Software Engineering and Applications 软件工程与应用, 2025, 14(2), 234-244
Published Online April 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.142022

Hans汉斯

基于Agent服务的ChatGPT处理多方对话任务

万　静

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年2月15日；录用日期：2025年3月27日；发布日期：2025年4月8日

摘　要

随着大型语言模型(LLMs)规模的显著扩展，它们在众多自然语言处理(NLP)任务中展现出了卓越的零样本学习能力，能够在无需针对特定数据集进行预训练的情况下执行任务。这些模型在多个语言相关领域，包括搜索引擎，显示了显著的泛化能力。然而，LLMs在处理多方对话(MPC)——一个涉及多个参与者进行复杂信息交流的场景——的能力尚未得到充分探索。本文旨在评估生成型LLMs，如ChatGPT和GPT-4，在MPC领域的应用潜力。我们通过在两个包含四个代表性任务的MPC数据集上对ChatGPT和GPT-4进行实证分析，具体评估了它们的零样本学习能力。研究结果表明，ChatGPT在多个MPC任务上的表现仍有提升空间，而GPT-4的结果则显示出积极的发展前景。此外，我们尝试通过整合MPC结构和代理机制，涵盖说话者架构以及与四个任务相关的代理方法，以增强模型性能。本研究全面评估了生成型LLMs在多方对话中的应用，并深入分析了构建更高效、更强大的MPC代理的理念与策略，为该领域的进步提供了新的洞见。最后，我们指出了LLMs在MPC应用中面临的挑战，特别是在解析复杂信息流和生成风格一致的响应方面，这些挑战可能会影响模型的实际应用效果。

关键词

多方对话，大语言模型，ChatGPT，智能代理

Agent-Based ChatGPT for Multi-Party Conversation Task Processing

Jing Wan

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science & Technology, Shanghai

Received: Feb. 15th, 2025; accepted: Mar. 27th, 2025; published: Apr. 8th, 2025

Abstract

With the significant expansion of Large Language Models (LLMs), they have demonstrated remarkable

文章引用: 万静. 基于 Agent 服务的 ChatGPT 处理多方对话任务[J]. 软件工程与应用, 2025, 14(2): 234-244.
DOI: 10.12677/sea.2025.142022

zero-shot learning capabilities across various NLP tasks, performing well without specific dataset pre-training. These models have shown strong generalization in language-related fields like search engines. However, their ability in Multi-Party Conversation (MPC) scenarios, where multiple participants exchange complex information, remains underexplored. This paper evaluates the application potential of generative LLMs, such as ChatGPT and GPT-4, in the MPC domain. Through empirical analysis on two MPC datasets with four representative tasks, we assess their zero-shot learning abilities. Results show that ChatGPT has room for improvement in various MPC tasks, while GPT-4 shows promising prospects. Additionally, we enhance model performance by integrating MPC structures and agent mechanisms, covering speaker architectures and agent methods related to the four tasks. This study comprehensively evaluates generative LLMs in MPC and analyzes strategies for building more efficient and powerful MPC agents, offering new insights for the field. Finally, we highlight challenges in LLMs' MPC applications, especially in parsing complex information flows and generating consistent responses, which may affect practical application.

Keywords

Multi-Party Conversation, Large Language Models, ChatGPT, Intelligent Agent

1. 引言

大型语言模型(LLMs)，如 ChatGPT [1]和 GPT-4 [2]，在自然语言处理(NLP)领域开启了一个新时代。这些模型通过在大量文本语料库上预训练，随后进行微调以确保遵循人类指令，显著提升了在语言理解、生成、交互和推理方面的性能[3] [4]。这些模型的高性能已被广泛记录[5] [6]。LLMs 能够处理各种 NLP 任务，通过适当的提示(prompt)和微调，这些模型能够根据指令执行新任务，这被视为人工通用智能发展的重要一步[7]。然而，尽管在某些情况下取得了合理的性能，LLMs 在零样本学习中仍然容易出错，且提示的格式对性能有重大影响。例如，添加“让我们一步一步地思考”这样的提示可以显著提高 ChatGPT 的性能[8]。这表明，提示的结构及内容，对 LLMs 性能的重要性。

多方对话阅读理解(MPC) [9]是自然语言处理领域中一个极具挑战性的研究方向。它模拟了真实世界中人类之间的互动交流，要求模型能够理解并处理多个参与者之间的复杂对话内容。近年来，已有大量研究致力于探索 LLMs 在各种 NLP 任务中的表现[10]。然而，尽管取得了显著进展，LLMs 在处理 MPC 方面的潜力仍未得到充分挖掘。为了填补这一研究空白，本研究在两个不同的 MPC 数据集上实施了四个任务(包括情感检测、说话人识别、响应选择和响应生成)，以系统地评估 LLMs (如 ChatGPT 和 GPT-4)在处理 MPC 方面的表现。实验结果表明，在 EmoryNLP 和 MELD 数据集上，ChatGPT 和 GPT-4 均能取得与监督学习方法相当的性能。在加入说话者架构后[11]，获得了一定性能的提升。然而，现有方法在处理复杂的多方对话场景时仍存在一些局限性。首先，这些方法往往依赖于静态的对话结构，难以适应动态变化的对话环境，导致在处理多变的对话参与者和意图时表现不佳。

针对多方对话中存在的缺陷，ChatGPT 的 Agent 功能提供了一种灵活的解决方案。Agent 功能是一种基于人工智能的智能代理模块，能够通过动态交互实时适应对话中的变化，从而增强对多方参与者意图的理解和响应生成的准确性。此外，Agent 功能能够更好地捕捉上下文中的复杂语义关系，生成更为连贯和自然的对话内容。在本研究中，我们将 ChatGPT 的 Agent 功能集成到我们的 MPC 结构合并策略中。

通过利用 Agent 的动态交互能力，模型能够更精准地捕捉多方参与者的意图，并生成更符合对话情境的响应。实验结果表明，使用 Agent 功能后，模型在 EmoryNLP 和 MELD 数据集上的性能显著提升，验证了 Agent 功能在提升 LLMs 处理 MPC 任务中的有效性。

总之，我们在本文中的贡献是三方面的：1) 进行了一项探索性研究，以检查 ChatGPT 和 GPT-4 在零样本学习情况下结合 agent 处理 MPC 的性能。这是同类研究中第一次研究这些 LLMs 结合 agent 后在 MPC 场景中的表现。2) 提出了一种 MPC-agent 方法，该方法增强了 ChatGPT 和 GPT-4 管理 MPC 的性能。该策略将 MPC 的说话者和 Agent 结构整合到 LLMs 中，从而大大提高了性能。3) 我们深入研究了 LLMs 在处理 MPC 方面的潜力，并阐明了未来研究中需要解决的挑战。这些讨论为进一步提高 LLMs 在 MPC 场景中的有效性奠定了基础。

2. 相关工作

2.1. 大语言模型

统近年来，LLMs 的开发取得了显著进展，GPT-3 [3]、Palm [12]、LLaMA [13]、ChatGPT [1]和 GPT-4 [2]等模型的出现，标志着自然语言处理领域的一个重要里程碑。这些模型不仅在语言生成和理解方面表现出色，还展示了新兴的能力，包括情境学习、数学推理和常识推理[14]。

最近的研究工作侧重于指令学习，主要通过以下两种方式提升 LLMs 的性能：一是生成高质量的指令数据[15] [16]，二是通过指令增强 LLMs [17] [18]。这些方法不仅提高了模型对复杂指令的遵循能力，还推动了 LLMs 在多种自然语言处理任务中的应用。

2.2. 多方对话阅读理解

处理多方对话阅读理解具有挑战性，因为对话中的每个话语都具有说话者角色的额外属性，说话者角色转换引起的复杂话语依赖关系打破了普通非对话文本的连续性。构建 MPC 系统的现有方法通常可以分为基于检索的方法和基于生成的方法。基于检索的方法主要集中在如何从大量候选回复中选择最合适的回答，而基于生成的方法则侧重于直接生成自然语言回复。这些方法在处理多方对话时面临诸多挑战，例如说话人角色的转换、话语之间的复杂依赖关系以及对话的动态性。

2.3. 处理 NLP 任务的大语言模型

LLMs 在 NLP 任务中的应用取得了显著进展，展示了其在多种任务中的强大能力。这些任务包括零样本学习、语言生成、信息检索、多方对话以及多模态对话等。例如，研究[5]对 ChatGPT 的零样本学习能力进行了评估，涉及推理、自然语言推理和摘要等 7 个代表性任务类别，揭示了模型在无监督场景下的强大泛化能力。研究[6] [19]探索了 ChatGPT 作为自然语言生成评估器的潜力，而研究[10]则针对信息检索中的相关性排序问题提出了基于 LLMs 的优化方案。

值得注意的是，尽管 LLMs 在多领域应用中表现突出，其局限性也逐渐显现。研究[20]通过系统性实验揭示了 ChatGPT 在问答任务中生成真实答案的不足，并据此提出提升输出真实性的方法论。

2.4. 大语言模型中的 Agent 功能

LLMs 中的 Agent 功能作为一种新兴的交互式智能体，逐渐成为 NLP 领域的重要研究方向。Agent 功能的核心在于其能够通过动态交互实时适应对话环境，精准理解用户意图并生成高质量响应。其关键组成部分[21]包括档案(Profile)、记忆(Memory)、规划(Planning)和行动(Action)，这些部分协同工作，使 Agent 能够完成复杂任务。

Agent 功能通常被设计为一个智能体，能够自主地与用户进行交互，并根据对话的上下文和历史信息做出决策。其主要特点包括动态交互能力、上下文理解和意图识别。Agent 能够实时响应用户的输入，并根据对话的动态变化调整其行为，这种能力使其在处理复杂的多轮对话时表现出色。此外，Agent 通过记忆功能整合历史需求与新问题，补充上下文信息后调用大模型，从而生成连贯、准确的回复。

3. 方法论

针对每一项任务，首先向大型语言模型(LLM)呈现一系列按时间顺序组织的多方会话内容。这些会话内容模拟了实际应用场景中多方交互的场景，为模型提供了丰富的上下文信息。随后，指示 LLM 基于特定于任务的提示来完成相应任务。这些任务提示是经过精心设计的，旨在引导模型准确地输出符合任务要求的结果。在此过程中，为确保输出结果的稳定性，将温度参数(temperature)设置为 0。在语言模型中，温度参数是一个控制输出随机性的超参数。当温度值为 0，时模型会输出最确定性的结果，即概率最高的答案，从而最大限度地减少不必要的词汇，避免对结果造成混淆。这种设置有助于提高模型输出的准确性和一致性。此外，为了提供更全面的解释，部分任务需要进行扩展说明，如图 1 中所示。这些扩展说明旨在对任务的背景、目标和具体要求进行更详细的阐述，以便更好地指导模型完成任务。

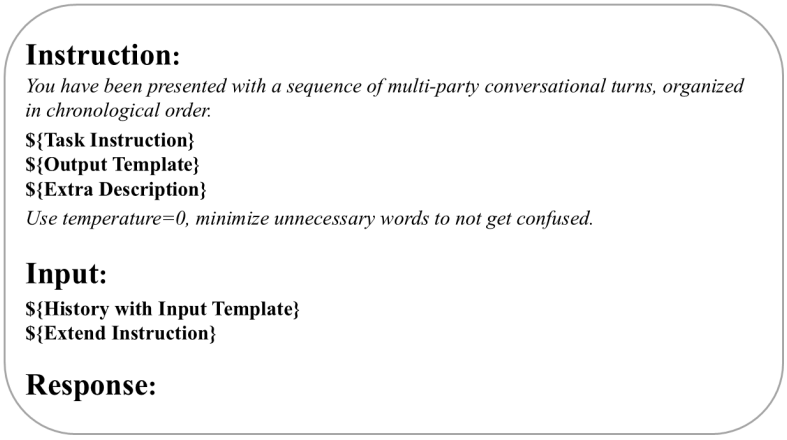

Figure 1. Task completion prompt template
图 1. 完成任务的提示模板

3.1. 特定任务的提示词

针对每个任务设计了不同的说明，以指导 LLM 完成任务。

情感检测(ED)：LLM 的任务是预测每个话语的情感，任务指令是“请使用以下 n 个标签评估对话中每个话语的情感：{...}”。以及输出模板“输出格式必须是：#{num} - {outdutation} // {emotion}”这里，n 是情感标签的数量，并且{...}是情感标签的列表。对话历史被形式化为“#{num} - {utteration}"。

说话人识别(SI)：LLM 的任务是预测最后话语的说话人，任务指令“请指出最后一句话的说话者。”和输出模板“输出格式应该只有一个说话者。”。与说话者的对话历史被形式化为“#{num} - {speaker}：{utterance}”。

响应选择(RS)：LLM 的任务是从候选人中选择最合适的响应，任务指令为“你的任务是从候选人集合中选择最合适的响应”。和输出模板“输出格式必须是：#{num} - {utterance}”。候选集被形式化为“#{num} - {utterance}”。因此，具有输入模板的历史被形式化为“Dialogue History: {conversation turns} Candidates: {candidates}”。如图 2 所示。

Instruction:

You have been presented with a sequence of multi-party conversational turns, organized in chronological order. Your task is to select the most appropriate response from the candidate set. The output format is "#{i} -- {speaker}: {utterance}".
Use temperature=0, minimize unnecessary words to not get confused.

Input:

Dialogue History:
Jade: Oh, Bob, he was nothing compared to you. I had to bite my lip to keep from screaming your name.
Chandler: Well, that makes me feel so good.
Jade: It was just so awkward and bumpy.
Ross: Bumpy?
Chandler: Well, maybe he had some kind of uh, new, cool style, that you're not familiar with.
Chandler: And uh maybe you have to get used to it.
--
Candidates:
#0 -- Jade: They're just ten blocks away, if I run, I can make it.
#1 -- Jade: Okay then.
#2 -- Jade: Julie! Julie, isn't that great? I mean, isn't that just kick- you-in-the-crotch, spit-on-your-neck fantastic?
#3 -- Jade: Like you, I don't even know where you work?
#4 -- Jade: I'll get everybody else , finally we can start celebrating my— I'm sorry, uh apparently I've opened the door to the past.
#5 -- Jade: Well there really wasn't much time to get used to it, you know what I mean?
#6 -- Jade: Two people screwing in there if you want to check that out.
#7 -- Jade: Come on! The time we were all waiting in line for
#8 -- Jade: No, I mean happy.

Response:

#5 -- Jade: Well there really wasn't much time to get used to it, you know what I mean?

Figure 2. Prompt and output for speaker-structured response selection task
图 2. 采用说话人结构的响应选择任务的提示和输出

响应生成(RG)LLM 的任务是生成响应，任务指令为“您的任务是生成最合适的响应”。不需要提供输出模板，因为生成任务是自由格式的。

其中说话者架构[11]被并入 LLM 以帮助理解话语。具体地，用“{speaker}：{utterance}”替换具有“{utterance}”的提示，以向 LLM 通知每个话语的说话者。

3.2. Agent 结构并入

每个任务都有一个对应可调用的功能函数，组成功能列表。Chatgpt-4 会根据提示词自动识别是哪个任务并且调用功能列表的相应函数去协助完成任务，四个功能函数设计具体如下。

情感检测(ED)：我们采用了经过微调的 DistilBERT 模型(distilbert-base-uncased-finetuned-sst-2-english)。该模型在大规模的情感分类数据集上进行了训练和微调，能够以高准确度识别出文本中的情感倾向(如积极、消极等)。将情感分析的结果作为额外的上下文信息，集成到提示词中获得最终输出结果，如图 3 所示。

Instruction:

You have been presented with a sequence of multi-party conversational turns, organized in chronological order. The multi-party dialogue turns have been annotated with 'positive' or 'negative' labels by the emotion detection model after each utterance. Please refer to the annotations and incorporate your own judgment. Please evaluate the emotions of each utterance in the dialogue using the following 7 labels: ('neutral', 'joyful', 'peaceful', 'powerful', 'scared', 'mad', 'sad'). The output format must be: #8 -- Monica Geller: That's right. // positive. If the speaker is 'Unknown', omit the positive, e.g., #8 -- Monica Geller: That's right.. Use temperature=0, minimize unnecessary words to not get confused.

Input:

#0 -- Monica Geller: Oh my God! // positive
#1 -- Paul the Wine Guy: I know, I know, I'm such an idiot. I guess I should have caught on when she started going to the dentist four and five times a week. I mean, how clean can teeth get? // negative
#2 -- Monica Geller: My brother's going through that right now, he's such a mess. How did you get through it? // negative
#3 -- Paul the Wine Guy: Well, you might try accidentally breaking something valuable of hers, say her- // negative
#4 -- Monica Geller: -leg? // negative
#5 -- Paul the Wine Guy: That's one way! Me, I- I went for the watch. // negative
#6 -- Monica Geller: You actually broke her watch? Wow! The worst thing I ever did was, I-I shredded by boyfriend's favorite bath towel. // negative
#7 -- Paul the Wine Guy: Ooh, steer clear of you. // positive
#8 -- Monica Geller: That's right. // positive

Response:

#0 -- Monica Geller: Oh my God! // scared
#1 -- Paul the Wine Guy: I know, I know, I'm such an idiot. I guess I should have caught on when she started going to the dentist four and five times a week. I mean, how clean can teeth get? // sad
#2 -- Monica Geller: My brother's going through that right now, he's such a mess. How did you get through it? // sad
#3 -- Paul the Wine Guy: Well, you might try accidentally breaking something valuable of hers, say her- // mad
#4 -- Monica Geller: -leg? // mad
#5 -- Paul the Wine Guy: That's one way! Me, I- I went for the watch. // mad
#6 -- Monica Geller: You actually broke her watch? Wow! The worst thing I ever did was, I-I shredded by boyfriend's favorite bath towel. // mad
#7 -- Paul the Wine Guy: Ooh, steer clear of you. // joyful
#8 -- Monica Geller: That's right. // joyful

Figure 3. Prompt and output for agent-structured emotion detection task
图 3. 采用 agent 结构的情感检测任务的提示和输出

说话人识别(SI)：我们使用了说话人频率计数的方法。这一设计带来的优势在于，它不仅提高了系统对于说话人身份的识别能力，还能为后续的对话管理提供重要的信息，使得系统能够更好地进行个性化对话。

响应选择(RS)：我们构造了响应选择功能功能。它通过计算对话历史与候选响应之间的相似度来选择最合适的响应。这个功能利用先进的算法来评估候选响应与当前对话上下文的匹配程度，从而选择出最最恰当的回应。

响应生成(RG)：我们构建了响应生成功能。它专注于从对话历史中提取关键信息，如主要人物、场景和主题。这个功能通过深入分析对话内容，理解对话的深层含义和上下文，生成与对话内容紧密相关的回复。

4. 实验

4.1. 数据集

EmoryNLP [22]数据集是一个基于经典美剧《老友记》构建的情感标注数据集，涵盖了 97 集、897 个场景和 12,606 条话语。该数据集的情感标注体系借鉴了 Willcox [23]的感觉轮理论，用了七种情绪标签：悲伤(sad)、愤怒(mad)、害怕(scared)、强大(powerful)、平静(peaceful)、快乐(joyful)和中性(neutral)。这种情感分类体系能够较为全面地覆盖人类情感的多样性，为情感分析提供了丰富的语义信息。

MELD [24]数据集每条话语都被标注为七种情绪之一：愤怒(Anger)、厌恶(Disgust)、悲伤(Sadness)、喜悦(Joy)、中立(Neutral)、惊讶(Surprise)和恐惧(Fear)。这种全面的标注方式有助于深入的情感分析。表 1 列出了我们实验中评估的两个数据集的统计数据。

Table 1. The statistics of the dataset evaluated in this paper
表 1. 本文评估的数据集的统计

数据集	Train	Dev	Test
EmoryNLP [21]	659	89	79
MELD [23]	1039	114	280

4.2. 基线模型的选择

(1) BERT [25]是一个双向语言表示模型，可以针对各种 NLP 任务进行微调。(2) GPT-2 [26]是一个单向的预训练语言模型。(3) BART [27]是一个去噪自动编码器，使用标准的基于 transformer 的架构，通过使用任意噪声函数破坏文本并学习重建原始文本来训练。(4) SPCL-CL-ERC [28]为会话中的情感识别任务引入了一种新的监督原型对比学习损失函数，通过对比学习的媒介解决了不平衡分类产生的问题，避免了大批量的必要性。它是 ED 任务上 EmoryNLP 和 MELD 数据集的 SOTA。(5) ChatGPT [1]，通过加入监督微调和来自人类反馈的强化学习方法确保模型和人类指令之间的无缝同步。(6) GPT-4 [2]是一个大规模的多模式模型，可以接受图像和文本输入并生成文本输出，在各种专业和学术基准上表现出人类水平的性能。

4.3. 实现细节

所有监督模型都使用 AdamW [29]方法进行训练。学习率初始化为 6.25e−5，并线性衰减到 0。批量大小设定为 128。模型在 10 个周期内训练。对于 ChatGPT 和 GPT-4，我们分别使用 OpenAI 提供的 API 端点 gpt-3.5 turbo-0301 和 gpt-4-0314。

4.4. 评价指标

为了评估 ED 任务，我们采用了加权 F1 评分，它是精确度和召回率的调和均值。为了评估 SI 任务，我们采用了准确性 ACC。为了评价 RS 任务，我们采用了 R10@1，即从 10 个候选者中选出的第一个正确答案的百分比。为了评估 RG 任务，我们采用了标准的基于字符串相似性的度量标准 SacreBLEU、ROUGE 和 METEOR。对于所有指标，值越高越好。

4.5. MPC 理解评价结果

如表 2 所示，EmoryNLP 和 MELD 对 ED 任务的 SOTA 来自研究[28]。表格中列出了监督模型(Supervised)和 LLMs 在多个任务上的性能，包括 ED、SI 和 RS。这些任务分别评估了模型在情感检测、

说话人识别和回应选择上的能力。在 SI 任务中，必须注意，在不提供明确的说话人信息的情况下，说话人信息检测是无法实现的。因此，我们在 SI 任务中对 ChatGPT 和 GPT-4 的评估仅限于提供说话者信息的情况。

在 EmoryNLP 数据集上，GPT-4 在 ED 任务中的 F1 得分为 39.38，高于 BERT 的 34.76 和 ChatGPT 的 37.16；而在结合 Agent 结构后，其性能进一步提升至 43.87，显示了 Agent 结构在捕捉对话动态的重要作用。在 SI 任务中，GPT-4 结合说话人结构和 Agent 结构后的准确率达到 69.27，远超 BERT 的 47.82 和 ChatGPT 的 51.42。在 RS 任务中得分同样优异。

在 MELD 数据集上，GPT-4 的表现同样突出。w/. Speaker & Agent 的结果在三个任务中均为 SOTA。在任务 RS 中，GPT-4 添加 Agent 结构，其性能进一步提升至 62.33，同比增长了 10.19%，进一步验证了代理结构的有效性。

Table 2. Evaluation results of the MPC comprehension task. Bold numbers indicate the best performance. Empty cells indicate that the values are not computable. The SOTA for the ED task is SPCL-CL-ERC
表 2. MPC 理解任务的评价结果。粗体数字表示结果达到最佳性能。空单元表示它们的不可计算性。ED 任务的 SOTA 是 SPCL-CL-ERC

模型 \ 任务	EmoryNLP			MELD		
	ED (F1)	SI (ACC)	RS (R10@1)	ED (F1)	SI (ACC)	RS (R10@1)
Supervised						
BERT	34.76	47.82	48.68	61.32	44.18	48.13
SOTA	40.94	-	-	67.25	-	-
LLMs						
ChatGPT	37.16	-	29.11	58.32	-	36.42
w/. Speaker	38.50	51.42	34.21	60.90	57.67	39.17
GPT-4	39.38	-	44.30	62.32	-	53.73
w/. Speaker	41.63	64.28	50.00	64.18	78.60	57.46
w/. Agent	43.87	-	56.24	67.93	-	59.21
w/. Speaker & Agent	**45.54**	**69.27**	**59.83**	**69.64**	**81.26**	**62.33**

4.6. MPC 生成的评价结果

如表 3 所示，对于监督模型，如 GPT-2 和 BART，它们在测试集上的表现通常优于基础 LLMs，这可能归因于它们在训练过程中对特定任务的优化。在指标上均取得了较高的分数，显示出其在生成与参考文本相似度较高的文本方面的优势。

当 GPT-4 模型结合了不同的信息结构，如 Speaker (发言者)、Agent (代理)以及两者结合时，其在不同测试集和评估指标上的表现呈现出明显的差异。具体来说，GPT-4 w/ Speaker 在 MELD 测试集上的 S-BLEU 分数提升至 0.9666，ROUGE_L 分数为 8.89，METEOR 分数为 13.06，这表明加入发言者信息对于提高模型在该测试集上的性能是有益的。当进一步结合 Agent 信息，即 GPT-4 w/ Agent 时，模型在 MELD 测试集上的 S-BLEU 分数进一步提高到 1.1021，ROUGE_L 分数为 9.97，METEOR 分数为 14.24，显示出代理信息的加入对于提升模型性能同样重要，尤其是在 METEOR 指标上。

最引人注目的是，当 GPT-4 同时结合了 Speaker 和 Agent 信息，即 GPT-4 w/ Speaker & Agent 时，其在两个测试集上的表现均达到了最佳，S-BLEU 分数为 1.2679，ROUGE_L 分数为 11.18，METEOR 分数

高达 14.67。这一结果验证了对话者和接收者信息整合对于提高模型性能的重要性，并且能够显著提升模型在多个评估指标上的表现。这些发现表明，通过整合更多的上下文信息，可以有效地增强模型在自然语言处理任务中的表现，尤其是在生成对话响应方面。

Table 3. Evaluation results of the MPC generation task. Bold numbers indicate the results achieved the best performance. S-BLEU stands for SacreBLEU. Empty cells indicate that the values are not computable
表 3. MPC 生成任务的评估结果。粗体数字表示结果达到了最佳性能。S-BLEU 是 SacreBLEU 的缩写。空的单元格表示它们的不可计算性

模型 \ 指标	EmoryNLP			MELD		
	S-BLEU	ROUGE-L	METEOR	S-BLEU	ROUGE-L	METEOR
Supervised						
GPT-2	0.6175	7.90	10.26	0.5160	6.01	7.74
BART	0.7009	8.63	11.86	1.0757	8.64	10.37
SOTA	-	-	-	-	-	-
LLMs						
ChatGPT	0.5358	9.03	11.30	0.9059	7.13	8.63
w/. Speaker	0.3082	8.95	11.45	0.9159	8.17	9.86
GPT-4	0.4608	8.60	12.22	0.9301	8.93	12.43
w/. Speaker	0.9049	9.99	14.61	0.9666	8.89	13.06
w/. Agent	1.1643	11.33	16.81	1.1021	9.97	14.24
w/. Speaker & Agent	**1.3432**	**11.89**	**17.58**	**1.2679**	**11.18**	**14.67**

5. 总结

本文深入探讨了大型语言模型(LLMs)在生成多轮对话(MPCs)方面的能力，这是一个尚未被充分研究的领域。通过对 ChatGPT 和 GPT-4 在 EmoryNLP 和 MELD 两个流行 MPC 数据集上的评估，我们发现这些模型展现出了与监督训练模型相媲美的性能。特别是在整合了发言者和代理信息后，GPT-4 在多个任务中表现出显著的性能提升，这表明对话者和接收者信息的整合对于提高模型性能至关重要。

提示设计在塑造大型语言模型(LLMs)如 ChatGPT 和 GPT-4 在多轮对话(MPC)任务中的表现方面扮演着关键角色，其重要性不言而喻。精心构造的提示不仅能够显著影响模型的输出质量，还能充分挖掘模型在处理复杂对话时的潜力。特别是当涉及到整合 Agent 信息时，提示的结构化方式对于提升模型的性能和有效性尤为关键。

目前的提示架构可能尚未充分发挥 LLMs 在 MPC 任务中的理想能力，尤其是在利用 Agent 信息方面。Agent 信息的整合对于模拟真实对话场景中的交互动态至关重要，它能够帮助模型更好地理解对话的上下文和目的。因此，对提示设计进行进一步的优化和改进，不仅有可能释放 LLMs 更高的性能，还能充分释放 ChatGPT 和 GPT-4 在处理 MPC 任务方面的全部潜力。

探索和增强提示架构，以确保在 MPC 任务领域中最大限度地利用这些强大语言模型的功能，是未来研究的一个重要方向。通过精心设计提示，我们可以更好地引导模型利用 Agent 信息，从而在多轮对话生成任务中实现更自然、更准确、更高效的对话交互。这不仅能够提升用户体验，还能推动 LLMs 在更广泛的应用场景中发挥更大的作用。

参考文献

[1] Open AI (2022) Introducing ChatGPT.

[2] Open AI (2023) GPT-4 Technical Report. arXiv: 2303.08774.

[3] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., *et al*. (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, 6-12 December 2020, Virtual.

[4] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., *et al*. (2023) Sparks of Artificial General Intelligence: Early Experiments with GPT-4.

[5] Qin, C.W., Zhang, A., Zhang, Z.S., *et al*. (2023) Is ChatGPT a General-Purpose Natural Language Processing Task Solver?

[6] Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., *et al*. (2023). Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *Proceedings of the* 4*th New Frontiers in Summarization Workshop*, Singapore, December 2023, 1-11. https://doi.org/10.18653/v1/2023.newsum-1.1

[7] Goertzel, B. (2014) Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, **5**, 1-48. https://doi.org/10.2478/jagi-2014-0001

[8] Ouyang, L., Wu, J., Mishkin, P., Zhang, C., *et al*. (2022) Training Language Models to Follow Instructions with Human Feedback.

[9] Cui, L., Wu, Y., Liu, S., Zhang, Y. and Zhou, M. (2020) Mutual: A Dataset for Multi-Turn Dialogue Reasoning. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, 1406-1416. https://doi.org/10.18653/v1/2020.acl-main.130

[10] Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., *et al*. (2023) Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2023, 14918-14937. https://doi.org/10.18653/v1/2023.emnlp-main.923

[11] Tan, C., Gu, J. and Ling, Z. (2023) Is ChatGPT a Good Multi-Party Conversation Solver? Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 2023, 4905-4915. https://doi.org/10.18653/v1/2023.findings-emnlp.326

[12] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., *et al*. (2022) PaLM: Scaling Language Modeling with Pathways.

[13] Touvron, H., Lavril, T., Izacard, G., Martinet, X., *et al*. (2023) Llama: Open and Efficient Foundation Language Models.

[14] Wei, J., Tay, Y., Bommasani, R., Raffel, C., *et al*. (2022) Emergent Abilities of Large Language Models.

[15] Wang, Y.Z., Kordi, Y., Mishra, S., Liu, A., *et al*. (2023) Self-Instruct: Aligning Language Model with Self-Generated Instructions. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics*, *ACL* 2023, (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 13484-13508.

[16] Peng, B.L., Li, C.Y., He, P.C., *et al*. (2023) Instruction Tuning with GPT-4.

[17] Chung, H.W., Hou, L., Longpre, S., Zoph, B., *et al*. (2022) Scaling Instruction-Finetuned Language Models.

[18] Taori, R., Gulrajani, I., Zhang, T.Y., Dubois, Y., *et al*. (2023) Stanford Alpaca: An Instruction-Following Llama Model. https://github.com/tatsu-lab/stanford_alpaca

[19] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C. (2023) G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, December 2020, 2511-2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[20] Zheng, S., Huang, J. and Chang, K.C.-C. (2023) Why Does ChatGPT Fall Short in Answering Questions Faithfully?

[21] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., *et al*. (2024) A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science*, **18**, Article 186345. https://doi.org/10.1007/s11704-024-40231-1

[22] Zahiri, S.M. and Choi, J.D. (2018) Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. *The Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 44-52.

[23] Willcox, G. (1982) The Feeling Wheel. *Transactional Analysis Journal*, **12**, 274-276. https://doi.org/10.1177/036215378201200411

[24] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. and Mihalcea, R. (2019) MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 527-536. https://doi.org/10.18653/v1/p19-1050

[25] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers

for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, NAACL-HLT 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[26] Radford, A., Wu, J., Child, R., Luan, D., *et al*. (2019) Language Models Are Unsupervised Multitask Learners. Open AI Blog.

[27] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., *et al*. (2020) BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, Ju8ly 2020, 7871-7880. https://doi.org/10.18653/v1/2020.acl-main.703

[28] Song, X., Huang, L., Xue, H. and Hu, S. (2022) Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, December 2020, 5197-5206. https://doi.org/10.18653/v1/2022.emnlp-main.347

[29] Loshchilov, I. and Hutter, F. (2017). Fixing Weight Decay Regularization in Adam. ArXiv: 1711.05101.