CCL24-Eval 任务8系统报告：基于提示工程和思维链的提示词构造

罗允[*]
北京交通大学
计算机科学与技术学院
23120391@bjtu.edu.cn

冯毅[*]
北京交通大学
计算机科学与技术学院
21112027@bjtu.edu.cn

景丽萍[†]
北京交通大学
计算机科学与技术学院
lpjing@bjtu.edu.cn

摘要

儿童故事常识推理与寓意理解评测任务旨在从常识推理和寓意理解两个任务多角度评价中文预训练语言模型和大型语言模型的常识推理和故事理解能力，这考察了模型的常识储备能力以及对文本内容的深入理解能力，因此极具挑战性。随着大语言模型的发展，其卓越的指令跟随能力显著提升了自然语言处理任务的效率和效果。然而，这也对提示词的设计提出了更高的要求，因为提示词的质量直接影响了大模型的表现和预测结果的准确性。因此，设计有效的提示词变得尤为重要，不仅需要理解任务的具体需求，还要具备对语言模型的深入认识和灵活运用能力。本文针对儿童故事常识推理与寓意理解评测赛道一的两个任务，提出了一种基于提示工程的提示词构造方法。首先，我们提出了一种基于融合提示工程、思维链的通用提示词构建框架；然后，我们针对具体的任务调整对应的提示词模板；最后，结合语言模型使用这些提示词进行结果预测。在本次评测中，我们的方法在赛道一的封闭数据条件下获得了第三名的成绩，这验证了我们方法的有效性，并展示了其在自然语言理解领域的应用潜力。

关键词： 提示工程；思维链；少样本学习；上下文学习

System Report for CCL24-Eval Task 8: Prompt Construction Based on Prompt Engineering and Chain-of-Thought

Yun Luo[*]
Beijing Jiaotong University
School of Computer Science and Technology
23120391@bjtu.edu.cn

Yi Feng[*]
Beijing Jiaotong University
School of Computer Science and Technology
21112027@bjtu.edu.cn

Liping Jing[†]
Beijing Jiaotong University
School of Computer Science and Technology
lpjing@bjtu.edu.cn

Abstract

Evaluation on Commonsense Reasoning and Moral Understanding in Children's Stories (CRMU), aims to evaluate the commonsense reasoning and story comprehension abilities of Chinese pre-trained language models and large language models from multiple perspectives, focusing on both Commonsense Reasoning (CR) and Moral Understanding (MU). This task tests the models' ability to retain commonsense knowledge and their deep understanding of textual content, making it extremely challenging. With the development of large language models, their excellent instruction-following capabilities have significantly improved the efficiency and effectiveness of natural language

[*]相同贡献
[†]通讯作者

processing tasks. However, this also raises the bar for prompt design, as the quality of the prompts directly affects the model's performance and the accuracy of the predictions. Therefore, designing effective prompts has become crucial, requiring not only an understanding of the specific task requirements but also an in-depth knowledge of and flexibility in using language models. In this paper, we propose a prompt construction method based on prompt engineering for the two tasks in Track 1 of the Children's Story Commonsense Reasoning and Implication Understanding Evaluation. First, we introduce a general prompt construction framework that integrates prompt engineering and chain of thought reasoning. Then, we adjust the corresponding prompt templates for each specific task. Finally, we use these prompts in conjunction with the language model to generate prediction results. In this evaluation, our method achieved third place under the closed data conditions of Track 1, demonstrating the effectiveness of our approach and its potential applications in the field of natural language understanding.

Keywords: Prompt Engineering , Chain of Thought , Few-shot Learning , In-context Learning

1 引言

近年来，随着预训练模型的兴起及比例定律 (Kaplan et al., 2020; Bahri et al., 2021)的提出，语言模型的规模界限不断被刷新。xAI公司推出的Grok-1模型，以其庞大的参数量，成为了迄今为止最大的开源语言模型。尽管这些语言模型在特定任务上的表现已经接近甚至超越了人类水平，但它们在处理需要推理能力和深层语义理解的文本任务时，仍显示出一定的局限性。

常识推理和寓意理解是自然语言处理领域的两大核心研究方向。常识推理聚焦于让模型具有人类一样的理解和运用广泛的常识性知识的能力，这些常识通常涵盖生物、物理、时间概念以及社会习俗等诸多维度(Davis and Marcus, 2015; Liu and Singh, 2004; Talmor et al., 2018)。该领域的目标在于使模型能够结合常识性知识，更好地模拟人类进行分析与推理。寓意理解的核心在于使模型能够敏锐地识别并理解隐藏在文本表层含义之下的深层语义。这意味着模型需要具备与人类相近的文本理解能力，能够捕捉到文本中的隐喻、讽刺等复杂语言现象 (Tanasescu et al., 2018; del Pilar Salas-Zárate et al., 2017)。因此，常识推理和寓意理解的研究不仅对于推动语言模型的理解能力至关重要，而且对于提升模型在复杂语言环境中的表现和适应性具有深远的影响。

当前已经有工作通过使用预训练模型来完成这些任务，比如Liu等人 (Liu et al., 2021)提出可以先从语言模型中生成与问题相关的常识知识，然后在回答问题时将这些知识作为额外的输入。Ling等人 (Ling et al., 2023)受到人类解释中的对比性质的启发，使用语言模型完成解释提示，这些提示根据证明正确答案所需的关键属性来对比不同的选项。通过将这些解释作为条件来指导模型的决策。这些研究采用了少样本学习的策略，通过为模型提供少量的常识性知识和学习样例，使模型能够学习到相关的知识和回答的格式。然而，在针对当前评测任务的应用中，这种方法展现出了一定的局限性。具体而言，由于本次评测任务中的每条数据均包含较长的故事文本、故事相关的问题以及多个选项，与常规问答任务中的数据相比，文本长度显著增长。当面临过长文本时，这些方法可能会遇到模型输入长度的限制问题，或者因文本信息过于复杂而导致信息遗忘，从而影响模型的性能。

因此，针对这些问题，本研究提出了一种结合提示工程与思维链技术的构建提示词策略，旨在利用大模型进行常识推理和寓意理解任务。为了充分激发大规模模型的潜在能力，本文建议采用提示工程的方法构建提示词框架，该框架涵盖任务描述、模型在任务中的角色定位、执行任务时需要注意的事项，以及评价结果的标准。在本研究中，我们采取了一种综合方法，将少样本学习技术 (Brown et al., 2020; Wang et al., 2020)与思维链技术相结合。为了应对过长输入内容可能导致的模型对提示词信息的遗忘问题，我们在少样本学习部分没有选择堆砌案例——而是为每个任务只提供一个具有代表性的案例。通过向模型展示问题的正确答案，并引

导模型对问题的各个选项进行深入的分析，以此作为提示词中的学习案例。最终，按照提示工程的要求，设计提示词的格式并利用分隔符进行引导，以促使模型生成规范且精确的答案。这种方法不仅提高了答案的准确率，保证了答案的格式规范性，也便于后续的处理与分析工作。

本文针对儿童故事常识推理与寓意理解评测赛道一的两个任务，提出了一种基于提示工程的提示词构造方法。鉴于两个任务在模型能力需求上的差异——常识推理任务侧重于模型结合常识性知识进行逻辑推断，而寓意理解任务则强调模型通过语言表层深入挖掘深层的语义内涵——我们率先构建了一个融合提示工程与思维链的通用提示词框架。随后，针对两个任务的具体特点，我们对提示词模板进行了精细化的调整，以确保模型能够运用相应的能力来完成任务。最后结合语言模型使用这些提示词进行结果预测。在本次评测中，我们的方法在赛道一的封闭数据条件下取得了第三名的成绩，充分验证了本研究所提出方法的有效性。

2 相关工作

2.1 常识推理

常识推理研究侧重于如何更好地运用外部知识来进行推理。之前的工作大多聚焦于如何更好地使用常识知识构建知识图谱，并使用神经网络结合知识图谱完成推理。Moghimifar等人 (Moghimifar et al., 2020)提出当时的方法面对没有见过的情况时无法识别多样化的隐含社群关系，从而无法估计正确的推理路径。因此设计了一种名为COSMO（Conditional SEQ2SEQ-based Mixture model）的条件序列到序列混合模型，该模型能够动态地生成多样化的内容，并用于形成即时的动态知识图谱，以支持常识推理。Sap等人 (Sap et al., 2019)构建了ATOMIC，这是一个包含推理知识的知识图谱，其中的信息以if-then的形式呈现事件、心理状态和角色之间的关系。常识推理通常依赖于预先构建的常识知识图谱，旨在通过结构化的方式整合和表示广泛的常识性信息，从而为推理过程提供必要的知识支持。然而，当面临全新的、之前未曾遇到过的问题形式或者特定的常识领域时，这类方法往往会展现出其局限性。

2.2 寓意理解

传统的机器阅读理解任务往往将关注的重心放在机器的推理能力上。比如Li等人 (Li et al., 2022a)提出了一种神经-符号方法，该方法通过在代表文本单元之间逻辑关系的图上传递消息来预测答案，在处理需要逻辑推理的机器阅读理解任务上显示出有希望的结果。而寓意理解在此基础上还需要对深层语义的理解能力。Guan等人 (Guan et al., 2022)通过构造寓意故事数据集STORAL并在传统的模型上进行了广泛的实验以展示此类任务的难度，并提出一种检索增强的算法，通过从训练集中检索相关概念或事件作为额外的指导来提高模型性能。本文的方法与Guan的方法相似，但我们的任务更注重使用提示工程设计提示词来对模型进行指导。

2.3 上下文学习

当预训练模型达到相当规模时，其将展现出优秀的上下文学习能力（亦称情境学习） (Min et al., 2022)。具体而言，针对一个预先充分训练的庞大模型，当面临迁移至全新任务时，无需对模型进行繁琐的微调操作。仅需提供数个简明的输入-输出对示例，该模型便能理解并适应新任务的具体要求，从而展现出其高度的适应性和学习能力 (Dong et al., 2022)。

提示工程是指设计与优化输入给人工智能模型的提示词，以确保模型能够更好地根据提示词生成预期的内容。比如，Kong等人 (Kong et al., 2023)提出了一种策略性设计的角色模拟提示方法，通过设计特定的角色扮演提示来引导大型语言模型进行推理。这些提示旨在激发模型扮演特定角色或实体的能力，从而模拟出更接近真实场景的交互和问题解决过程。基于此，我们也在提示词中加入了角色扮演的部分，与Kong等人做法的不同之处在于我们为完成常识推理此类任务设计了多个角色，激发模型全面地思考问题的能力。

对于一些较为复杂的任务，比如算术、常识和符号推理等任务上，让模型直接生成最后的答案效果可能会很差，这种情况下我们可以使用思维链来激发模型的推理能力。Wei等人 (Wei et al., 2022b)受到之前使用形式语言生成中间步骤以及模型从上下文中可以进行少样本学习等工作的启发，提出了思维链的方法。通过在提示词中给出推理的中间步骤，来引导模型在之后的生成过程中生成中间步骤来更好地完成推理任务，在GSM8K任务上使用PaLM 540B结合思维链的方法实现了当时最先进的性能。在完成选择题时，分析选项-排除错误选项-得出正确选

项是一种产生中间步骤的方法，因此我们在提示中加入了由大语言模型生成的对各个选项的分析，来鼓励大模型分步解决问题。

在过去的一年中，大型语言模型技术取得了显著的学术与工业进展，市场上涌现出了一系列卓越的大型语言模型产品。以清华大学研发的GLM-4、百度公司精心打造的文心大模型4.0，以及OpenAI公司推出的GPT-4为例，这些模型不仅在文本生成领域展现出了卓越的能力，它们在机器翻译、命名实体识别和主题抽取等任务中也表现出了优秀的性能。鉴于此，本研究旨在探究将优秀的大型语言模型与精心设计的提示词相结合，在处理复杂任务时可能取得的效果。

3 任务描述

儿童故事常识推理与寓意理解评测分为常识推理和寓意理解两个子任务，旨在评价模型的常识推理与故事理解的能力。数据集的规模如下表所示：

数据集	开发集	测试集	总计
常识推理(CR)	400	1692	2092
寓意理解(MU)	252	1056	1308

Table 1: 数据集规模展示

常识推理：常识推理子任务的问题和答案由人工标注，问题涉及到的常识类型包含社会常识、生物常识、时间常识、空间常识以及物理常识。如Figure 1所示：

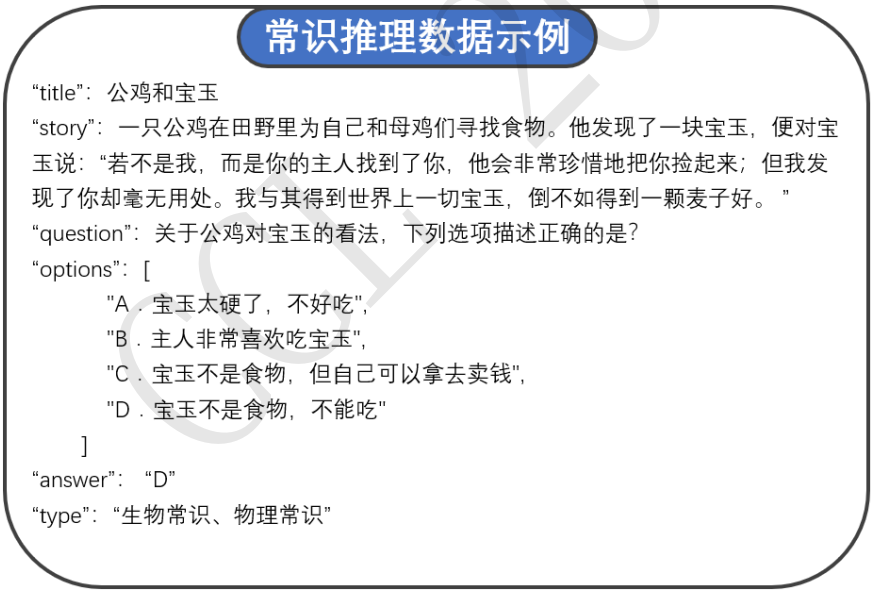

Figure 1: 赛道一常识推理数据示例

想要正确地回答图一中的问题，首先需要有一定的生物知识，即宝玉是一种矿物，它并不在公鸡的食谱上；其次，具备的物理常识让我们知道，宝玉质地坚硬，公鸡不吃宝玉并非是因为宝玉不好吃，而是其不能吃，因此对于公鸡来说宝玉还不如一粒麦子。这样可以推理出选项D是正确答案。

寓意理解：寓意理解子任务的问题和答案采用自动构建和人工标注结合的方式。题目一般要求从四个候选选项中选择最符合故事情节的寓意，如Figure 2所示：

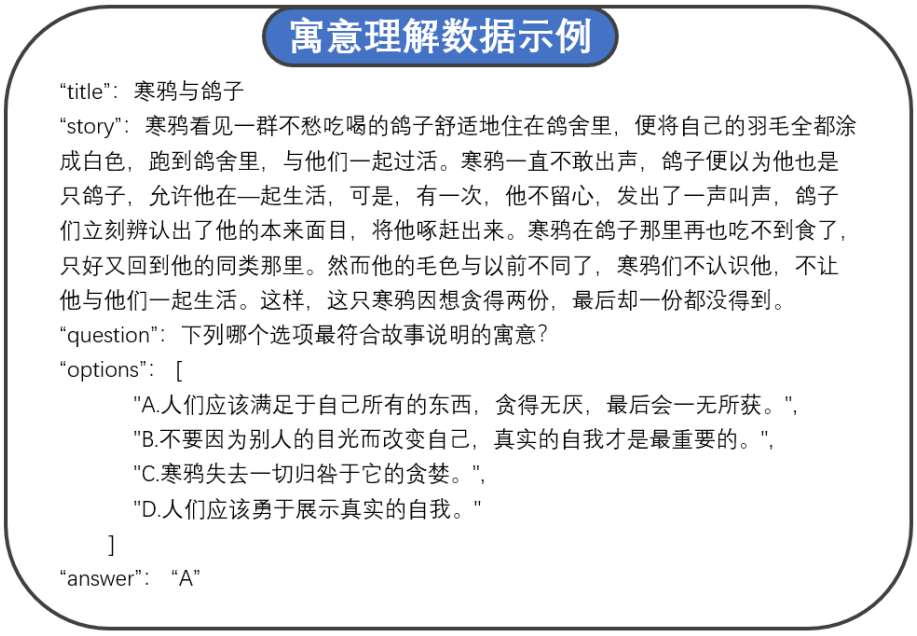

Figure 2: 赛道一寓意理解数据示例

要想推理出正确答案，首先需要理解故事。在这个故事里，寒鸦因为自己的贪念，在看见鸽子舒适的生活后伪装自己混入鸽群，在被鸽子识破之后失去了一切，最后甚至不被自己的族群所接受。然后，结合文章开始进行推理——B选项虽然涉及到“改变自己”这一故事情节，但与本文的主题无关；C选项只是重复了情节，浮于文本的表面而没有触及深层的语义；D选项与B选项类似，虽然“真实的自我”与故事的文本相关，但并非是故事的寓意。寓言故事往往通过虚构情节和以动物作为主人公来向阅读者传达哲理或警示，这则寓言故事正是在告诫我们知足是一种美德，贪婪可能会葬送我们所拥有的一切。因此，A选项是正确的选项。

4 提示词设计方法

4.1 明确任务和角色

在本研究中，我们在提示词的初始部分明确指出了大模型所执行任务的具体名称。随后，遵循OpenAI在其官方网站上发布的关于提示词设计的方法，我们进一步指导大模型在任务中应扮演的角色及其需完成的具体任务。以故事寓意理解任务为例，我们期望大模型不仅能够深入理解故事内容，还能进行批判性分析，并最终做出决策。因此，我们将这些要求明确地体现在对大模型角色的设定上：

在提示词的开篇明确任务以及角色：

#任务名称#：儿童寓言故事理解题（选择最佳寓意）
#扮演角色#：
故事解读者：精确理解故事的情节及其所要传达的核心思想。
批判性分析者：分析故事背后的深层含义，并批判性地评估各个选项与故事的一致性。
决策者：在理解故事的基础上，做出最符合故事意旨的选择。

4.2 任务详细指导

在明确了任务要求和角色定位之后，本研究进一步阐述了完成该任务所需遵循的指导原则。类比于教育领域中，教师在考试前向学生传授解题技巧，本研究也在提示词设计中向大模型传达了执行任务的关键策略。以故事寓意理解任务为具体案例，研究者作为教育者的角色，

基于Oakhill等人 (Oakhill et al., 2014)提出的儿童在阅读理解中所需的能力，我们提炼出了三条实用的指导原则。这些原则被纳入到提示词中，以指导大模型在执行任务时的策略选择：

在注意事项中为模型提供任务的详细指导：

#完成任务的注意事项#：
1. 理解故事重点：要精准把握故事所要表达的重点，进而推断其蕴含的深层道理。
2. 深入理解而非仅看表面：不能仅仅围绕故事的表面内容进行分析。寓言故事的目的通常是以简单的故事传递深层的人生或道德理念，因此，需要超越故事的字面意义，挖掘更深的含义。
3. 确保答案选项与故事的关联性：正确的选项必须与故事内容紧密相关，且需要围绕故事的核心思想进行论述，避免选择那些与故事无关或偏离故事主旨的答案。

4.3 任务标准定义

在机器翻译任务中，Li等人 (Li et al., 2022b)提出可以通过修改提示词来使模型生成的翻译中包含某些词汇或者风格更符合要求。基于此，本研究进一步将评价标准纳入提示词设计之中，以期优化模型输出的质量和风格。参考Yang等人 (Yang and Klein, 2021)在可控文本生成中从主题符合程度、文本质量以及多样性三方面来全面地评价受控文本的质量，我们也提出了对任务完成程度的评价角度。以故事寓意理解任务为例，我们认为准确性、深度、关联性、适宜性与一致性是衡量任务完成程度的五个维度，因此我们将其整合进提示词，旨在引导大模型生成更符合预期的翻译结果：

在提示词中明确任务完成的标准：

#完成任务的标准#：
1. 准确性：所选答案必须精确反映故事的核心寓意或教训。
2. 深度：答案需要体现对故事深层次意义的理解。
3. 关联性：选择的答案必须与故事情节和主题直接相关，且能恰当地体现故事的教育意义。
4. 适宜性：确保所选寓意适合儿童的认知水平，并能为其提供有价值的教育意义。
5. 一致性：在类似的测试场景中，所选答案应保持一致的评判标准和解释逻辑。

4.4 融入思维链与少样本策略

对于多步骤的推理问题，可以通过思维链技术让大模型将较为复杂的问题分解成可以一步步解决的子问题，然后再依次求解来提高模型的推理效果。OpenAI的研究人员发现大模型的推理能力能够通过思维链获得较大的提升，在与运动有关的常识推理上，运用了思维链的PaLM (Chowdhery et al., 2023)表现甚至超过了运动爱好者。

因此，在本研究中，我们采用的提示词设计融合了少样本学习和思维链策略：首先，我们从基线模型在训练集中表现不佳的题目中精选了一个具有代表性的例子。我们认为，挑选此类题目将激励模型进行更深入的思考。其次，受到思维链理论的启发，我们认为在多项选择题中对每个选项进行详尽解释是一种有效的问题分解策略。基于这一理念，我们决定利用GLM-4模型，结合问题和答案，对这道题的四个选项进行详尽分析，并将其整合为提示词的一部分，旨在提升模型的推理和决策能力：

在提示词中融入思维链和少样本学习的策略：

#示例#：
##寓言故事##：
（寒鸦与鸽子）寒鸦看见一群不愁吃喝的鸽子舒适地住在鸽舍里，便将自己的羽毛全都涂成白色，跑到鸽舍里，与他们一起过活。寒鸦一直不敢出声，鸽子便以为他也是只鸽子，允许他在一起生活，可是，有一次，他不留心，发出了一声叫声，鸽子们立刻辨认出了他的本来面目，将他啄赶出来。寒鸦在鸽子那里再也吃不到食了，只好又回到他的同类那里。然而他的毛色与以前不同了，寒鸦们不认识他，不让他与他们一起生活。这样，这只寒鸦因想贪得两份，最后却一份都没得到。
##问题##：
下列哪个选项最符合故事说明的寓意?
A．不要为了一时的利益而放弃自己的原则和尊严。
B．贪婪只会让人得到短暂的快乐，却失去长久的幸福。
C．虚伪和伪装最终会被识破，因而失去信任和尊重。
D．真诚和坦率的态度才是与他人建立真正联系的关键。
##答案##：选B
##分析##：选项B是最符合这个寓言故事的寓意，因为它直接指出了贪婪行为所带来的后果。故事中的寒鸦因为追求短期的利益而伪装自己，最终不仅失去了新的生活环境，也被自己原本的同类排斥。这显示了寒鸦的贪婪导致了短暂的满足后的长期不幸，体现了贪婪可能带来的瞬间快乐和长远的失落。因此，选项B“贪婪只会让人得到短暂的快乐，却失去长久的幸福”最准确地反映了故事的主旨。其他选项虽然也可以从某种程度上解释故事中的某些方面，但没有选项B那样直接和全面地反映了故事的核心教训。选项A侧重于原则和尊严，选项C关注于伪装和识破，选项D强调真诚和坦率，而B选项则是直接指向了故事的主题——贪婪的后果，这是故事最核心的寓意。因此，B选项是对这个故事寓意最恰当的表达。

4.5 运用分割符

在本研究中，提示工程的设计原则贯穿于我们提示词的构建过程。OpenAI在其发布的提示词设计指南中建议，应恰当地使用分隔符来区分输入的不同部分，比如使用三重引号（”””标题”””）来明确标识。此外，为了便于后续处理中答案的提取，我们遵循提示工程的原则设计了格式，引导模型首先输出答案，随后提供相应的分析：

在提示词中善用分隔符引导模型：

##寓言故事##：题目：寒鸦与鸽子
故事：寒鸦看见一群不愁吃喝的鸽子舒适地住在鸽舍里.....
##问题##：
问题：下列哪个选项最符合故事说明的寓意?
选项：
A．不要为了一时的利益而放弃自己的原则和尊严。
B．贪婪只会让人得到短暂的快乐，却失去长久的幸福。
C．虚伪和伪装最终会被识破，因而失去信任和尊重。
D．真诚和坦率的态度才是与他人建立真正联系的关键。
##答案##：
##分析##：

5 实验

5.1 提示词设计与模型选择

我们按照在第4章提供的方法，针对常识推理和寓意理解这两个任务，精心构建了相应的提示词。由于我们在提示词设计中融入了思维链技术，因此对模型的规模提出了一定的要求：根据Wei等人 (Wei et al., 2022b)的发现，思维链提示是一种取决于模型尺度的涌现能力 (Wei et al., 2022a)。对于大多数参数量小于10B的小型模型，思维链提示会导致模型性能的损害。只有在与参数量较大的模型一起使用时才会产生性能提升。我们在一些模型上进行了评测，包括ERNIE-speed-128k、deepseek-chat、Yi-34B-chat、glm4-air以及GPT-4。我们的实验结果如下表所示：

队伍	常识推理	寓意理解	总计
Baseline	0.688	0.561	0.612
Team-1	0.865	**0.744**	**0.793**
Team-2	0.834	0.734	0.774
RENIE-speed-128k	0.657	0.562	0.600
deepseek-chat	0.826	0.645	0.717
Yi-34B-chat	0.727	0.602	0.652
glm4-air	0.768	0.412	0.554
Our Method（GPT4）	**0.869**	0.708	0.773

Table 2: 封闭数据的赛道一评测提交结果对比

根据Table 2所展示的数据，GPT-4模型结合本研究采用的方法在赛道一的两个子任务中均展现出了优秀的性能。特别是在常识推理任务上，我们的方法在赛道一上获得了第一名的成绩，这充分地证明了我们的方法的有效性。

5.2 实验结果分析

为了深入探究在本研究所提出的提示词中哪些组成部分对任务性能具有显著影响，本研究进一步开展了消融实验。该实验旨在系统地评估和比较各个组成部分对模型性能的具体贡献。消融实验涵盖了以下关键要素：任务与角色（TR）、详尽的指导（DG）、任务的标准（CT）以及学习样本（SL）。以下是消融实验的结果：

提示词	常识推理	寓意理解	总计
Our Prompt	**0.869**	0.708	0.773
-TR	0.861	**0.721**	**0.777**
-DG	0.856	0.705	0.766
-CT	0.865	0.715	0.775
-SL	0.832	0.688	0.745

Table 3: 消融实验结果对比

由Table 3的实验结果，我们得到了以下结论：

(1) 针对常识推理任务，本研究设计的提示词在实验中取得了显著的高分，这一结果表明，所设计的提示词能够有效地促进模型对常识知识的理解和应用。提示词的设计在提升模型对常识推理任务的处理能力方面发挥了关键作用；

(2) 在寓意理解任务上，本研究设计的提示词得分略低于“-TR”以及“-CT”。经分析，这一现象可能源于提示词中角色和评价标准的定义过于复杂，导致模型未能充分集中注意力于深层语义的理解，从而影响了其性能表现；

(3) 在常识推理和寓意理解两个子任务中，“-SL”的得分普遍较低，这一结果突显了在提示词中融入任务样例的必要性。这表明，任务样例的加入对于提升模型在相关任务上的表现具有显著影响。

第二十三届中国计算语言学大会论文集，第336页-第345页，太原，中国，2024年7月25日至28日。
卷3：评测报告

6 总结

本文针对儿童故事常识推理与寓意理解评测赛道一的两个任务，提出了一种基于提示工程的提示词构造方法。我们首先提出了一种融合提示工程和思维链的通用提示词构建框架，然后针对具体任务对提示词模板进行了调整，最后结合语言模型使用这些提示词进行结果预测。在本次评测中，我们的方法在赛道一的封闭数据条件下取得了第三名的成绩，证明了该方法的有效性。通过这项研究，我们展示了提示工程在提升大语言模型性能方面的重要性，尤其是在处理复杂自然语言理解任务时。此外，我们的方法不仅强调了提示词设计的关键性，还展示了在实际应用中对任务需求和模型特性的深刻理解和灵活运用的必要性。然而，我们也意识到当前提示词设计方法存在的局限性，尤其是在需要人工设计模板且步骤较为繁琐的情况下。展望未来，我们计划继续改进提示词构建方法，以期实现更高效的自动化设计流程。同时，我们也期待将这一方法应用于更广泛的自然语言处理任务中，以探索其更深远的应用潜力。我们相信，随着技术的不断进步和研究的深入，提示工程将为自然语言处理领域带来更多创新和突破。

参考文献

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*.

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

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

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

María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Miguel Ángel Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. *arXiv preprint arXiv:2204.09438*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Xiao Li, Gong Cheng, Ziheng Chen, Yawei Sun, and Yuzhong Qu. 2022a. Adalogn: Adaptive logic graph network for reasoning-based machine reading comprehension. *arXiv preprint arXiv:2203.08992*.

Yafu Li, Yongjing Yin, Jing Li, and Yue Zhang. 2022b. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590.

Chen Ling, Xuchao Zhang, Xujiang Zhao, Yifeng Wu, Yanchi Liu, Wei Cheng, Haifeng Chen, and Liang Zhao. 2023. Knowledge-enhanced prompt for open-domain commonsense reasoning. In *1st AAAI Workshop on Uncertainty Reasoning and Quantification in Decision Making*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

第二十三届中国计算语言学大会论文集，第336页–第345页，太原，中国，2024年7月25日至28日。
卷3：评测报告

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Farhad Moghimifar, Lizhen Qu, Yue Zhuo, Mahsa Baktashmotlagh, and Gholamreza Haffari. 2020. Cosmo: Conditional seq2seq-based mixture model for zero-shot commonsense question answering. *arXiv preprint arXiv:2011.00777*.

Jane Oakhill, Kate Cain, and Carsten Elbro. 2014. *Understanding and teaching reading comprehension: A handbook*. Routledge.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Chris Tanasescu, Vaibhav Kesarwani, and Diana Inkpen. 2018. Metaphor detection by deep learning and the place of poetic metaphor in digital humanities. In *The thirty-first international flairs conference*.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*.