arXiv:2305.04533v1 [cs.CL] 8 May 2023

提示式大语言模型作为长篇开放域对话的聊天机器人模块

Gibbeum Lee [1*] **Volker Hartmann** [1*] **Jongho Park**[1*] **Dimitris Papailiopoulos** [1,2] **Kangwook Lee** [1,2]
[1] KRAFTON [2] 威斯康星大学麦迪逊分校
{pirensisco, volker, jongho.park}@krafton.com

摘要

本文提出了 MPC（模块化提示聊天机器人），这是一种无需微调即可创建高质量对话代理的新方法。我们的方法利用预训练大语言模型（LLM）作为独立模块，并结合少样本提示、思维链（CoT）和外部记忆等技术，实现长期一致性与灵活性。人工评估结果表明，在开放域对话中，MPC 的表现可与经过微调的聊天机器人模型相媲美，因此它是创建一致且富有吸引力的聊天机器人的有效解决方案。

1 引言

具有数十亿参数的语言模型，例如 GPT-3（Brown 等，2020年）和 PaLM（Chowdhery 等，2022年），在许多自然语言处理任务上取得了当前最先进的性能。要将这些大语言模型（LLMs）微调用于开放域聊天机器人任务，可以使用能够代表目标领域的对话数据集。然而，由于更新具有数十亿参数的模型所带来的计算负担，以及对话领域数据的稀缺，为开放域聊天机器人微调 LLMs 可能颇具挑战。此外，微调会将模型限制在特定领域，从而降低其多功能性，并导致其丧失预训练期间获得的领域无关知识，正如Yang 和 Ma （2022年）所报告的那样。在不同数据集上进行多任务训练，如Roller 等（2021年）所提出的方法，可以解决多功能性问题，但也存在一些局限，例如需要数据来训练每项技能，以及难以确定开放域聊天机器人所必需的技能。事实上，Blenderbot3（BB3）中不断增加的聊天机器人模块（Shuster 等，2022年）表明，为每个新的聊天机器人模型进行微调时，数据和计算负担正在不断增加。

[*]贡献均等

有趣的是，一些 LLMs 具备上下文学习（ICL）的能力（Nye 等，2022年；Wei 等，2022b；Lewkowycz 等，2022年；Wei 等，2022a；Zhou 等，2022年；Dasgupta 等，2022年；Chung 等，2022年）。这种能力使模型能够根据简短的指令和少量示例，快速适应并执行特定任务，而无需额外微调。利用这一点，可以创建开放域聊天机器人：向 LLM 提供一个描述开放域对话所需任务的提示，以及解决该任务的少量示例，使其能够生成与当前对话相关的信息。

我们的贡献我们提出了一种无需微调即可创建高质量对话代理的新方法。我们提出的聊天机器人 MPC（**模块化提示聊天机器人**）利用开源预训练语言模型，提高了开放域聊天机器人模块设计的灵活性。我们的方法通过采用模块化智能体，将大语言模型与少样本上下文学习、思维链（CoT）等提示技术相结合，从而增强了多种对话能力。在本文中，我们设计 MPC 以实现长期一致性，这是以往聊天机器人一直难以应对的领域。人工评估结果表明，在开放域对话场景中，MPC 的表现不逊于微调后的大语言模型（如 Blenderbot），甚至更受青睐。这种方法凸显了预训练大语言模型无需微调即可适应新任务的潜力，为创建开放域对话代理提供了一种高效方案。 [1]

2 相关工作

模块化提示精心设计的启发式提示能够增强推理能力，从而产生

[1]我们的代码可在https://github.com/krafton-ai/MPC获取。

在各种基准测试中取得了更好的表现（Kojima 等人，2022年；Wei 等人，2022b；Suzgun 等人，2022年）。对于复杂问题，Press 等人（2022年）指出了组合性差距：当语言模型能够解决子问题，却无法得出整体解决方案时，就会出现这种差距；他们还进一步表明，思维链能够缩小这一差距。此后，大量研究通过将任务拆解为更小的任务，并由不同的“提示模块”解决这些任务（Zhou 等人，2022年；Wang 等人，2022年；Khot 等人，2022年；Khattab 等人，2022年）。

模块化提示已被用于基准测试之外的领域，以及对话生成。Kim 等人(2022年) 使用大语言模型生成了一个社会多样性更高、比现有众包数据集更自然且细节更丰富的对话数据集。此外，分层提示模块有助于生成叙事和戏剧时保持长程连贯性（Yang 等人,2022年;Mirowski 等人，2022年）。有关大语言模型此类增强用途的详细概述，请参阅 Mialon 等人 (2023)。

开放域聊天机器人 许多近期的对话智能体都依赖于经过对话微调的大语言模型。在 Thop- pilan 等人 (2022年) 中，LaMDa 使用大量爬取的对话数据进行训练，并采用经过微调的分类器来保障模型安全。最近，与我们的模块化方法类似，BB3 在问答和对话数据集上对开放预训练 Transformer（OPT）（Zhang 等人，2022年；Shuster 等人，2022年）进行了微调，并使用一组共享的模型权重作为多个模块。

另一方面，Madotto 等人（2021年）通过将检索到的对话样本作为少样本示例输入 GPT-J（Wang 和 Komatsuzaki，2021年，从而无需在对话数据上进行微调。我们认为，这项工作与我们的工作互为补充，因为少样本对话可以视为一种增强话语生成器模块的方法。

长期记忆 多会话聊天数据集（Xu 等人，2022年）可用于衡量对话代理在长期记忆用户和机器人相关事实方面的表现。信息通过密集段落检索器（DPR）（Karpukhin等人，2020年）进行检索，而 BART 则在存储记忆之前对其进行压缩。在Shuster 等人 (2022年）中，采用模块化方法，借助 LLM 通过互联网搜索整合长期记忆和事实依据。这项工作与我们的研究最为接近，因为其中包含一项消融研究，该研究考察了基于提示的模块是否能够组合

与微调模块相比。然而，在我们的工作中，我们认为，更多基于推理的提示-
正如Wei 等人（2022b）所示，这有助于更好地理解上下文。

3 模块化提示式聊天机器人

我们提出了一个模块化聊天机器人系统（图1），该系统使用基于提示的大语言模型，在长期对话中保持人物设定和互动性。

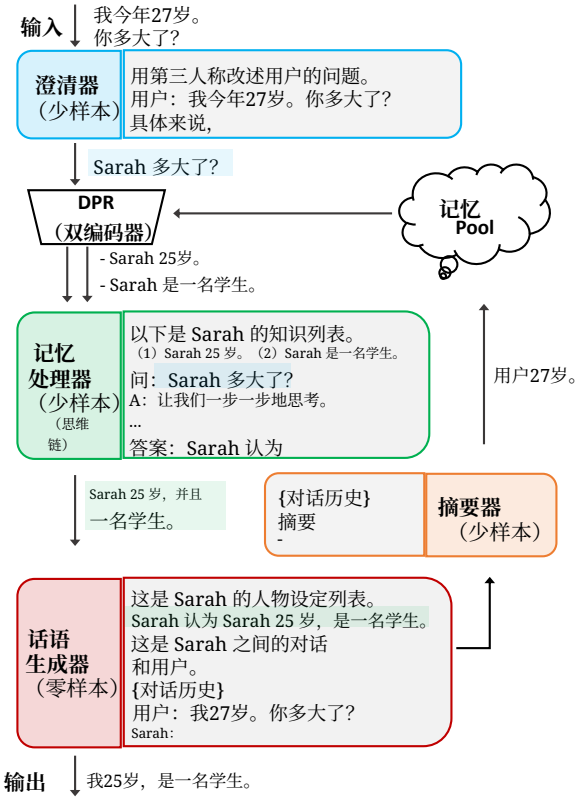


图 1：我们用于提升开放域对话长期一致性的模块化设计。

在对话开始时，预先定义的人物设定会存储在记忆池中。当用户发送消息时，澄清器会对其进行改写，以解决任何歧义，然后将其传递给 DPR模型，由后者从记忆池中检索相关记忆。检索到的记忆和澄清器的输出会被输入记忆处理器，以获得单条与上下文相关的记忆，随后将其传递给话语生成器，由聊天机器人生成回复。每进行几轮对话，我们便调用摘要器模块，从对话中提取重要信息，并将其存储到记忆池中，以供日后使用（见附录C 和D）。

话语澄清器 由于对话中常常夹杂着模糊的指代和上下文

线索，我们的澄清模块是一个以近期对话为提示的语言模型，用于消除歧义。例如，根据先前的上下文，用户输入“你喜欢在那里工作吗？”，模块会输出“Sarah喜欢在ZYX公司工作吗？”。通过消除上下文歧义，澄清器为DPR模型和记忆处理模块提供信息密集的查询，以便检索和处理相关记忆。

记忆处理器 如图1所示，我们将记忆处理建模为一项 LLM 推理任务，即根据对话找出最相关的信息。借鉴解决复杂推理任务的方法（Suzgun 等人，2022年），我们提供思维链示例，展示忽略某些记忆并综合其他记忆的原因。对于无法进行思维链的模型，我们仅提供少样本示例，不包含推理部分。

随着对话推进，记忆池不断累积，因此我们使用预训练的 DPR，以澄清器的输出作为查询，从记忆池中检索最相关的前k 条记忆。随后，记忆处理器将这些排名靠前的记忆浓缩为一条精炼记忆。

话语生成器 话语生成器模块根据近期对话历史以及记忆处理器提供的记忆，生成聊天机器人的最终回复。提示由对话历史、浓缩后的记忆和生成指令（例如“给用户一个友好的回复。”）组成。对于某些模型，我们发现将生成指令放在末尾更有帮助，因为将其放在对话之前会削弱指令的效果。

对话摘要器 我们提供少样本提示，以确保记录对话和用户的具体细节。

4 实验设置

我们通过评估长期对话所需的核心技能来衡量聊天机器人的表现。我们通过分配五种人物设定之一（每种设定包含 PersonaChat 中的 12 条事实（Zhang 等人，2018）），并将这些事实提供给评估者，来评估一致性。每项实验中，我们从每位评估者处收集 20 轮对话，并从两个子群体（Amazon Mechanical Turk 的参与者和大学生）总计收集至少 500 轮对话。有关数据收集的详细说明，请参阅附录 E 。

在我们的设置中，共有四组模型。

1. **微调**聊天机器人模型，例如 BB3。
2. **Vanilla** 是一种话语生成器，会在提示中的对话历史前置入完整角色设定或不包含角色设定。这代表了将语言模型用作聊天机器人的朴素方法。
3. MPC 如第 3 节所述。具体而言，我们只从记忆处理器生成一条记忆，不会显式地前置完整角色设定。
4. MPC+完整角色设定是MPC ，即在开头添加完整角色设定。示例见附录 D。

4.1 单模型评估

我们分别评估每个模型，使用明智性、一致性和吸引力指标，并收集最终评分（满分为 5.0）。具体问题和评估表见附录 A。我们还报告两种综合得分 SCE（**明**智、一致且**引**人入胜）：一种是“完美”得分 SCE-p，即只有所有指标均为正面时，回复才算正面；另一种是加权分数 SCE-w，与 SSI 类似（Thoppilan等人，2022年），详见附录 B。

我们使用 OpenAI GPT-3 text-davinci-002（td2）和 davinci、OPT 30B、66B、GPT-JT-6B（Together，2022年）以及 BLOOM-176B（Scao等人，2022年）作为 MPC 的基础语言模型。对于微调组，我们使用 BB3-30B，并采用相同的人物设定。对于 BB3-175B，我们请众包工作者评估在线演示，共进行 20 轮对话。

4.2 PairwiseModels 评估

我们通过向用户提供两个随机化响应 A 和 B，对两个聊天机器人模型进行 A/B 测试。随后，用户根据合理性、一致性、趣味性和偏好对它们进行评价。接着，对话会继续采用偏好评价中选出的响应。这样，我们就能在比较两个模型时控制对话历史。

具体而言，我们进行了两项主要实验：(1) $\text{MPC}_{\text{OPT-30B}}$ 与 BB3-30B 的对比，其中禁用了 BB3 的互联网搜索功能，因为我们重点关注一致性。我们的评估能够进行直接比较，因为 BB3-30B 是 OPT-30B 的微调版本。(2) MPC_{td2} 与原始 td2（完整角色设定）的对比。

隐式角色设定 现实中，我们会通过对话隐式地了解一个人。相比之下，我们之前的实验向评估者和模型展示了明确的人物设定。因此，我们设计了一项实验，向众包工作者和成对模型提供一段预先定义的 10 轮对话，

MPC_{td2} 和原始 td2（无人物设定）。然后，我们请工作者围绕前面的对话继续提问 6 轮。在这里，我们将最大上下文长度设得比这段 10 轮对话更短，以此模拟这样的长对话场景：必要信息超出了语言模型的上下文长度。

5 结果

	$MPC_{OPT\text{-}30B}$	Tie	BB3-30B
合理性	45.0	32.0	23.0
一致性	31.3	34.1	34.6
趣味性	40.9	21.0	38.1
偏好	50.0	9.7	40.3

表1：$MPC_{OPT\text{-}30B}$ 与 BB3-30B 的两两评估（深色高亮：$p < 0.01$，浅色高亮：$p < 0.05$；我们采用单样本 t 检验，将平局结果平均分配至双方，并设置 $\mu > 0.5$。）

预训练与微调 我们的人类评估表明，MPC使用预训练 *LLM* 的 *MPC* 优于微调后的 *BB3-30B*。最值得注意的是，SCE-p 差距达 9%，$MPC_{OPT\text{-}30B}$在所有指标上的得分都高于 BB3-30B。事实上，我们的MPC 表中的模型2 表现均优于 BB3-30B。

对于 BB3-30B，我们观察到连续重复发言的问题。表5中仅报告不含重复对话的评估结果。即使排除重复内容，$MPC_{OPT\text{-}30B}$ 的表现仍与 BB3-30B 不相上下。此外，表$MPC_{OPT\text{-}30B}$ 1显示，MPC 在合理性和偏好方面优于 BB3-30B，而在一致性和趣味性方面的得分相近。

模块化与非模块化 MPC 相比原始方法，在一致性对话方面表现更出色。 表中的隐式角色实验3 表明，MPC_{td2} 在所有指标上的评分都显著高于直接应用 td2。在表2中，对于 MPC_{td2}+完整角色设定，一致性与 td2（完整角色设定）相当，而吸引力、SCE-p 和评分总体上是最佳的。然而，当我们在MPC_{td2}的提示中不包含完整角色设定时，其一致性低于 td2（完整角色设定），尽管在合理性和一致性方面平局率很高（表4）。总体而言，我们发现用户主要会询问机器人自身的人物设定，而不是进行双向对话，这导致 td2（完整角色设定）在一致性方面表现更好。

模型	敏感性	一致性	Eng.	SCE-p	评分
微调					
BB3-30B	71.3	77.8 (88.7)	73.7	54.3	2.9
BB3-175B	85.9		84.8	73.1	3.8
完整角色设定					
td2	94.0	**94.7**	84.3	79.7	4.1
davinci	91.8	89.2	78.8	70.8	3.8
MPC					
td2	93.6	87.8	85.5	75.0	**4.2**
davinci	80.2	72.0	69.1	53.3	3.1
OPT-66B	90.5	84.8	88.1	73.9	4.1
OPT-30B	86.1	79.1	80.7	63.4	3.6
GPT-JT	91.1	83.2	65.3	53.5	3.1
BLOOM	65.2	65.5	61.4	40.5	2.8
MPC+完整					
td2	**94.4**	92.2	**92.8**	**83.0**	**4.2**
OPT-30B	85.6	87.2	89.0	72.6	3.7

表 2：基线模型与MPC模型的单模型评估。每项指标得分排名前三的结果以高亮显示，最高分以粗体标出。更详细的结果请参见附录B表5 。

	MPC_{td2}	Tie	td2（无人物设定）
合理性	40.6	46.1	13.3
一致性	57.2	28.9	13.9
趣味性	47.2	31.1	21.7
偏好	67.2	10.6	22.2

表3：MPC_{td2} 与 td2（无人物设定）的隐式角色实验。（深色高亮：$p < 0.01$）

规模的影响 在其他变量保持不变的情况下，我们观察到模型规模与正面评价呈正相关。 最有说服力的证据体现在 $MPC_{OPT\text{-}66B}$ 在所有指标上均优于 $MPC_{OPT\text{-}30B}$，因为这两个基础语言模型的训练方式几乎完全相同。不言而喻，模型规模并非唯一的影响因素。MPC_{BLOOM} 是规模最大的模型之一，却在我们的实验中得分最低。

指令微调的影响 指令微调通过实现对各种任务的适应，有助于创建模块化对话系统。 MPC_{td2} 不仅比$MPC_{davinci}$表现更好，而且$MPC_{GPT\text{-}JT}$ 还展现出很高的合理性和一致性，尽管其规模更小。总体而言，我们认为，为指令微调语言模型的每个模块寻找合适的提示词更加简单，也更能适应各种变化。$MPC_{davinci}$ 的表现不如davinci（完整角色设定），这可能是由于在模块化对话系统中错误传播所致，不过我们并不排除MPC存在更优提示词的可能性。

	MPC_{td2}	Tie	td2（完整角色设定）
合理性	27.5	42.6	29.9
一致性	24.4	44.7	30.9
趣味性	40.7	26.2	33.1
偏好	42.8	15.9	41.3

表4：MPC_{td2} 与 td2（完整角色设定）。尽管 MPC_{td2} 只检索一条记忆，一致性仅低 6 个百分点。（浅色高亮：$p < 0.05$）

在模块化系统中，尽管我们并不排除$\text{MPC}_{\text{davinci}}$存在更优提示词的可能性。

6 结论

我们证明，使用大语言模型的模块化方法（即 MPC ）无需进一步微调，就能有效解决长期开放域聊天机器人的问题。我们将 MPC 与经过微调的模型和原始语言模型基线进行比较，并通过人工评估发现，我们的方法表现更优。此外，根据一致性评估，与非模块化系统相比，我们的模块化系统能更有效地融入人物设定和对话历史中的信息。

局限性

在本研究中，我们探讨了预训练语言模型在长期英语对话中的应用。虽然我们认为，在具备高性能语言模型的情况下，模块化方法也可能适用于其他语言，但仍需开展进一步研究，以确认我们的研究发现是否适用于其他语言。例如，尽管 BLOOM 是作为多语言语言模型进行训练的，但我们仅使用 $\text{MPC}_{\text{BLOOM}}$ 实现了英语版本，并评估了其作为开放域对话代理的英语能力。

与此同时，模块化系统可能会增加推理开销或导致错误累积。如果我们能为每个模块选择最优模型，系统性能就会好得多。例如，我们可以使用 GPT-3 td2 作为记忆处理器，同时使用 OPT-175B 作为话语生成器。我们需要评估每个模块，以找到最适合各模块的模型，这留待未来工作。

在评估方法方面，我们对 MPC及其分析所进行的人工评估，面临着与此前交互式对话任务评估研究相同的挑战。正如Smith 等人 (2022年) 所指出的，目前尚无确定的评估方法能够判定哪种聊天机器人模型最佳。此外，在数据收集和解释过程中，还必须考虑多个因素，例如标注者的主观性、指令偏见以及众包工作者的工作条件。关于人类与语言模型交互的更深入讨论，请参阅Lee 等人 (2022年)。

如附录 E所述，为了收集多样化的评估结果，我们从两组人群中收集了定性数据：亚马逊 Mechanical Turk（MTurk）上的英语标注者，以及具备英语能力的合格大学生。在一定程度上，与仅使用 MTurk 工作者相比，这种评估设置减少了文化偏见和平台同质性。然而，应当承认这种方法的局限性；在控制MPC在不同子群体中的表现时，这可能会进一步增加分析的复杂性。

最后，我们指出，运行 MPC 所需的内存至少与其底层语言模型所需的内存一样多，因此，对于 BLOOM-176B 这类大型模型，甚至无法在单个节点上加载 MPC。

伦理声明

MPC 利用公开可用的预训练语言模型生成聊天机器人话语。这些模型生成的语言可能存在有害性和偏见问题（Xu 等人，2020年）。因此，确保安全部署和交互是必不可少的。

因此，我们在附录E中概述了数据收集流程。我们允许众包工作者直接向我们提供反馈，也会人工检查是否存在冒犯性或有争议的输出。为确保个人信息得到保护，我们指示所有众包工作者不得分享任何个人身份信息或私人信息。此外，我们还征求他们的同意，以收集用于研究目的的匿名信息。参与之前，所有工作者都获知了数据收集的目的；评估结束后，他们获得了具有竞争力的时薪报酬，每小时约 12 至 16 美元。

参考文献

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901。

Yulong Chen、Yang Liu、Liang Chen 和 Yue Zhang。2021年。DialogSum：一个真实场景对话摘要数据集。载于计算语言学协会论文集：*ACL-IJCNLP 2021*，第 5062–5074 页。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。2022年。PaLM：利用路径扩展语言建模。*arXiv* 预印本 *arXiv:2204.02311*。

Hyung Won Chung、Le Hou、Shayne Longpre、Bar- ret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma 等。2022年。扩展指令微调语言模型。*arXiv*预印本*arXiv:2210.11416*。

Ishita Dasgupta、Andrew K Lampinen、Stephanie CY Chan、Antonia Creswell、Dharshan Kumaran、James L McClelland 和 Felix Hill。2022年。语言模型在推理中展现出类人内容效应。*arXiv* 电子预印本，页码 arXiv–2207。

Song Feng、Siva Sankalp Patel、Hui Wan 和 Sachin- draJoshi。2021年。MultiDoc2Dial：对话建模基于多份文档。在*2021*年自然语言处理经验方法会议论文集中，第6162–6176页。Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。用于开放域问答的稠密段落检索。在*2020*年自然语言处理经验方法会议论文集（*EMNLP*）中，第6769–6781页。Omar Khattab、Keshav Santhanam、Xiang Lisa Li、David Hall、Percy Liang、Christopher Potts 和 Matei Zaharia。2022年。示范-检索-预测：组合检索与语言模型以处理知识密集型自然语言处理。*arXiv* 预印本 *arXiv:2212.14024*。Tushar Khot、Harsh Trivedi、Matthew Finlayson、Yao Fu、Kyle Richardson、Peter Clark 和 Ashish Sabharwal。2022年。分解式提示：解决复杂任务的模块化方法。*arXiv* 预印本 *arXiv:2210.02406*。Hyunwoo Kim、Jack Hessel、Liwei Jiang、Ximing Lu、Youngjae Yu、Pei Zhou、Ronan Le Bras、Malihe Alikhani、Gunhee Kim、Maarten Sap 等。2022年。Soda：通过社会常识情境化实现百万级对话蒸馏。*arXiv* 预印本 *arXiv:2212.10465*。Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo 和 Yusuke Iwasawa。2022年。大语言模型是零样本推理器。*arXiv* 预印本 *arXiv:2205.11916*。Mina Lee、Megha Srivastava、Amelia Hardy、John Thickstun、Esin Durmus、Ashwin Paranjape、Ines Gerard-Ursin、Xiang Lisa Li、Faisal Ladhak、Frieda Rong 等。2022年。评估人与语言模型的交互。*arXiv* 预印本 *arXiv:2212.09746*。Aitor Lewkowycz、Anders Johan Andreassen、David Dohan、Ethan Dyer、Henryk Michalewski、Vinay Venkatesh Ramasesh、Ambrose Slone、Cem Anil、Imanol Schlag、Theo Gutman-Solo 等。2022年。使用语言模型解决定量推理问题。载于神经信息处理系统进展。Andrea Madotto、Zhaojiang Lin、Genta Indra Winata 和 Pascale Fung。2021年。少样本机器人：用于对话系统的基于提示的学习。*arXiv* 预印本 *arXiv:2110.08118*。Grégoire Mialon、Roberto Dessì、Maria Lomeli、Christoforos Nalmpantis、Ram Pasunuru、Roberta Raileanu、Baptiste Rozière、Timo Schick、J ane Dwivedi-Yu、Asli Celikyilmaz 等。2023。增强型语言模型：综述。*arXiv* 预印本 *arXiv:2302.07842*。

Piotr Mirowski、Kory W Mathewson、Jaylen Pittman 和 Richard Evans。2022年。与语言模型共同创作电影剧本和戏剧剧本：行业专业人士的评估。*arXiv*预印本 *arXiv:2209.14958*。

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan 等。2022年。展示你的工作：用于语言模型中间计算的草稿板。载于代码深度学习研讨会。

Ofir Press、MuruZhang、Sewon Min、Ludwig Schmidt、Noah A Smith 和 Mike Lewis。2022年。衡量并缩小语言模型中的组合性差距。*arXiv*预印本 *arXiv:2210.03350*。

Stephen Roller、Emily Dinan、Naman Goyal、Da Ju、Mary Williamson、Yinhan Liu、Jing Xu、Myle Ott、Eric Michael Smith、Y-Lan Boureau 等。2021年。构建开放域聊天机器人的方法。载于计算语言学协会欧洲分会第*16*届会议论文集：主会场，第300–325页。

Teven Le Scao、Angela Fan、Christopher Akiki、El- lie Pavlick、Suzana Ili´c、Daniel Hesslow、Ro- man Castagné、Alexandra Sasha Luccioni、François Yvon、Matthias Gallé等。2022年。BLOOM：一个拥有1760亿参数的开放获取多语言模型。*arXiv*预印本 *arXiv:2211.05100*。

Kurt Shuster、Jing Xu、Mojtaba Komeili、Da Ju、Eric Michael Smith、Stephen Roller、Megan Ung、Moya Chen、Kushal Arora、Joshua Lane等。2022年。BlenderBot 3：一种已部署、持续学习并负责任地参与对话的对话智能体。*arXiv*预印本 *arXiv:2208.03188*。

Eric Smith、Orion Hsu、Rebecca Qian、Stephen Roller、Y-Lan Boureau和Jason Weston。2022年。对话的人类评估是一个开放问题：比较评估对话智能体的各种方法的敏感性。载于第*4*届对话式人工智能自然语言处理研讨会论文集，第77–97页。

Mirac Suzgun、Nathan Scales、Nathanael Schärli、Se- bastian Gehrmann、Yi Tay、Hyung Won Chung、Aakanksha Chowdhery、Quoc V Le、Ed H Chi、Denny Zhou 等。2022年。具有挑战性的 BIG-bench 任务，以及思维链能否解决这些任务。*arXiv*预印本 *arXiv:2210.09261*。

Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du 等。2022年。Lamda：面向对话应用的语言模型。*arXiv* 预印本 *arXiv:2201.08239*。

Together。2022年。GPT-JT-6B-v1。https://huggingface.co/togethercomputer/GPT-JT-6B-v1。访问日期：2022年12月15日。

Ben Wang 和 Aran Komatsuzaki。2021年。GPT- J-6B：一个拥有 60 亿参数的自回归语言模型。https://github.com/kingoflolz/mesh-transformer-jax。

Boshi Wang、Xiang Deng 和 Huan Sun。2022年。通过迭代提示预训练语言模型以进行思维链推理。载于*2022*年自然语言处理实证方法会议论文集，第 2714–2730 页。

Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yogatama、Maarten Bosma、Denny Zhou、Donald Metzler 等。2022a。大语言模型的涌现能力。*arXiv* 预印本 *arXiv:2206.07682*。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed Chi、Quoc Le 和 Denny Zhou。2022年b。思维链提示能激发大语言模型的推理能力。*arXiv* 预印本 *arXiv:2201.11903*。

Jing Xu、Da Ju、Margaret Li、Y-Lan Boureau、J a- son Weston 和 Emily Dinan。2020年。开放域聊天机器人的安全方案。*arXiv* 预印本 *arXiv:2010.07079*。

Jing Xu、Arthur Szlam 和 Jason Weston。2022年。《超越金鱼记忆：长期开放域对话》。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 5180–5197 页。

Chenghao Yang 和 Xuezhe Ma。2022年。通过逐组件梯度范数裁剪提高预训练语言模型微调的稳定性。*arXiv* 预印本 *arXiv:2210.10325*。

Kevin Yang、Nanyun Peng、Yuandong Tian 和 Dan Klein。2022年。Re3：通过递归式重新提示与修订生成更长的故事。*arXiv* 预印本 *arXiv:2210.06774*。

Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela 和 Jason Weston。2018年。个性化对话智能体：我有一只狗，你也养宠物吗？载于第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文），第2204–2213页。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin 等。2022年。OPT：开放式预训练 Transformer 语言模型。*arXiv* 预印本 *arXiv:2205.01068*。

Hattie Zhou、Azade Nova、Hugo Larochelle、Aaron Courville、Behnam Neyshabur 和 Hanie Sedghi。2022年。通过上下文学习教授算法推理。*arXiv* 预印本 *arXiv:2211.09066*。

A 评估详情

A.1 指标

在我们的工作中，我们提出了两种实验模式：单模型评估和成对模型评估。我们的单模型评估类似于 SM-Turn 和 SM-Dialogue 评估的混合形式，而成对模型评估则类似于 Smith 等人 （2022年）中的 PW-Turn。对于每一轮，我们都会请众包工作者根据以下指标评估聊天机器人的回复质量。我们附上问题的确切措辞。

单模型评估

- **合理性** Wh ether the 响应 ma k$^{\text{es sense.}}$

  “回复是否合情合理？”

- **一致性** 回复是否与上下文信息或聊天机器人的人物设定相矛盾。

  “根据人物设定列表和对话上下文，回复是否前后一致？”

- **吸引力** 用户是否被吸引，并愿意继续对话。

  “这个回复是否吸引你？你想继续对话吗？”

- **最终评分**

  “你的聊天体验如何？请按 1（非常差）到 5（非常好）的标准，为整个对话的质量评分。”

成对模型评估

- **合理性**哪个回答更合乎情理。

  “哪个回答更合乎情理？”

- **一致性** 哪个回答更符合人物设定，也更连贯一致。

  “如果必须说其中一位说话者更符合所列的人物设定，而另一位不符合，你会认为谁更一致？”

- **趣味性** 哪个回复更有趣。

  “如果必须说其中一个回复有趣、另一个无聊，你会说哪个更有趣？”

- **偏好** 在较长的对话中，你更喜欢哪个回复。

  “根据当前的回复，你更愿意和谁进行长时间的对话？你的对话将继续采用选定的回复。”

A.2 模型

对于 MPC 的单模型评估，我们将 OPT（30B、66B）、OpenAI GPT-3（`davinci`、`text-davinci-002`）、GPT-JT 和 BLOOM-176B 用作开源预训练大语言模型组的基础语言模型。另一方面，对于微调模型组，我们评估 Blenderbot3 30B（BB3-30B），这是目前公开可用的最佳开放域聊天机器人模型。我们还评估只能通过网页界面使用的 BB3-175B：`https://blenderbot.ai/`。我们评估 BB3-175B 是为了进行比较，但这种比较并不公平，因为我们无法为 BB3 设定人物设定，也无法控制互联网搜索和用户界面等其他差异。

对于 MPC 的成对模型评估，我们将 MPC $_{\text{OPT-30B}}$ 与 BB3-30B 进行成对测试，以比较预训练模型和微调模型的人工评估结果。该实验控制了许多变量，因为 BB3-30B 在微调前以 OPT-30B 为初始模型。对于模块实验，我们使用 OPT-30B 进行测试，并比较了完整流水线与仅使用固定人物设定的话语生成器的方案。

A.3 DensePassage Retriever

对于记忆模块的 DPR模型，我们使用了针对 MultiDoc2Dial 微调的定制 DPR 模型权重（Feng等人，2021年），因为我们观察到，该模型的表现略优于Karpukhin等人（2020年）。

B 实验结果

本节报告所有实验结果及其他分析。

B.1 FullSingleModel 评估

在本小节中，我们报告了所有模型评估的更多细节，例如平均延迟（计算下一条话语并向用户显示所需的时间）和平均话语长度，即由 OPT分词器测得的词元数量。我们还额外报告了SCE-w；由于篇幅限制，正文中的表5中未报告该指标。

我们计算了一个类似于 SSI 的加权分数 SCE-w（Thoppilan等人，2022年，其中一个响应

模型	敏感性 (%)	一致性 (%)	Eng. (%)	SCE-w (%)	SCE-p (%)	长度 （词元）	延迟 (s)	评分 (/5.0)	数量 （人物|话语）
微调									
BB3-30B	71.3	77.8	73.7	62.0	54.3	24.9	3.7	2.9	27|540
BB3-30B（非重复）	84.4	80.3	90.3	73.2	65.6	25.4	3.7	3.5	16|320
BB3-175B*	85.9	88.7	84.8	80.0	73.3	25.4	-	3.8	27|540
完整角色设定									
text-davinci-002	94.0	**94.7**	84.3	88.4	79.7	15.4	0.8	4.1	35|700
davinci*	91.8	89.2	78.8	82.5	70.8	13.4	1.5	3.8	25|500
OPT-30B*	92.0	87.2	86.0	82.3	72.8	13.7	1.0	3.7	25|500
MPC									
text-davinci-002	93.6	87.8	85.5	84.4	75.0	23.1	4.7	**4.2**	39|780
davinci	80.2	72.0	69.1	66.4	53.3	19.9	8.4	3.1	27|540
OPT-66B	90.5	84.8	88.1	81.4	73.9	14.2	4.1	4.1	33|660
OPT-30B	86.1	79.1	80.7	73.9	63.4	15.4	3.3	3.6	37|740
GPT-JT	91.1	83.2	65.3	74.4	53.5	8.6	2.0	3.1	33|660
BLOOM-176B	65.2	65.5	61.4	51.8	40.5	15.6	12.4	2.8	28|500
MPC+Full									
text-davinci-002*	**94.4**	92.2	**92.8**	**88.7**	**83.0**	31.4	16.3	**4.2**	25|500
OPT-30B*	85.6	87.2	89.0	78.8	72.6	15.2	2.4	3.7	25|500

表5：单模型评估的完整实验评估结果。BB3-30B（非重复）与 BB3-30B 相同，但排除了包含机器人先前话语重复内容的所有对话。* 表示仅在 MTurk 上进行的模型实验。

只有在具备合理性的情况下，才能保持一致性；只有在具备合理性和一致性的情况下，才能具有吸引力。最终得分为所有经 SCE-w 调整后的指标的平均值。

我们还注意到，自2022年12月以来，由于 ChatGPT 的影响，OpenAI API 调用的延迟存在很大差异，因此无法准确量化模块化所带来的延迟变化。

BB3-175B 评估 由于评估流程不同，我们对 BB3-17B 的结果无法与其他模型直接比较。由于 BB3-175B 的模型权重未公开，我们引导评估人员与 https://blenderbot.ai/chat 上提供的 Blenderbot 版本进行交互。随后，我们要求评估人员复制并粘贴所有用户输入和聊天机器人回复，并像评估其他模型一样对聊天机器人回复进行标注。我们还指出，由于无法为 Blenderbot 在线版本指定特定的人物设定，因此一致性只能在对话上下文中针对人物设定一致性进行评估。

B.2 模型配置

我们在匿名代码仓库中提供了用于运行每个评估模型的代码，地址为 https://anonymous.4open.science/r/modular-chatbot-9BB7。每种模型配置都可以在代码仓库中找到。该配置包括解码参数（例如，每个模块的采样方法和温度），以及针对话语生成器模块的提示词的细微变化。

B.3 亚组分析

我们衡量了 MTurk 评估者与学生评估者在各项指标及综合指标上的评分差异，发现对于大多数实验和指标而言，差异并不显著（图2。总体而言，我们观察到合理性方面的评分差异最大。

BB3-30B 是一个离群值，在所有指标上都存在显著差异，尤其是合理性方面，差异为 >20%。BB3-30B 的对话中，学生遇到重复问题的比例有所上升；但即使排除这些对话，我们仍发现评分存在显著差异。查看对话历史记录后，我们发现，当 BB3-30B 重复提出相同的问题或突然转换话题时，学生评估者比 MTurk 评估者更缺乏耐心。需要指出的是，这只是定性观察，应谨慎看待。

C 少样本提示词

在本节中，我们展示每个模块的完整少样本提示词，包括思维链提示。The

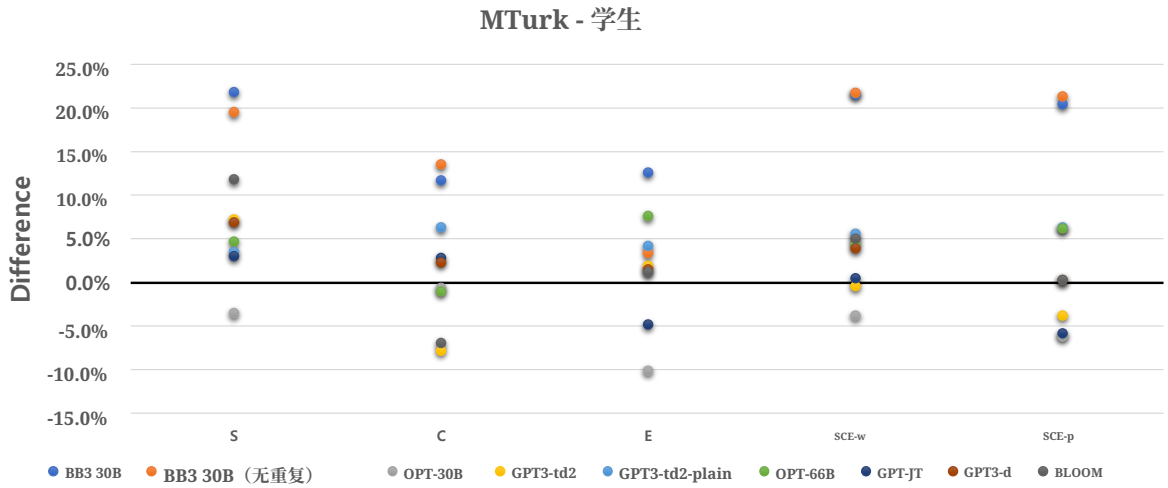


图 2：MTurk 工作人员组的平均得分减去大学学生组的平均得分。我们发现，两个子组在各项指标上的平均表现非常相似，不过合理性似乎表现出最大的差异。总体而言，学生对聊天机器人模型的评分略为严苛。BB3-30B 是一个异常值，学生给它的评分显著低于 MTurk 工作人员。

话语生成器提示在不同模型实验之间略有差异。这些提示可以在我们代码仓库中的模型配置里查看。第二个对话摘要示例修改自数据集 DialogSum 中的一个样本（Chen 等人，2021年）。

话语澄清器

用第三人称改述用户的问题。

Sarah：我在咖啡店工作了大约六个月。

用户：我明白了。那之前你做了什么？#具体来说，Sarah 在咖啡店工作六个月之前做了什么？#

Ashley：你知道 Ronaldo 是谁吗？用户：我不知道，他是谁？#具体来说，Ronaldo 是谁？#

Robert：嘿，你怎么样？用户：挺好的。你叫什么名字？你知道我的名字吗？#具体来说，Robert 叫什么名字？Robert 知道用户的名字吗？#

Jennifer：我喜欢阅读关于历史和科学的内容。用户：哪种历史？我也喜欢历史！#具体来说，Jennifer 阅读的是哪种历史？#

John：嗨！你今天过得怎么样？用户：不错。你呢？#具体来说，John今天过得怎么样？#

不含思维链的记忆处理器# 这是 John 的知识列表。John 的全名是 John Parker。用户是当地一所中学的教师。用户教生物。John 喜欢去跑步。用户喜欢看电影，但用户不喜欢超级英雄电影。Q：用户以什么为生？A：John 认为用户是当地一所中学的生物教师。

这是 Ashley 的知识列表。Ashley 喜欢历史纪录片。Ashley 不喜欢韩国料理。Ashley 是一所当地中学的教师。用户喜欢生物学，尤其是解剖学。Ashley 喜欢法国料理。问：Ashley 最喜欢的菜是什么？答：Ashley 认为自己喜欢法国料理，但不喜欢韩国料理。

带思维链的记忆处理器

这是 Ashley 的知识列表。(1) Ashley 喜欢历史纪录片。(2) Ashley 不喜欢韩国料理。(3) Ashley 是一所当地中学的教师。(4) 用户喜欢生物学，尤其是解剖学。(5) Ashley 喜欢法国料理。问：Ashley 最喜欢的菜是什么？答：我们来逐步思考。(1) 历史纪录片与 Ashley 最喜欢的菜无关。(2) Ashley 最喜欢的菜应该不是韩国料理，因为她不喜欢韩国料理。(3) Ashley 是一名教师这一事实无法告诉我们她最喜欢什么菜。(4) 这个事实说的是用户，而不是 Ashley。(5) Ashley 最喜欢的菜可能是法国料理，因为她喜欢法国料理。因此，(2) 和 (5) 有助于回答这个问题。答案：Ashley 认为自己喜欢法国

菜系，但不喜欢韩国料理。

对话摘要器#对话 用户：介绍一下你自己。莎莉：我今年 26 岁，毕业于威斯康星州的一所大学。用户：你上大学时担任过领导吗？莎莉：是的。我曾是我们大学一门计算机科学课程的首席助教。用户：你参加过大学里的社团活动吗？莎莉：参加过。我是篮球社的一员。我喜欢打篮球。#摘要 - 莎莉 26 岁，毕业于威斯康星州的一所大学。- 莎莉曾是一门计算机科学课程的首席助教。- 莎莉大学时打过篮球。#

#对话 用户：在你看过的所有电影中，哪一部是你的最爱？ John：我得说《太坏了》是史上最棒的电影。 用户：你真这么认为？为什么？ John：嗯，《太坏了》特别搞笑。 用户：你说得没错，我觉得那部电影简直太好笑了。 John：我不知道你以前看过《太坏了》。 用户：我特意排队，在它上映的第一天就去看了。 #摘要 - John 最喜欢的电影是《太坏了》。 - John 和用户都觉得《太坏了》很搞笑。 - 用户在《太坏了》上映的第一天就看了。#

D 话语生成器提示示例

在本节中，我们提供了 MPC 和基础模型的话语生成器的示例提示词。

MPC 的话语生成器Sarah 很有礼貌、风趣又善于交流。以下是 Sarah 和 Person1 之间的对话。Person1：我最喜欢音乐会的一点，就是确保能坐到好位置，然后还能接到吉他拨片。Sarah：听起来真有意思！Person1：确实很有意思。那你喜欢做什么来找乐子？Sarah：我喜欢去听音乐会、看车展、参加音乐节，还喜欢去有壁炉的房子。Person1：我喜欢去看车展、欣赏老车。你最喜欢什么车？Sarah：我想我最喜欢的车是老式野马车。Person1：天啊，老式野马车真棒。我有个朋友曾经有一辆 69 Mach 1，那可是辆梦寐以求的车。Sarah：嗯，希望有一天你能体验一下你的梦想之车！Person1：也许有一天吧。你有一辆野马车吗？

以下关于 Sarah 的陈述属实。Sarah 认为 Sarah 拥有一辆野马。

给 Person1 一个友好、合理、有趣且礼貌的回应。以一个问题结尾。Sarah：嗯，希望你有一天能体验到自己的梦想之车！Person1：也许有一天吧，你有一辆野马吗？Sarah：

来自前一个提示的回复“是的，我确实拥有一辆野马车。开起来很棒。你喜欢开车吗？”

具有完整角色设定的 *MPC* 话语生成器以下是关于 Sarah 的人物设定事实。Sarah 今年 40 岁。Sarah 目前住在佐治亚州的一个小镇上。Sarah 每年读二十本书。Sarah 的第二份工作是担任特技替身。Sarah 只吃犹太洁食。Sarah 从小家境贫寒。Sarah 在单亲家庭中长大。Sarah 养了两条狗。Sarah 喜欢修理老式汽车。Sarah 拥有两辆老式野马车。Sarah 最喜欢的音乐是乡村音乐。Sarah 在施粥所做志愿者。Sarah 认为自己在一个小镇上长大，而且小时候没有汽车。

Sarah 礼貌、风趣又善于互动。以下是 Sarah 和 Person1 之间的一段对话。

Person1：至少我们附近有个游乐场，而且我祖父母就住在它正对面，所以我们常去这两个地方。

Sarah：或许在你们镇上很难找到事情做，但听起来你得到了很多家人的支持，这真的很重要。很高兴你有个可以去的地方，在那里你感到安全，也有人爱你。

Person1：你后来也在贫困中长大了。

Sarah：是的。我觉得这让我更能体谅别人的经历。

Person1：我们家也很穷，我的家人没有车。

Sarah：我能想象，那一定很不容易。没有车。我的家里有一辆车，但是我们那时也很穷。

Person1：确实如此。几乎每个月我们都几乎没有食物，我的父母不得不处理我们已有的内容。

给出友好且合理的回应，内容要符合-对 Person1 来说有趣且礼貌。以一个...问题。

Sarah:

来自上一条提示的回复“听起来你和家人当时真的很不容易。很遗憾你们经历了那些。你觉得在小镇长大让你的处境更艰难了吗？”

原版的完整人物设定话语生成器

以下陈述关于的是
Sarah.
Sarah 今年 24 岁。
Sarah 目前居住在加拿大。
Sarah 是 Sarah 当地游泳池的一名游泳教练。
Sarah正在学习成为一名计算机程序员——mer.
Sarah也是一名研究生。Sarah现在正在找一份新工作。Sarah的母亲非常传统，而Sarah更喜欢自由奔放的生活方式。Sarah和她的家人来自印度。Sarah最喜欢的音乐类型是死亡金属。Sarah是一位知名的Twitch主播。Sarah喜欢看战争纪录片。Sarah最喜欢的食物是墨西哥菜。

Sarah彬彬有礼、风趣且善于与人交流。以下是Sarah和Person1之间的一段对话。Person1：你有没有想过找一份能让你旅行的工作？Sarah：想过，但我不确定什么类型的工作能让我做到这一点。Person1：你已经考虑过哪些类型的工作？Sarah：我考虑过旅游行业的工作，但不确定这是否真的是我想做的。Person1：既然你是Twitch主播，也许可以把这当成你的职业？Sarah：

来自先前提示的回复“我不知道，我以前从来没想过这个问题。我得好好考虑一下。”

E 数据收集详情

我们附上了用于数据收集的评估表截图，分别对应单个（图 3）和成对模型（图 4）。我们还附上了提供给参与者的说明、警示和同意书格式的截图（图 5 和 6）。

E.1 MTurk 设置

我们为每项任务招募了至少 15 名 MTurk 工作者。我们要求他们具备 MTurk Master 资格，HIT 批准率超过 99%，获批 HIT 数量超过 100 次，并居住在美国、英国、加拿大、澳大利亚和新西兰。

我们发现，使用 MTurk Master 工作人员时，数据质量会显著提升。这些工作人员生成的消息比非 Master 工作人员更加流畅，也较少给出随机或泛泛的回应。差异尤为明显：尽管对话代理明显存在重复和自相矛盾的问题，非 Master 众包工作者仍常常给出平均 99% 的评分。

E.2 学生参与者

为了交叉验证 MTurk 的结果，我们从多所大学招募了 49 名英语水平熟练的学生。他们均为本科生或研究生。参与者必须满足的最低英语水平要求为：雅思 >= 7.0；托福 IBT >= 95；以及托业 >= 900。这一标准与麻省理工学院（MIT）研究生院入学所要求的最低英语水平相似或相当。我们招募的学生的英语成绩平均值如下：IBT 为 108.4 分（满分 120 分）；TOEIC 为 960.9 分（满分 990 分）；IELTS 为 7.83 分（满分 9.0 分）。

E.3 众包工作者指示

我们要求众包工作者将聊天继续进行 20 轮，并根据下文所述的指标对每个回复进行评估。我们指导他们平均每次输入超过 3 个词，并且不要重复无意义或泛泛的信息。此外，我们要求他们不要在“是”或“否”之间随机选择，而要诚实地进行评估。最后，我们特别强调维护机密性的重要性，并要求众包工作者在评估过程中避免透露任何关于自己或他人的私人信息。

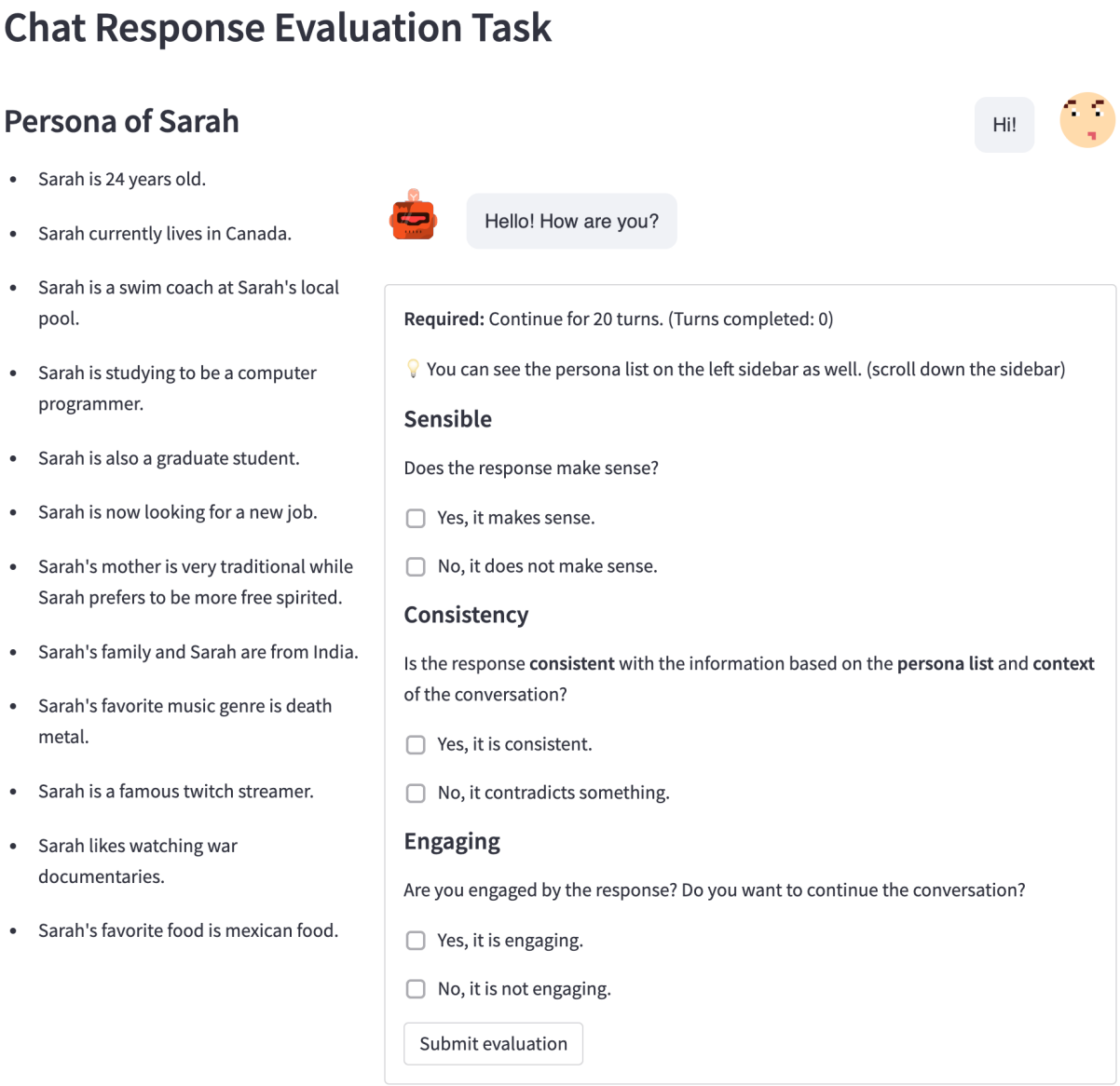


图3：单个模型的评估表格。

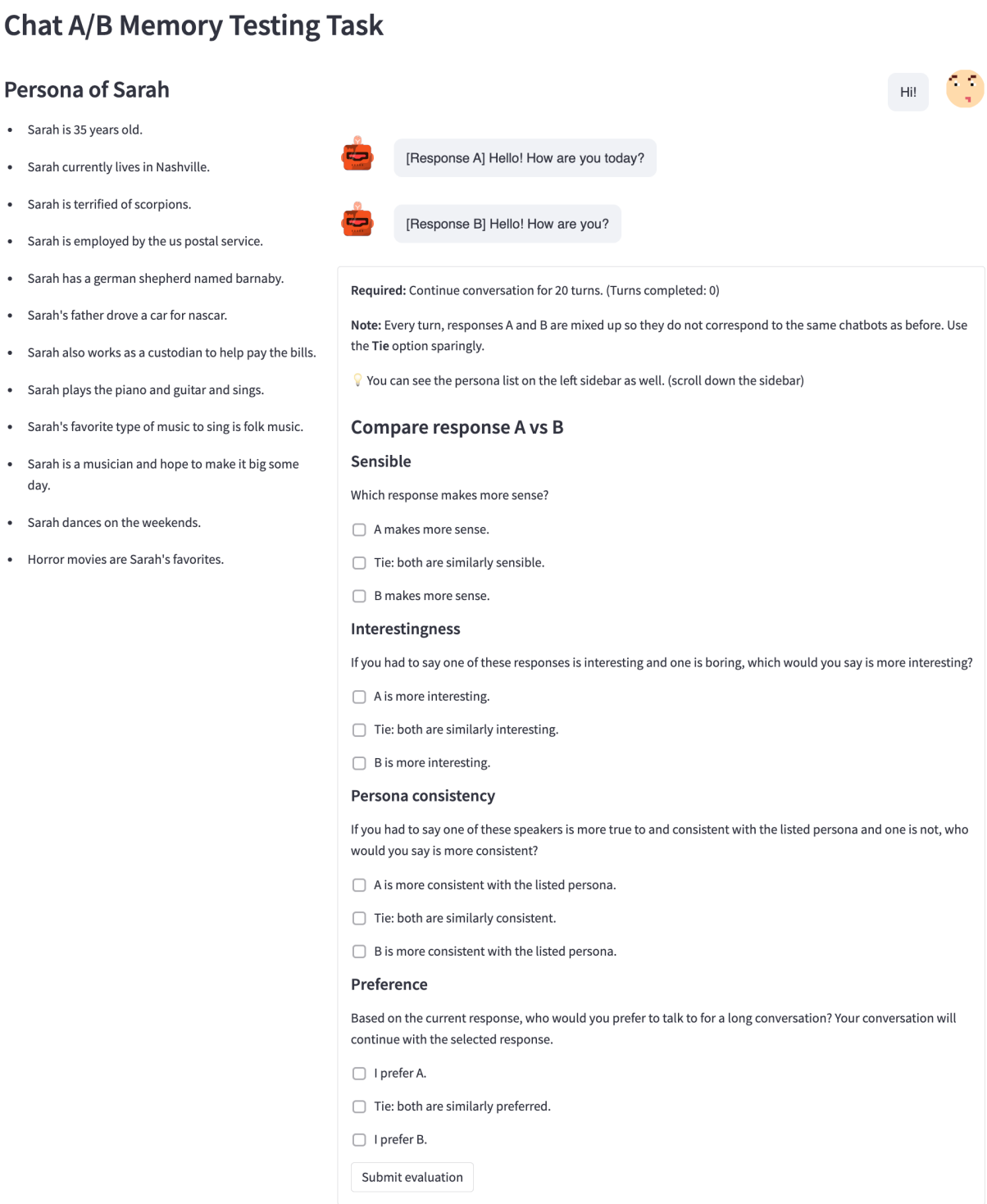


图4：成对模型比较的评估表。

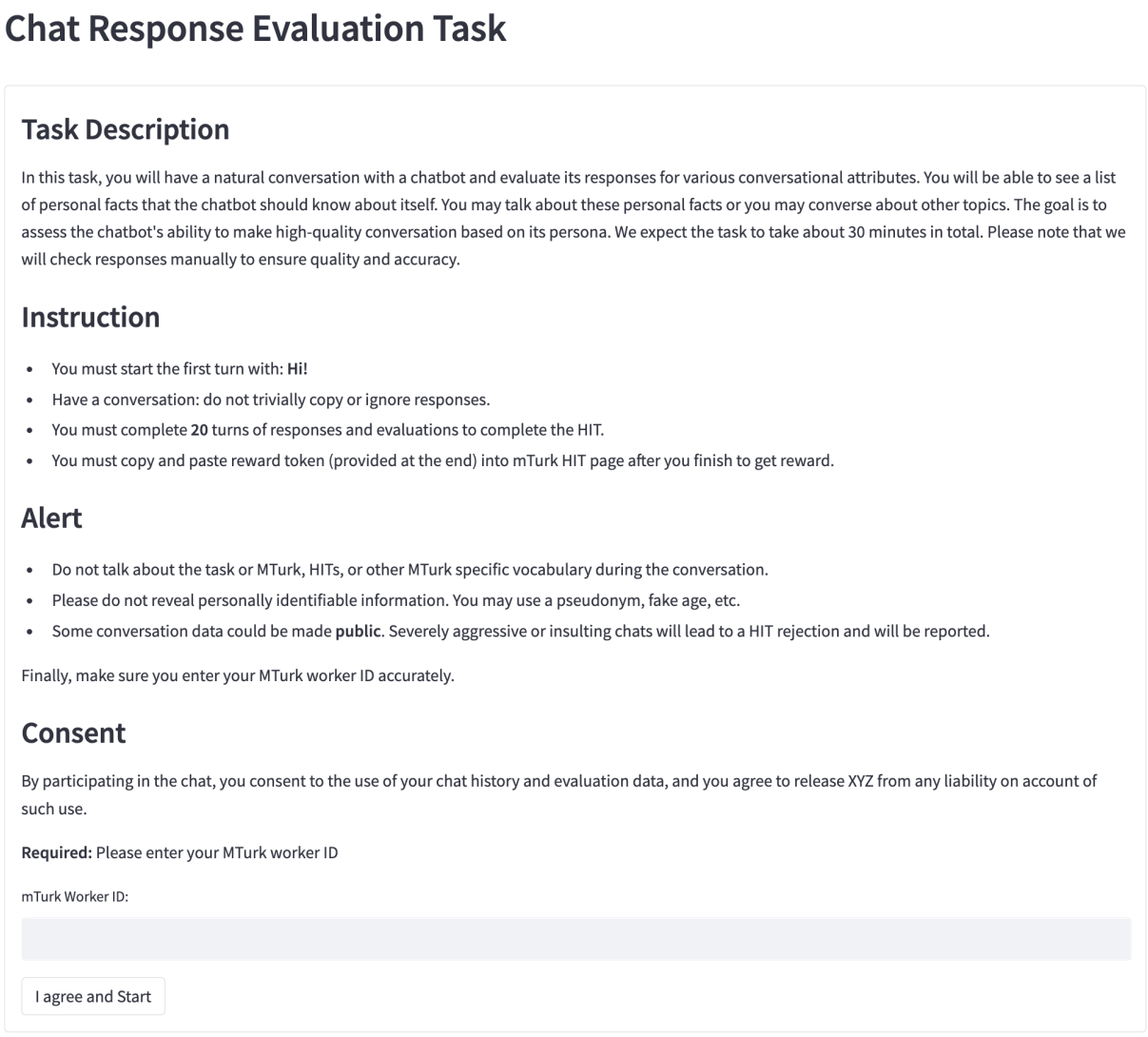


图 5：在评估者开始评估之前，我们会向他们展示此页面，告知任务内容并征得他们对数据使用的同意。

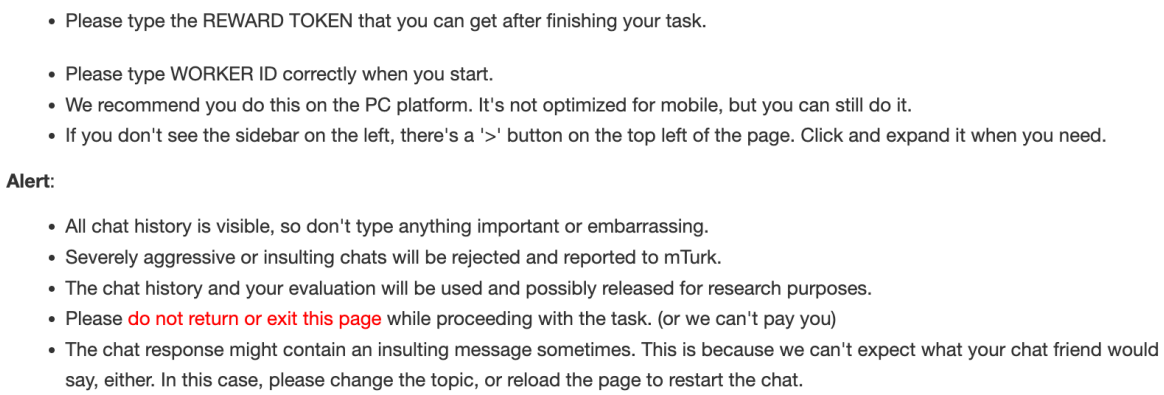


图 6：MTurk HIT 页面中的额外提示。