**78.27**	**50.34**	**36.27**	**49.71**
CSMI-真实标签	**50.41**	**41.26**	**79.85**	**54.87**	**38.56**	**54.19**

表 4: 生成任务对比实验结果

对于对话回复生成任务，以上述分类模型、决策模型和大语言模型ChatGLM为骨干，本文构建了以下变体模型：（1）微调版本（ChatGLM）：直接使用CSConv数据集微调ChatGLM。（2）基于思维链微调（CLC）：使用思维链方法微调大语言模型。（3）多任务融合方法（CSMI）：使用基于提示的方法微调大语言模型。此外，本文亦通过对CSConv数据集上的先前研究成果进行对比，来体现本研究方法的先进性。包括：在LCCC不同版本数据集上训练的CDialGPT$_{base}$(Wang et al., 2020)和CDialGPT$_{large}$(Wang et al., 2020)；在闲聊语料上训练的GPT2-chitchat；在CLUECorpusSmall语料库上训练的Distil-Cluecorpussmall(Radford et al., 2019)和Cluecorpussmall(Radford et al., 2019)，以及Jiang et al. (2023)提出的中文认知刺激对话系统CSD。

如表4所示，CSMI方法的评分高于前人的工作，由此证明多任务融合方法的有效性与先进性。同时，CSMI与不同决策模型的评分随着决策准确率的提升而提高，表明决策信息对于生成模型有着正向的引导作用，证明了多任务融合方法的有效性与先进性。但是，CSMI与不同决策模型的评分略低于ChatGLM和CSMI-真实标签的评分，因为如果决策策略与真实值不同，则生成的回复与真实回复会有很大的差异。由此可见，在没有提供真实标签的情况下，学习决策信息是很重要的。

4.4 消融实验

为了进一步验证决策模型和提示模板在大语言模型上的表现，本文设计了一系列消融实验。

本文通过对话编码中是否融入情感信息和决策信息作为控制条件进行实验，表5展示了以LSTM构成的决策模型DM$_{LSTM}$，针对每一部分设计，在认知刺激决策任务上的消融实验结果。可以看到，历史对话信息、情感信息和历史决策信息，都对决策模型DM$_{LSTM}$产生贡献。其中情感信息和历史决策信息的贡献更为突出，表明模型可以有效的利用情感和决策信息，进

模型名称	历史轮数	含情感信息	含历史决策信息	ACC	F1
DM_{LSTM}	1	否	否	63.51	63.79
DM_{LSTM}	3	否	否	65.26	65.65
DM_{LSTM}	3	是	否	66.67	66.32
DM_{LSTM}	**3**	**是**	**是**	**67.37**	**67.14**

表 5: DM_{LSTM}模型在认知刺激决策任务上的消融实验结果

一步提升模型决策能力。

本文还通过控制对大语言模型不同的提示模板信息为条件进行实验。如表6所示，加入了回复提示模板微调的模型优于直接微调的模型，可见把三个任务间的推理过程作为回复提示，进一步提高了大语言模型对于标签信息的理解能力。在增加了任务提示模板后，BLEU-4和ROUGE-L分数均又提高了1.5左右，由此证明基于提示的方法可以激发大语言模型对任务深层次的理解和推理能力，表明了基于提示方法的有效性。

模型名称	回复提示模板	任务提示模板	BLEU-4	ROUGE-L
ChatGLM	否	否	39.32	49.51
CSMI	是	否	39.62	52.62
CSMI	**是**	**是**	**41.26**	**54.19**

表 6: 基于提示模板的大语言模型消融实验

4.5 人工评测

对话生成任务中，自动化评测指标往往存在局限性，因为其无法反映出模型所生成的结果与参考目标之间在深层语义及流畅度等方面的差距。因此本文开展了人工评测，从下述三个主要评价角度衡量模型的生成能力：

- 流畅度：机器回复是否流畅并符合中文语法。
- 支持性：模型回复能否提供帮助。
- 决策匹配度：机器回复是否能根据给出的决策信息，生成对应的回复内容。

对话主题	流畅度	支持性	决策匹配度	
			认知刺激	情感支持
出题游戏	4.61	3.64	4.24	4.07
回忆过往	4.28	4.06	4.67	4.39
分享生活	4.43	3.98	4.41	4.56
平均分	**4.44**	**3.89**	**4.62**	**4.34**

表 7: 人工评测结果

本文选取CSMI-DM_{LSTM}作为人工评测的模型对象，保留其在测试集上的生成结果。本研究从测试集中选取了三个出现频次较多的对话主题作为评测内容，分别是出题游戏、回忆过往和分享生活，共三段完整的多轮对话。以上评测标准的衡量标准均为1-5分，分别表示不可接受-优异。本研究邀请5名评测人员参加人工评测，结果如表7所示。总体来看，流畅度和决策匹配度都达到了良好的水平。支持性评分表明，评测人员认为机器在对话中给出的回复可以提供些许帮助。决策匹配度评分相对较高，说明决策信息作为一种指令，可以很好的引导大语言模型生成回复。

4.6 样例分析

历史对话	U_1：我昨天做了个小手工。 R_1：做了个什么呢? U_2：拿了干花粘在了木框上做成了相框。 R_2：很有意思诶。
当前输入	U_3：但是我不会搭配颜色，粘的很乱。
情感分类结果	emo：伤心
决策结果	cst：安慰用户；str：肯定优势和能力，提供鼓励
模型回复	**R_3：没有啊！我觉得很有创意，真是太厉害了。**

表 8：样例分析

本文从测试集生成的结果中选取了一个对话样本作为分析样例，同时保存了其生成时情感分类以及决策的中间结果，如表8所示。老人分享了自己做手工的事情，但是觉得自己做的不完美，给出了“但是我不会搭配颜色，粘的很乱。”的回复。对于老人的这一回答，情感分类模型对于当前句分类为伤心，决策模型给出了安慰用户以及提供鼓励的决策结果，生成模型根据以上结果给出回复“没有啊！我觉得很有创意，真是太厉害了。”。模型的回复在安慰老人的同时又给出了自己认为很有创意的肯定和鼓励。由此可见，情感信息和决策信息可以作为一种指令，来引导大语言模型生成与它们高度相关的回复。同时也表明了情感和决策判断的正确与否，会对后续生成产生影响。表明了学习情感信息和决策信息的重要性，证明了多任务融合方法的有效性。

5 总结

本文将认知刺激对话任务视为具有分类、决策和生成任务的多任务融合推理过程。重新整理了决策数据集，设计了带有分层编码器的决策模型。并将情感分类模型、决策模型和生成模型融合到一起，提出多任务融合方法。决策实验以及生成实验，都证明了本研究方法的有效性和先进性。本文研究的这一进展，为认知刺激对话任务提供了新的视角，不仅拓展了对认知刺激对话任务的理解，而且为相关领域的未来研究提供了有价值的参考和启示。同时也为认知刺激对话在结合大型语言模型方面开辟了新的方向。由于我们方法灵活多样，未来将着重设计更加强大的决策模型，并且探索结合不同的大语言模型。这些努力将为中文认知刺激对话任务的发展做出新的贡献。

参考文献

T. C. De Oliveira, F. C. Soares, L. D. De Macedo, D. L. Diniz, N. V. Bento-Torres, and C. W. o Diniz. 2014. Beneficial effects of multisensory and cognitive stimulation on age-related cognitive decline in long-term-care institutions. *Clin Interv Aging*, 9:309–320.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10628–10640, Toronto, Canada, July. Association for Computational Linguistics.

Jeongsim Kim, EunJi Shin, KyungHwa Han, Soowon Park, Jung Hae Youn, Guixiang Jin, and Jun-Young Lee. 2021. Efficacy of smart speaker–based metamemory training in older adults: Case-control cohort study. *Journal of medical Internet research*, 23(2):e20177.

Minha Lee, Sander Ackermans, Nena Van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Javier Navarro, Faiyaz Doctor, Víctor Zamudio, Rahat Iqbal, Arun Kumar Sangaiah, and Carlos Lino. 2018. Fuzzy adaptive cognitive stimulation therapy generation for alzheimer's sufferers: Towards a pervasive dementia care monitoring platform. *Future Generation Computer Systems*, 88:479–490.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Jeong-Mo Park, Mi-Won Kim, and Hee-Young Shim. 2019. Effects of a multicomponent cognitive stimulation program on cognitive function improvement among elderly women. *Asian Nursing Research*, 13(5):306–312.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Hyeyoung Ryu, Soyeon Kim, Dain Kim, Sooan Han, Keeheon Lee, and Younah Kang. 2020. Simple and steady interactions win the healthy mentality: designing a chatbot service for the elderly. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–25.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online, November. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Aimee Spector, Lene Thorgrimsen, BOB Woods, Lindsay Royan, Steve Davies, Margaret Butterworth, and Martin Orrell. 2003. Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia: randomised controlled trial. *The British Journal of Psychiatry*, 183(3):248–254.

Seiki Tokunaga, Katie Seaborn, Kazuhiro Tamura, and Mihoko Otake-Matsuura. 2019. Cognitive training for older adults with a dialogue-based, robot-facilitated storytelling system. In Rogelio E. Cardona-Rivera, Anne Sullivan, and R. Michael Young, editors, *Interactive Storytelling*, pages 405–409, Cham. Springer International Publishing.

Seiki Tokunaga, Kazuhiro Tamura, and Mihoko Otake-Matsuura. 2021. A dialogue-based system with photo and storytelling for older adults: toward daily cognitive training. *Frontiers in Robotics and AI*, 8:644964.

Stefano Valtolina and Liliana Hu. 2021. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter*, pages 1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 91–103. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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