arXiv:2304.00350v1 [cs.CL] 1 Apr 2023

当众包人群遇上角色设定：构建大规模开放域角色对话语料库

Won Ik Cho[*1†]，**Yoon Kyung Lee**[*2]，**Seoyeon Bae**[2],**Jihwan Kim**[1†],
Sangah Park[3‡]，**Moosung Kim**[4],**Sowon Hahn**[2],**Nam Soo Kim**[1]
[1] 韩国首尔大学电气与计算机工程系 [2] 韩国首尔大学心理学系[3] 韩国首尔
DeepNatural AI [4] 韩国板桥 Smilegate AI

摘要

构建自然语言数据集需要谨慎，因为词语语义很容易受到文本细微变化或所标注概念定义的影响。这种倾向既可见于问答和对话生成等生成式任务，也可见于主题分类或情感分析等构建基于类别划分的语料库的任务。开放域对话涉及两名或更多众包工作者围绕任意话题自由交谈，而收集此类数据尤其困难，原因有二：1）出于隐私顾虑，数据集应当通过“创作”而非“获取”来构建；2）有偿创作此类对话时，众包工作者的表现可能与其在真实环境中的行为有所不同。本研究在构建大规模开放领域角色对话语料库的过程中应对这些议题，其中，角色设定意味着对话由几名具有固定角色设定的演员与来自不特定众包人群的用户侧工作人员共同开展。

引言

现代对话代理通常会被赋予特定的角色设定，使用户觉得它们具有人类般的特征（Seeger、Pfeiffer 和 Heinzl（2021））。在设计过程中，还会考虑个性、长期记忆和共情特质等认知与情感特征，以增强用户与这些代理互动的投入感，从而促成更长时间的对话（Schuetzler 等（2018））。

具有特定角色设定的代理经常会面临棘手的情境：它们需要记住对话历史，并保持角色设定一致，以回答用户的问题（Xu、Szlam 和 Weston，2021年）。例如，如果代理被设定为一个喜欢弹钢琴的高中男生，那么当用户问他作为职业女性的经历时，就可能产生冲突。此外，在一段对话中，代理对对话伙伴的态度通常也应保持一致。如果对话一直以友好的方式进行，气氛却突然转冷，可能会令人尴尬。所有这些因素都有助于形成流畅的对话流程。

通常，要让代理流畅地延续对话，需要运用各种启发式方法，例如理解问题、生成答案、参与闲聊、绕过问题或使用回退指令（McTear（2018））。这些启发式方法与对话代理的功能层面有关，而非代理的认知和心理层面，例如维持“自我”或避免与先前话语相矛盾（Shuster 等，2021年）。因此，我们认为，构建一个能记住自身角色并在整个对话中保持一致的对话数据集，将有助于对话代理的实现与分析。然而，从头开始构建对话数据集，并不像收集日常对话那样简单。

为什么创建对话数据集很困难

创建包含两名或更多参与者的对话数据集是一项颇具挑战的工作，因为两名参与者要顺利展开对话，需要满足若干条件，例如确认共同基础（Stalnaker，2002年）、建立融洽关系（Cassell、Gill 和 Tepper，2007年），以及协调沟通风格（Tannen 等，2005年）。对于没有特定话题的开放领域对话，构建过程会复杂得多，因为参与者的共同基础和兴趣通常各不相同，导致对话迅速结束。

难以将任务导向型对话中采用的策略用于开放领域对话，例如基于手册的对话（Wizard-of-Oz）（Wen 等，2017年）或自我对弈（Shah 等，2018年）。手册可能无法涵盖开放领域对话中可能出现的各种话题（Godfrey、Holliman 和 McDaniel，1992年；Li 等，2017年；Zhang 等，2018），而自我对弈可能受到内容和自然度方面的限制。因此，在构建开放领域对话时，不可避免地需要多位说话者参与（Dinan 等，2019年；Rashkin 等，2019年；Roller 等，2021年；Xu、Szlam 和 Weston，2021年）。

开放领域角色对话

如果多位说话者参与一场对话，通常的对话会以参与者围绕日常生活话题交谈、交换信息或增强社会联结的方式进行（Li 等，2017年）。这种设定赋予参与者同等的义务，双方都努力引导对话，以免对话陷入缺乏有意义的互动的尴尬

[*]这些作者贡献相同。
[†]毕业前完成的工作。
[‡]在 DeepNatural AI 任职期间完成的工作。版权所有 © 2022 年，美国人工智能促进协会（www.aaai.org）。保留所有权利。

氛围。然而，如果其中一位参与者被赋予与那些先与其交谈的人进行对话的义务，就像服务中的对话代理或在线游戏中的非玩家角色一样，说话者之间的关系和角色可能会不同于默认设定。我们将其视为角色对话的一种特殊形式。

在文献中，角色人设对话被视为具有代表性的开放域互动，并对此进行了研究和构建（Roller 等，2020）。在这方面，研究者采用参与者之间的双人扮演方法，创建了多个角色对话数据集。还有一些角色对话数据集采用了参与者之间的双人扮演策略（Zhang 等，2018）。尽管“角色设定”通常被视为说话者的一种长期社会特征，但大多数研究都侧重于在每段对话开始时设定一个特定情境及若干属性，然后要求参与者在对话阶段遵循给定条件（Zhang 等，2018；Xu、Szlam 和 Weston，2021）。这使得人们能够遵循大量角色设定条件并开展对话，但未必涵盖另一位参与者可能感受到融洽关系或友善度的长期社会特征。我们认为，要创建具有长期社会特征的角色对话数据集，还需要采用另一种角色扮演形式。同时，即使参与者彼此感到友善，也应避免他们对彼此粗鲁无礼或造成冒犯；鉴于这种融洽关系是在人工情境中建立的，应保护参与者免受人际互动可能造成的心理伤害。

我们的研究

我们的研究致力于创建一个大规模、开放领域的韩语角色对话数据集，确保参与者之间进行安全且不肤浅的对话。我们的研究分为三个方面：

设置 我们首先假设这样一种设置：角色参与者与用户参与者进行交谈（Zhang 等，2018；Roller 等，2020），并且用户根据角色档案的详细信息首先发起对话。我们为双方准备了对话指南，同时赋予角色一方更多义务和主导角色。角色参与者应接受访谈，以便被聘为演员并参与对话。

收集 完成所有设置后，我们在众包平台的工作者社区中进行宣传，招募对与演员交谈感兴趣的用户参与者。对话由用户的消息发起，我们构建了一个网络应用程序，使参与者能够进行对话，同时允许众包平台的管理者检查对话，以进行角色与用户之间的内容审核。

分析 每次对话结束后，我们都会让参与者填写一份调查表，询问他们对这次对话的满意度，以及趣味性、友善度和联结感等方面的情况。此外，我们还会访谈演员和主持人，了解他们对对话流程的看法，包括他们觉得有趣或遇到困难的地方。这一过程为我们的策略提供了佐证，也使我们的方案具有可持续性。

通过这些设置、数据收集方案和分析，我们旨在表明，我们的策略能够让一小群演员和众包用户参与者共同构建大规模的开放领域角色对话语料库，并让双方在对话即收益范式中都获得满意的体验。我们在此提出三个研究问题：

- **研究问题 1：** 在构建成功的对话数据集时，应考虑哪些因素？哪些因素对角色人设对话至关重要，又有哪些因素会使这一过程充满挑战？
- **研究问题 2：** 主持人在大规模对话数据集构建中扮演什么角色？这与其他自然语言处理数据集有何不同？
- **研究问题 3：** 上述考量能否帮助实现预期的构建流程和产出？它们能否在确保参与者获得满意体验的同时，保留语料库中多样化的角色特征？

相关工作

对话数据集

对话长期以来一直备受关注，其构建方案也发生了很大变化。早期方法从真实对话中摘录脚本，仅做少量清理；而近期方法则从头构建对话数据集，并添加一些相关标签。

Switchboard Switchboard（Godfrey、Holliman 和 McDaniel，1992）是早期的多说话人会话语音和文本语料库，由围绕预设话题开展的电话对话组成。该语料库收录了来自美国各地的 500 位说话者进行的 2,500 段对话，这些对话通过德州仪器的 T1 线路自动采集。此外，每段录音中的词语都经过时间对齐，以便进行词级转录。在此基础上，人们还开展了对话行为标注和语用分析等后续研究，但自其发表至今已约 30 年，因此该数据集略显过时。此外，无论电话服务用户是否已同意，直接从线路中收集随机对话都使该数据集容易面临隐私和许可问题。

DailyDialog DailyDialog（Li 等，2017年）是一个由人工撰写、噪声较少的多轮对话集合。该数据集反映了日常交流方式，涵盖与日常生活相关的各种话题，包括校园生活、工作、旅游、人际关系等。此外，话语还标注了情感和对话行为，为数据集分析提供了丰富的信息。然而，原始数据是从各种网站抓取的，这些网站旨在帮助英语学习者练习日常生活中的英语对话，因此该数据集相比其他从社交媒体抓取的数据集更加正式。此外，对话不一定能反映每位说话者的个人特征，导致其中大多数

对话彼此相似。在这种情况下，由于缺少展示出的角色设定，很难判断说话者在对话过程中是否彼此投入、感到趣味性或产生共情，因为这可能导致缺乏融洽关系，从而使内容流于表面。

开放领域角色对话研究

开放域对话代理生成的自由闲聊正变得越来越连贯、越来越有吸引力。然而，现有模型无法生成较长的对话（不超过五轮），只会重复用户的评论，还会忘记用户先前说过的话（Xu、Szlam 和 Weston，2021）。众所周知，生成模型还会通过生成错误信息来臆造知识，而且无法理解其有限模型参数之外的外部知识（Roller 等，2021）。近期的工作包括由众包工作者参与的角色扮演游戏（Shuster 等（2021））、创建一千多组人设（Zhang 等，2018）、维持较长的对话（Xu、Szlam 和 Weston，2021），以及生成共情式回应（Rashkin 等，2019；Smith 等，2020）。

PersonaChat PersonaChat 数据集（Zhang 等，2018）是一组个人闲聊对话。每对说话者都基于特定的角色设定进行对话。该数据集分为三个阶段：人设、修订后的人设和角色聊天。一个角色设定由五个句子表达，每个句子都会经过改写，以避免重复。之后，两对众包工作者通过扮演特定角色进行交谈，生成了数据（超过 16 万条话语）。

共情对话和维基百科向导共情对话（Rashkin 等，2019 年）包含带有共情表达的对话数据。当听到说话者在特定情境中使用特定的情感词语时，倾听者会以共情方式回应。维基百科向导（Dinan 等，2019年）包含从维基百科页面抓取的文本。担任专家角色的工作人员围绕多个话题进行交流，并将每条话语与一个维基百科句子关联起来。因此，在这些数据上训练的模型能够将来自互联网的知识运用于自由聊天。开放域聊天机器人的关键在于维持具有吸引力且持续时间较长的对话。包含全部三个组成部分的模型生成了更加自然的回应，尽管定性指标仍然较差。要让拥有自然且连贯对话能力的对话代理满足用户，仍然十分困难。目前还不足以断定，用户参与度和类人程度的高分就能代表用户满意度。因此，为了维持引人入胜且富有吸引力的互动，应当适当明确交互式代理的更高层次概念，例如个性和共情品质。

多轮次聊天 多轮次聊天（Xu、Szlam 和 Weston，2021年）是一个由众包工作者进行的多轮次聊天集合，其中对话会被中断并重新开始，间隔时间短至 7 小时，长至 7 天；相比之下，此前的简短对话持续时间不到一小时。现有开放域对话模型所学习的讨论越短，对话就越无法继续，记忆也会随之丢失。因此，需要能够处理更长上下文的模型。为了回忆先前会话中讨论的内容，每个情节中的对话还标注了简短回顾。这样，后续跟进的另一名众包工作者便能迅速回忆起之前的话语。使用上述数据集训练的生成模型能够生成比以往模型更连贯的话语。然而，仅训练所推荐的长期记忆和摘要，并不能直接提高用户对吸引力的评分。对话中哪些方面能够吸引或引起用户兴趣，仍需进行广泛研究。

方法

我们的数据构建流程由四个步骤组成：对话指南、试点研究、演员[1] 招募，以及对话收集[2]。

对话指南

令角色设定指角色参与者，令用户指众包工作的用户参与者。角色设定侧和用户侧都有相应的对话指南。两侧的整体组织方式相似，但角色设定侧包含更多规范。

个人资料角色设定侧需要一份能够代表参与者的个人资料，其涵盖范围略广于文献中角色设定的概念（Zhang 等，2018；Roller 等，2020）。个人资料包括角色设定的姓名、图片，以及三条以上关于该角色设定的信息。此时，角色设定可以不受其生命性或人性影响而自由选择，但参与者必须准备一份能够在保持角色设定一致性的同时进行对话的个人资料。此外，个人资料不得包含能够识别参与者个人信息，或引发政治或社会争议的内容。个人资料图片应恰当地呈现角色设定，但不得使用违反许可政策的素材。角色档案示例（*HungrySloth*）见图1。

角色设定准备的个人资料会公布给用户，用户根据个人资料决定是否与每个人设进行对话。决定与某个人设交谈的用户应先发起对话，而该人设有义务与这类用户继续对话。

注意事项与禁忌 在对话中，用户和角色设定会收到类似的指示。以下是双方通常应遵循的注意事项与禁忌，其中一些灵感来自会话准则（Grice，1975）。所有参与者都应：

[1]我们不指定参与者是男性还是女性演员，但此处使用“演员 ”一词统称两者，无意偏向任何生理性别。[2]“对话”和“交谈”之间存在细微差别，但此处两词可互换使用。

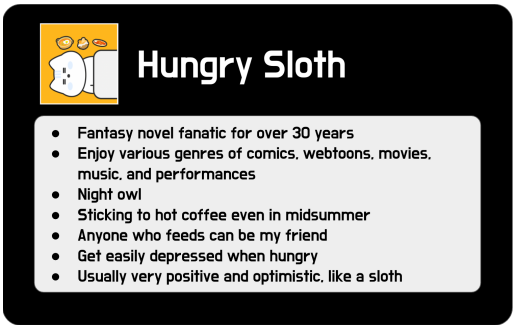


图1：角色档案示例。

.

• 在对话中保持尊重。请记住，你交谈的对象也是一个具有情感和思想的个体。

• 无论是否使用敬语，都要彼此礼貌相待[3]。

• 对话时要真诚。尽可能记住并提醒彼此对话的上下文。

• 对话时要相互配合。聊聊双方都可能感兴趣或觉得愉快的话题。

此外，参与者不应：

• 除非对方主动透露，否则不要询问对方的身份。请记住，对话是远程以文字形式进行的，因此对方可能不愿透露自己的人口统计属性或个人信息。

• 在对话中包含敏感信息（自己或他人的个人信息、政治立场等）。请记住，如果语料库公开发布或用作训练数据，此类内容可能造成伤害。

• 发表带有社会偏见或有害表达的仇恨言论。社会偏见是指根据一个人的身份，将其视为少数群体或代表性不足群体而加以评判（Waseem and Hovy，2016年）；有害表达则包括侮辱性言论、性骚扰，以及讽刺和不道德表达等冒犯性语言[4]（Davidson et al.，2017年；Moon、Cho 和 Lee，2020年）。

对话策略 为帮助引导对话流畅进行，我们向角色设定提供了特别建议。以下策略涵盖了多种技巧，可用于应对尴尬、令人恼火的情境，或对话无法顺利进行的情况。

• 在不泄露角色设定的个人信息或身份的前提下，充分利用角色设定正在经历的情境。

• 浏览与用户的对话历史，并思考接下来推进对话的故事情节。

• 考虑用户会积极回应哪类话题。

• 如果设定了特定职业（如教授、演员、导演等）的角色档案，就要利用背景信息或常识来进行对话。

• 设定好独特的人设后，务必确保对话内容不会枯竭，也不要过早地因过度强调这一点而削弱对话张力。

此外，虽然人设应当回应用户一方的对话，但也要求他们讲述许多自己的故事。这是为了检验他们能否始终如一地维持人设，也是一个在对话中明确或隐含地插入有关自身内容的过程。

其他事项 以下说明是给角色参与者的，他们承担着更多引导对话的责任。

• 以角色设定的视角参与对话，而不是以现实世界中的自我身份参与。

• 对话中的行为或言谈至少应符合参与者在对话开始前确定并告知的角色设定。

• 如果用户发来的消息很少或已经停止发送，应不时主动发送消息[5]。建议每三次主动发送一次。

为便于与人设进行对话，也为用户参与者提供了一些指导。

• 可以提及特定的名人、品牌或艺术作品（电影、音乐等），但不得带有强烈的批评或广告意图。

• 不要以难以维持对话张力的话题开启对话。

• 任何情况下都不允许侮辱或骚扰；即使某些表达并无冒犯之意，也不得使用可能被误解为脏话的说法。

上述策略和说明适用于通用对话指南，不受合作对象用户群体或众包平台的影响。关于角色扮演者和用户侧工作人员参与时长及聊天限制的补充说明，将在主要对话收集阶段中介绍。

试点研究

本研究的目标之一，是辨析特定角色设定与用户之间对话的特点。因此，在试点研究中，我们通过少量参与者进行的少量对话，验证了对话指南是否恰当。

[3]韩语中的一个重要问题是，社会关系会影响话语的句法成分（Byon，2006年）；而对于没有此类句法成分的其他语言，这或许指的是正式程度。[4]我们参考了 PyCon KR 行为准则 https://pycon.kr/ 2020/about/coc/，以制定进一步的指南。

[5]在韩语中，这称为“선톡”（seon-tok），反映了说话者（或发送者）对受话者（接收者）的兴趣。

参与者 试点研究邀请了三位角色设定（下文称为演员）和五位用户（下文称为工作人员）参与者[6]，主要是对该项目感兴趣的研究人员和本科生。每位演员都准备了一份个人资料，并通过与作者进行对话来熟悉相关说明。工作人员也获得了对话手册。

对话 我们在试点研究中使用了一款公开可用的即时通讯软件。我们使用了Kakao公司提供的KakaoTalk[7]，这是一款大多数韩国公民都熟悉的移动端和桌面端应用，并提供匿名聊天服务。具体而言，在公开资料聊天中，用户可以设置用于匿名聊天的个人资料，包括姓名、照片和简短简介。每位演员创建的个人资料会展示给用户，工作人员据此开始对话。

每段对话向演员支付12,000作为报酬，而工作人员则根据自身兴趣自愿参与。对话的边界定义为对话明确结束并转入下一段对话起点的时刻。

调查 工作人员在每次对话结束时，就友善度、喜爱度、联结感和趣味性开展评分调查[8]。共进行了20段对话，并通过对上述调查进行分析得出了以下结果。

- 如果对话中展现出引人注目的观点，工作人员就会觉得这段对话或角色设定很讨喜。
- 如果对话中展现出大量共情，例如分享体验、表达赞赏或认同等，工作人员就会觉得这段对话或角色设定很讨喜。

- 自我披露、提问与共情之间保持平衡，对流畅的对话至关重要。
- 角色设定在高层次上的一致性和鲜明特征，并不一定能保证其趣味性和讨喜度。

演员招募

在演员招募过程中，我们为主要的对话收集招募角色参与者。在这一过程中，共有 20 名演员申请，经过访谈后选出了 11 名。

在这一阶段，为了实现可扩展且可复现的数据收集，会有一位主持人 负责联系工作人员和研究人员（作者）。这位主持人相当于众包平台的经理，该平台设有人力资源池，参与者从中招募而来。主持人协助招募参与者并与他们沟通，同时管理招募工作和工作者培训（Yang、Jang和Cho，2022年）。

[6]“参与者”一词用于指代参与实验的人群。在主要构建阶段，该词与“演员”或“工作人员”互换使用；这些人员通过招募、雇用或众包方式获得，以便构建可扩展的语料库。[7] https://play.google.com/store/apps/details?id=com.kakao.talk[8]请注意，此调查流程不同于主要对话收集过程中开展的调查。



图 2：面向众包人群的公告，其中包含角色档案。

.

具体而言，主持人会访谈每位角色设定申请者，研究人员则根据访谈数据作出决定。演员筛选时会考虑以下标准：

- 角色设定是否讨人喜欢，或者角色设定概念是否始终保持一致?

- 对话中是否呈现出良好的乒乓式互动，并能流畅地运用回应式复述或反问?
- 是否具备长时间引导紧凑（不松散）对话的能力?

- 自我披露是否充分且自然?
- 擅长表现出同理心或真诚吗?

经过评审后，研究人员最终选定了11位演员。每位演员的个人资料（申请时提交）都会公布给工作人员群体，让工作人员了解该项目，并对各个角色设定产生兴趣（图2）。

对话收集

在主要收集阶段，招募的演员会在真实场景中与工作人员交谈。我们重新设计了聊天平台，并针对参与时长、报酬等方面制定了其他几项适用于大规模构建的说明。

平台与流程 我们采用的用户界面与试点研究中的不同（图3）。这是由众包平台开发的一款应用程序，其结构允许工作人员与演员开启一对一聊天室。此外，主持人可以监控这些开放的聊天室，并利用该监控系统处罚言行不当的工作人员，或调节整体工作量。具体来说，输入句子后按回车键会换行，但只有当对方开始聊天时，轮次才会转到对方。说话者信息会与换行信息一同导出，时间戳也会随对话一同记录。

工作人员阅读招募信息，并参与面向各个角色设定开放的项目。因此，分配给每位演员的工作量

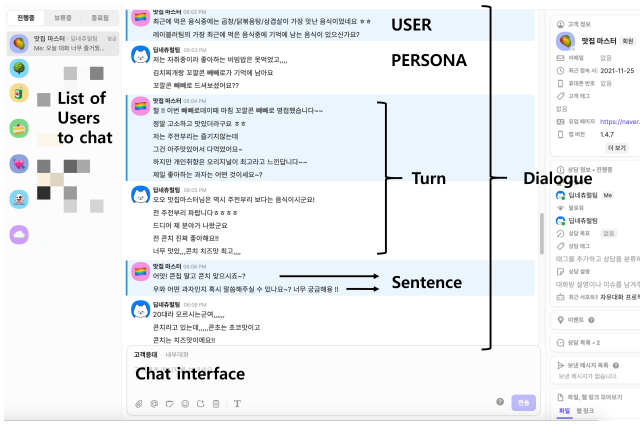


图3：聊天用户界面的屏幕截图。

.

各不相同，工作人员也没有义务与所有演员进行对话。不过，考虑到演员与用户之间参与者数量的不平衡，我们告知工作人员，演员的回复有时可能会延迟。此外，还要求他们不要催促演员回复，并在每次对话结束时留下结束语（例如，我现在要去吃晚饭了！），以免演员一直等待后续对话。

基本上，角色设定和用户端的对话指南与试点研究中的类似。不过，为了确保数据的一致性和可持续性，我们向工作人员提供了以下注意事项。

- 允许出现拼写错误、勘误，以及聊天过程中自然产生的不合语法的句子。
- 简单的感叹词或填充词（如 ㅋㅋ/ㅎㅎ，即“哈哈”）可以使用，但不视为有效的句子输入。
- 请避免使用照片、表情符号、颜文字等，因为它们无法在多种数据格式中保持一致的显示效果。

接下来，我们将介绍为适应众包条件下的对话流程而添加的一些其他功能。

人工时长 我们以 Xu、Szlam 和 Weston（2021年）为基准，并指导演员在至少进行四轮次对话后插入人工时长。人工时长可以从数小时（1-7）自由延长至数天（1-7），参与者应在插入该时长后继续进行恰当的对话。

引入人工时长有两个目的。首先，人工时长使我们能够生成长期对话，而在有限时间内生成对话时，这类对话难免受到限制。尽管插入的时长是人为设定的，参与者仍能复现他们在过去对话中习得的行为。其次，允许角色设定一方拥有人工时长，可以让演员在不愿继续对话时暂停一下。这有助于缓解角色扮演者疲劳，也可以被视为向用户发出的一种警告信息。

我们不会强制要求角色扮演者为每位用户同步相同的对话历史，因为除了保留个人资料信息之外，所有对话都被视为独立事件。例如，角色设定可能会在与用户 A 对话时透露口语化事件或旅行经历，但与用户 B 对话时则不必保留这些设定。这是因为每段对话都发生在异步时空中。通过插入人工时长，我们弥补了独立场景中可能出现的时间窗口狭窄这一限制。此时，为了让对话自然流畅，我们要求用户考虑实际插入的时间，继续进行对话，不必因演员提及的人工时长而感到尴尬。

奖励系统 在众包中，我们指导演员在特定期限内（约一个半月）完成一定数量的对话（每位演员300次），或者按每次对话获得特定报酬。因此，他们的工作被视为有偿工作，其生产率较少受到与他人交谈意愿的影响。然而，用户参与者会在存在适当奖励的情况下参与对话。因此，我们通过严格定义对话、轮次，和句子（或类句子片段），建立了适当的奖励系统。

对话 由15至30轮次对话组成，包括完整的文本行，从工作人员的问候开始，到对话的结束部分为止。其中至少包含三次人工时长。轮次 指工作人员和演员随后说出的两组类句子片段。在此，类句子片段可以由多行句子组成，而其是否为语法完整的句子并不重要。句子 表示组成类句子片段的多行文本。因此，一轮次最终由工作人员和演员说出的两行以上句子组成。

工作人员完成五段对话并填写调查表后，即可获得奖励，每段对话约可获得 1.2 美元。调查表参考了试点阶段使用的版本，并包含对话另一方的评价。演员也需要填写类似的调查表，但对演员而言，完成调查表并不直接关系到奖励。

每段对话中，工作人员必须至少进行15轮对话。作为激励措施，完成15轮对话后，还可以继续聊15轮。此时，工作人员每多聊五轮即可获得0.4美元的额外奖励。通过这种方式，我们设计了一套系统，既能保证对话达到最低轮数，也能让工作人员根据自己的兴趣或热情获得额外奖励。

调查表 与试点研究类似，我们要求参与者在每次对话后填写调查表。下面介绍我们希望通过调查表了解的内容。我们开展了一项投票，让参与者评估彼此的态度以及互动内容。我们使用从1（“非常低”）到4（“非常高”）的量表，衡量参与者与对方对话时的趣味性

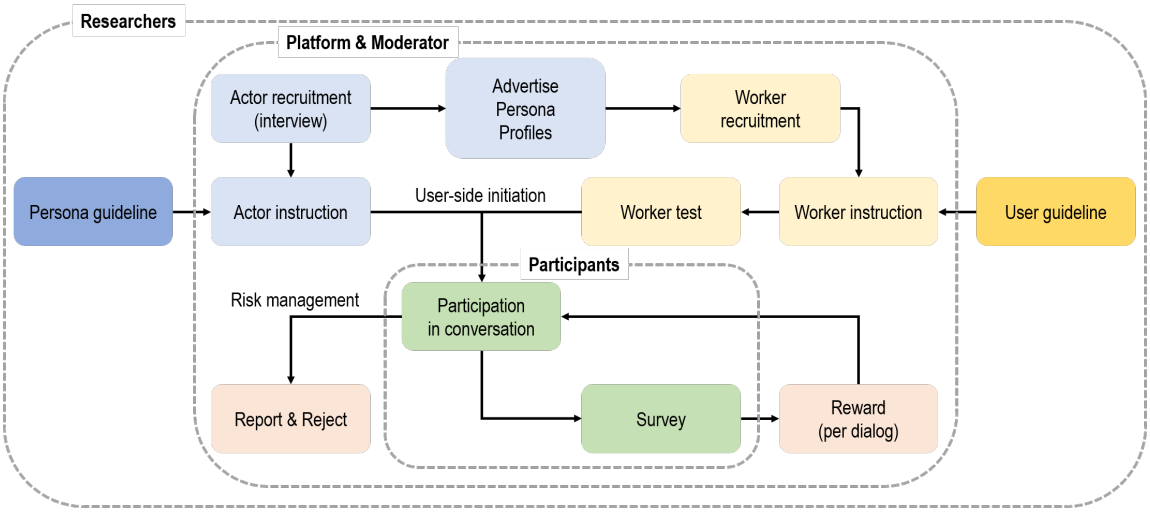


图4：所提出的构建方法的整体流程。这里，演员指参与角色设定对话的参与者，角色设定；工作人员则指扮演用户的一方。只有演员需要设置具体的个人资料，而对话由工作人员发起。

（趣味性；“与对话伙伴的对话有多有趣？”）、对话流程的自然度（自然度；“对话流程有多自然？”），以及对话伙伴个性的吸引力（吸引力；“你的对话伙伴有多吸引人？”）。此外，参与者还需要评估彼此的共情能力。我们采用了巴雷特-伦纳德关系量表（BLRI），该量表由咨询师和来访者用于评估彼此在咨询中的共情能力（Barrett-Lennard 1981年，2015年）。BLRI调查采用从1（“极低”）到4（“非常高”）的量表。

项目流程

整个项目的流程如图4所示。参与其中的利益相关方共有三类：制定角色设定和用户对话指南的**研究人员**、招募演员和工作人员并指导其进行角色对话的**平台（主持人）**，以及参与构建工作的**参与者**。所有对话均由用户发起。演员与工作人员之间的部分对话会在报告与拒绝流程后终止；否则，参与者完成调查后即可获得奖励。

开始对话前，工作人员会了解相关说明，并通过熟悉角色设定介绍来准备对话。对话以问候语开场，建议问候语体现各个角色设定的特点，而不是简单地打声招呼。

对话持续15轮。每轮由工作人员和演员各说的话组成；一方说完后，另一方再接着聊。无论按下回车键换行多少次，这样的聊天都视为一个句子。

如果工作人员决定退出对话，就会结束对话。有时，演员如果感到疲倦，或因与工作人员交谈而感到不自在，也会结束对话。双方都必须在每次对话后完成调查。工作人员完成调查表后会获得奖励。当演员完成分配给每人的对话数量时，整个项目便结束，每个角色设定约有 300 段对话。

访谈

为确保所提出的方案具有可复现性，我们访谈了平台主持人以及四位扮演角色设定（#1 -#4）的演员。我们对访谈内容进行了总结，以了解参与者对所提出方案的评价。访谈所用的问题集及完整文本见附录。

扮演者反馈

角色对话的特点 与平常和朋友或家人的对话不同，#1 感到有义务继续对话。#1 和 #2 觉得这更像是在与新人或同事交谈，而不是与朋友交谈，因此会更加谨慎地选择措辞（#2），或进行更仔细的自我反思（#4）。

与真实用户交谈的挑战 演员因必须同时参与大量对话而感到负担沉重（#2、#3）。面对不友善、粗鲁或欠考虑的交流对象时，他们也会遇到一些挑战（#1）；有些用户甚至过于悲观，或让演员感到沮丧（#2、#4）。有些用户总是提出毫无意义的问题（例如，你现在在做什么？）（#2），或不断追问（#2、#4），让演员觉得自己像是在接受审问（#4）。有些用户甚至无视规则或不听通知，并在人工时间结束后仍尴尬地继续对话（#2）。

对人工时长的看法 演员使用人工时长来避免无聊的对话（#1）、获得休息时间，例如午餐时间（#2），或让对话与其他工作并行进行（#4）。甚至还有一个案例，其中

对话情境或流程需要人工时长（例如，虚拟参观艺术展览）（#3）。其中一些人希望减少强制插入的次数（#2、#3），并获得更自由的使用方式。

塑造角色设定的难点 演员根据用户的个人资料，难以回答用户提出的所有相同问题（#3、#4）。他们必须做大量准备，才能维持自己的角色设定概念（#1、#3）。#1 准备了一份问答列表，#3 则学习了一些与自身个人资料相关的新知识。#2 想要改变自己的角色设定，因为其概念过于宽泛；#4 则难以在不同角色之间找到平衡。

项目在现实生活中的影响 一些演员表示，他们逐渐适应了自己设定的人设特征（#1、#2）。#1 有时会不经意地把一些琐事也告诉朋友，因为他们觉得自己就像项目中的用户；#4 也因为这个项目，开始在现实对话中立即作答。此外，得益于交流对象的建议，他们拓展了个人爱好或艺术品味（#2、#3）。

主持人的反馈

组织大规模角色对话收集的挑战主持人认为，最大的挑战在于激励演员并理解他们的情感。演员在与用户参与者进行对话时感到困难，通常会通过私信向主持人倾诉。主持人表示，当工作人员以演员觉得恼人的方式与他们交谈时，演员尤其会感到为难；但为了符合自己的角色设定，他们又不能表现出不悦。此时，主持人既要体谅他们的困境，也要提供切实可行的解决方案。主持人还认为，如果演员不坦率地表达意见或情感，就很难了解他们的内在状态并提出解决方案。

演员招募与报酬方面的建议本项目的演员招募由研究人员和主持人共同负责，但访谈由主持人进行。主持人表示，如果演员的角色设定与其真实个性差异很大，沟通就会变得困难；鼓励演员在塑造角色时保留一定程度的自我披露或分享意见，或许有助于改善沟通。在报酬方面，主持人表示，演员更愿意按对话计酬，而不是在固定期限内完成指定数量的工作。主要原因是，固定工作安排会带来目标达成压力，而按对话计酬则能让时间分配更加灵活。

项目的可复现性主持人报告称，主持人在该项目中的工作量约占 10% 至 15%，而其最重要的职责是促进工作人员之间的沟通，并关心他们的心理状态。这得出的结论是，主持人最重要的品质是能够体谅工作人员遇到的不便并给予安慰。找到合适的主持人人选，并让其了解任务，以便与研究人员和工作人员顺畅沟通，将有助于建设项目顺利开展。

实验

完成语料库的整体构建后，我们开展了若干实验，以考察数据集的特征。首先，我们查看从调查结果中获得的总体分布和语料库统计数据。接下来，我们使用词向量聚类，考察关于各个角色设定的对话内容是否反映了其个人资料。最后，我们检验大型语言模型是否能够进行少样本对话生成，这意味着该语料库可以按照我们的预期方式加以利用。

调查分析

我们分析了参与者在每次对话后报告的调查结果。共分析了1,658份调查数据。我们使用斯皮尔曼相关性（斯皮尔曼 1904年）来衡量对话统计数据中的因素与调查结果之间的相关性。

对话双方的体验并不相同。在七项属性（趣味性、自然度、参与度、BLRI、讨喜度、尊重度和联结感）中，演员和工作人员报告的相同情感结果均未显示出显著相关性（即，用户与角色设定对于“与对方进行对话有多有趣？”这一问题的回答之间的相关性为：0.044，p值为：0.073）。然而，在各自回答中的七项属性之间，部分属性存在较高相关性（即，用户对于“与对方进行对话有多有趣？”和“对话过程中对方有多讨喜？”这两个问题的回答之间的相关性为：0.708，p值为：1.59e-252）。

由于要求演员在对话进展不顺畅时使用停顿，因此，演员停顿得越多，就越容易对对话产生负面感受（即，角色设定对用户的自然度感受与停顿次数之间的相关性为：-0.228，p值：5.35e-21）。值得注意的是，当停顿次数较多时，工作人员的对话体验反而更为积极（即，用户对角色设定的自然度感受与停顿次数之间的相关性为：0.118，p值：1.54e-6）。

此外，我们使用韦尔奇t检验进行了进一步分析（韦尔奇 1947年）。演员认为，工作人员越年轻，搭档就越讨人喜欢（p值：0.041）。经历开放聊天的群体（占总数的75%）与未经历开放聊天的群体之间存在显著的年龄差异（p值：4.75e-27）。当工作人员经历开放聊天时，演员会觉得与工作人员的互动更加投入（p值：0.045）。

主题聚类

首先，我们使用开源Python库 MeCab 分词器（Kudo，2006）提取韩语名词（例如普通名词、专有名词等），以便



图5：词语的语义可视化。

随后，我们删除了一些停用词，包括依存名词、感叹词和前置词。接着，我们使用 Gensim 库（Rehurek 和 Sojka，2011）训练了 Word2Vec 模型（Mikolov 等，2013），以获得最终的分词词语列表（N = 452,560）。最后，我们使用 Skip-gram 创建词嵌入，并通过 K 均值聚类对嵌入向量进行分类。

图5展示了使用 t-SNE 对嵌入向量进行分类的结果（Van der Maaten 和 Hinton，2008）。最终的前8个主题簇如下：日常作息（话题1）、食物（话题2）、工作/教育或众包项目本身（话题3）、圣诞节和节日季（话题4）、反思性想法（话题5）、音乐（话题6）、大众媒体（话题7）以及旅行（话题8）。

少样本对话生成

除了简单列出所收集对话的量化特征外，我们还希望展示所得语料库如何应用于当今的少样本对话生成范式（Brown 等，2020）。为此，我们从集合中选取对话并进行少样本生成，以考察：i）生成的对话是否与原始脚本一致；ii）添加角色信息是否会带来不同结果；以及 iii）所创建的数据集是否有助于未来的对话生成系统。

以下是*Hungry sloth*（树懒，角色设定）与*Green carrot*（胡萝卜，用户）进行对话的原始脚本。请注意，粗体部分将用作后续生成过程的输入查询，斜体部分则是参与者创作的原始脚本。

CARROT：嗨，我联系你是因为你现在应该在线。
SLOTH，你在做什么呢？……还在睡觉吗？SLOTH：我在呢，CARROT，最近怎么样！
CARROT：终于等到你了！我还以为你睡了三天呢。
SLOTH：抱歉，我一直在处理一项紧急工作。CARROT：**哦，我明白了……那现在工作结束了吗？**
SLOTH： 是啊……不过，连续三天过着不同的生活模式，我的状态不太好……我需要点什么让自己振作起来CARROT： 要不要好好睡一觉？SLOTH： 我现在开始睡可以吗？哈哈 [3 天过去了]

为了了解大语言模型如何通过少样本学习生成对话，我们采用韩国 GPT-3 API Hy-perCLOVA[9]（Kim 等，2021年）设计了两个场景[10]。首先，我们根据上述对话的前四句话生成后续对话，然后将其与额外使用角色档案的情况进行比较。

CARROT/SLOTH： [同上]
CARROT： **哦，我明白了，那工作现在已经结束了 ？**
SLOTH： 是啊，我今天要休息一下CARROT： 干得好 *:)* 工作这么辛苦可能会伤到你——
自我。顺便说一句，这几天天气变得好冷*:(*SLOTH： 确实如此……

我们注意到，大模型给出的回答序列相当自然。在这里，我们将图1中的角色档案添加到对话开头。只需对个人资料文本进行简单增强，就能得到略有不同的输出，更好地体现生成对话中的角色设定。

HUNGRY SLOTH 的个人资料：痴迷奇幻小说逾30 年。喜欢各种类型的漫画、网络漫画、电影、音乐和表演。夜猫子。即使在盛夏也坚持喝热咖啡。任何愿意喂食的人都可以成为我的朋友。饿了就很容易沮丧。通常非常积极乐观，像树懒一样。

CARROT/SLOTH： [同上]
CARROT： **哦，我明白了，那工作现在结束了？** 你愿意和我共进晚餐吗？我请客SLOTH： 哦，你真好！！我现在饿死了CARROT： 耶，待会儿见 *:D*

请注意，SLOTH 的整体回应似乎反映了该角色设定本身的特征。在原始脚本中，考虑到此前已经连续几天有紧急工作，演员要求吃点能让自己振作起来的东西，并询问自己是否可以睡觉，这体现了其与 SLOTH 的相似之处。在没有角色档案的情况下生成的脚本中，SLOTH 休息了一会儿，而 CARROT 对此感到担心，呈现出与原始脚本相当的对话。相比之下，使用角色档案生成的脚本使该角色急于

[9]https://clovastudio.ncloud.com/playground[10]尽管高性能大语言模型（LLM）是在 2022年11月于 HCOMP 展示本文非存档版本之后才问世的，但我们没有对它们进行额外实验，因为它们在韩文文本方面的性能无法得到保证，而且其中一些模型会轻易地将自我表现为智能代理，这使得我们难以将它们应用于我们的数据集。

吃点东西（因为用户会请他/她吃饭），突出了其特征中“饥饿”的一面。

在生成的段落中，我们检查了生成的对话是否与原始版本一致，并发现添加角色信息突出了角色特征的另一面，从而使对话流程略有变化。原始脚本与生成脚本之间的一个区别是，与演员一方相比，用户一方的对话内容丰富得多。这并非预期的变化，而且模型似乎认为生成用户一方的脚本同等重要，甚至更为重要。不过，我们认为，根据当今的少样本和提示范式，这种多样性可以得到研究并被积极利用。我们相信，我们的数据集能够在验证其可行性方面发挥重要作用。

讨论

通过上述设置和数据收集方案，我们旨在展示我们的策略如何帮助利用一小群演员和众包用户参与者，构建一个大规模的开放领域角色对话语料库，并让两组参与者在对话即收益范式中都获得令人满意的体验。

研究问题 1： 构建成功的对话数据集时，应考虑哪些因素？哪些因素对角色对话很重要，又有哪些因素使这一过程充满挑战？

组织角色对话与组织日常对话有很大不同。尤其是角色设定仅分配给演员（由工作人员发起对话）时，妥善应对可能令演员感到尴尬或烦恼的意外和不受欢迎的情况至关重要。减少或缓解演员的心理疲劳非常重要，而允许对话持续一段人为设定的时长是解决方案之一。主持人在这方面也发挥了重要作用。

研究问题 2： 主持人在大规模对话数据集构建中扮演什么角色？这与其他自然语言处理数据集有何不同？

在自然语言数据集构建中，主持人的常见职责是了解研究人员的任务并指导工作人员，同时调解冲突和管理经费。在其他（语言学）标注任务中，适应上述职责可能有助于项目取得成功。然而，大规模对话数据集构建中的主持人可能还需要关注参与者之间的冲突，因为对话不可避免地涉及两名以上个体之间的互动。尤其是在角色对话中，演员需要控制自己对谈话对象的社交态度，这可能造成心理压力。主持人的职责是体谅他们面临的困难，有时提供解决方案并给予鼓励，促使演员继续对话。

研究问题 3： 上述考量是否有助于实现预期的构建流程和产出？它们能否在保留语料库中角色特征多样性的同时，确保工作人员获得满意的体验？

首先，从主题聚类（Van der Maaten 和 Hinton，2008）中，我们可以看出，聘用具有特定个人资料的角色扮演者，并让他们与不同用户进行对话，可以确保对话话题的多样性。尽管有两个主要背景因素：参与者处理多样话题的能力，以及招聘演员时对人设广泛覆盖面的考量，但我们认为，我们的方法可以避免数据集中只包含单一类型的对话。其次，从少样本学习示例（Brown 等，2020；Kim 等，2021）中，我们可以确认，人设信息能够编码到角色扮演者一方的回复生成过程中，这表明所构建的数据集在当今提示范式中的实用性。谈到工作人员的体验，调查结果显示，对话的趣味性以及工作人员年轻与否，会影响对话对象的讨喜度；而开放聊天的体验会让对话对象感受到吸引力。这表明，众包工作者对工作环境的熟悉程度会影响对话氛围，并可能最终影响满意度。我们认为，创作对话是一项具有代表性的工作，要想产出高质量成果，既需要工作人员通力合作，也需要他们拥有令人满意的工作体验；我们的结论也可为面临类似挑战的其他众包领域提供参考。

结论

本文提出了一套大规模角色对话数据集的构建方案，并介绍了构建过程、众包工作者访谈以及用于验证的实验。我们围绕研究发现提出了三个研究问题，证实了我们的方法能够回应有关参与者困难、主持人角色以及产出成果的认可与评价等尚未解决的问题。我们希望本研究能够为促进成功且以工作人员为中心的语料库构建研究迈出一步，为所有参与项目的利益相关者带来满意的体验。所收集对话的公开版本可在 GitHub 仓库[11]，获取，相关材料也将为未来研究持续维护和改进。

参考文献

Barrett-Lennard, G. T. 1981年。《共情循环：对一个核心概念的完善》。《咨询心理学杂志》，28(2)：91–100。

Barrett-Lennard, G. T. 2015年。 关系量表：完整的资源与指南。 JohnWiley&Sons。

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.;Neelakantan, A.; Shyam,P.; Sastry, G.; Askell, A.; 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。

Byon, A. S. 2006. 语言间接性和敬语在韩语请求中实现语言礼貌所发挥的作用。2(2):247–276。

[11]https://github.com/smilegate-ai/OPELA

Cassell, J.; Gill, A.; and Tepper, P. 2007. 对话中的协调与融洽关系。载于具身语言处理研讨会论文集，41–50。

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017年。自动仇恨言论检测与冒犯性语言问题。载于国际*AAAI*网络与社会媒体会议论文集，第11卷，512–515。

Dinan, E.;Roller, S.; Shuster,K.; Fan, A.;Auli, M.; and Weston, J. 2019年。维基百科向导：知识驱动型对话代理。载于国际学习表征会议。

Godfrey, J. J.; Holliman, E. C.; and McDaniel, J. 1992年。SWITCHBOARD：用于研究与开发的电话语音语料库。载于声学、语音与信号处理国际会议，第1卷，517–520。IEEE计算机学会。

Grice, H. P. 1975年。逻辑与对话，41 – 58。荷兰莱顿：Brill。ISBN 9789004368811。Kim, B.; Kim, H.; Lee, S.-W.; Lee, G.; Kwak, D.; Hyeon, J. D.; Park, S.; Kim, S.; Kim, S.; Seo, D.; 等。2021年。大规模语言模型能带来哪些变化？关于HyperCLOVA的深入研究：数十亿规模的韩语生成式预训练变换器。载于*2021*年自然语言处理实证方法会议论文集，3405–3424。

Kudo，T. 2006年。Mecab：一种词性与形态分析器。*http://mecab. sourceforge. jp*。

Li，Y.；Su，H.；Shen，X.；Li，W.；Cao，Z.；以及 Niu，S. 2017年。DailyDialog：一个人工标注的多轮对话数据集。载于第八届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），986–995。

McTear，M. 2018年。聊天机器人的对话建模：当前方法与未来方向。语言传播研究文本：电子语音信号处理，175–185。

Mikolov，T.；Sutskever，I.；Chen，K.；Corrado，G.；以及 Dean，J. 2013年。词语和短语的分布式表示及其组合性。载于第*26*届神经信息处理系统国际会议论文集（第*2*卷），3111–3119。

Moon, J.; Cho, W. I.; and Lee, J. 2020年. BEEP! 用于有害言论检测的韩语在线新闻评论语料库。发表于第八届社交媒体自然语言处理国际研讨会论文集，25–31。

Rashkin, H.; Smith, E.M.; Li, M.; and Boureau, Y.-L. 2019年. 迈向共情开放域对话模型：一个新的基准和数据集。发表于计算语言学协会第*57*届年会论文集，5370–5381。

Rehurek, R.; and Sojka,P. 2011年. Gensim——用于向量空间建模的 Python 框架。马萨里克大学信息学院自然语言处理中心，布尔诺，捷克共和国，3(2)：2。 Roller, S.; Boureau, Y.-L.; Weston, J.; Bordes, A.; Di- nan, E.; Fan, A.; Gunning, D.; Ju, D.; Li, M.; Poff, S.; 等。2020年。开放域对话代理：当前进展、未解决的问题与未来方向。*arXiv* 预印本 *arXiv:2006.12442*。

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; 等。2021年。构建开放域聊天机器人的方法。在计算语言学协会欧洲分会第*16*届会议论文集：主会场，300–325。Schuetzler, R. M.; Giboney, J. S.; Grimes, G. M.; 和 Nunamaker Jr, J. F. 2018年。对话代理具身化和对话相关性对社会期许性作答的影响。《决策支持系统》，114：94–102。Seeger, A.-M.; Pfeiffer, J.; 和 Heinzl, A. 2021年。与类人对话代理进行短信交流：为拟人化而设计。《信息系统协会期刊》，22(4)：8。Shah, P.; Hakkani-T¨ur, D.; T¨ur, G.; Rastogi, A.; Bapna, A.; Nayak, N.; 和 Heck, L. 2018年。通过对话自博弈，一夜之间构建对话代理。*ar Xiv* 预印本 *arXiv:1801.04871*。Shuster, K.; Urbanek, J.; Szlam, A.; 和 Weston, J. 2021年。我是我，还是你？最先进的对话模型无法维持身份。*arXiv* 预印本 *arXiv:2112.05843*。Smith, K. A.; Bishop, F. L.; Dambha-Miller, H.; Ratnapalan, M.; Lyness, E.; Vennik, J.; Hughes, S.; Bostock, J.; Morrison, L.; Mallen, C.; 等。2020年。提升医疗咨询中的共情——一项干预措施的二次分析。《普通内科学杂志》，35(10)：3007–3014。Spearman, C. 1904年。两个事物之间关联的证明与测量。《美国心理学杂志》，15(1)：72–101。Stalnaker, R. 2002年。共同基础。《语言学与哲学》，25(5/6)：701–721。Tannen, D.; 等。2005年。对话风格：分析朋友间的谈话。牛津大学出版社。Van der Maaten, L.; 和 Hinton, G. 2008年。使用 t-SNE 可视化数据。《机器学习研究杂志》，9(11)。Waseem, Z.; 和 Hovy, D. 2016年。仇恨符号还是仇恨之人？推特仇恨言论检测的预测特征。在 *NAACL* 学生研究工作坊论文集，88–93。Welch, B. L. 1947年。当涉及若干不同总体方差时，“学生”问题的推广。《生物统计学》，34(1-2)：28–35。Wen, T.-H.; Vandyke, D.; Mrk˘si´c, N.; Gasic, M.; Barahona, L. M. R.; Su, P.-H.; Ultes, S.; 和 Young, S. 2017年。一种基于网络、端到端可训练的面向任务对话系统。在计算语言学协会欧洲分会第*15*届会议论文集：第*1*卷，长篇论文，438–449。Xu, J.; Szlam, A.; 和 Weston, J. 2021年。超越金鱼记忆：长期开放域对话。*arXiv* 预印本 *arXiv:2107.07567*。

Yang, K.; Jang, W.; 和 Cho, W. I. 2022年。APEACH：通过分析众包生成的仇恨言论评估数据集来攻击贬损性表达。*arXiv* 预印本 *arXiv:2202.12459*。Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; 和 Weston, J. 2018年。个性化对话代理：我有一只狗，你也养宠物吗？在计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），2204–2213。

致谢

本研究得到 Smilegate AI[12] 的支持。我们感谢所有众包工作者和 DeepNatural AI[13] 创建了高质量的数据。

[12]https://smilegate.ai/ [13] https://deepnatural.ai

完整访谈文本

问题集

给主持人的问题

1. 您以前是否有过管理大规模对话语料库构建工作的体验？如果没有，您是否管理过任何自然语言数据集的创建工作？2. 在大规模对话数据集构建过程中，最具挑战性的是什么？又有哪些方面让您觉得付出是值得的？3. 演员面临的主要困难是什么？您是如何应对这些处境的？(a) 与演员的沟通的关键是什么？4. 您是否曾在与演员的沟通中感到受阻？(a) 哪些方面让您感到困难？(b) 这些困难是否与您在访谈角色扮演者时观察到的某些特点相吻合？(c) 在什么条件下，主持人与演员之间的沟通可以得到改善？5. 演员的报酬是如何确定的？他们对奖励是否满意？6. 您如何看待主持人在这一建设项目中的重要性？(a) 您认为主持人的职责中，哪些方面最重要？(b) 项目中是否有哪些方面可以改进？

给演员的问题

1. 你以前有过匿名聊天的体验吗？2. 这种对话与和身边的人进行的日常对话，主要区别是什么？(a) 哪种对话与这个项目最接近？（比如与朋友、同事、父母等的对话）3. 与真实用户的对话和访谈中的练习有区别吗？(a) 与固定且预期的对话对象进行对话，和与不确定且可能出乎意料的人群进行对话，两者之间有区别吗？(b) 你是否有自己的策略来让对话顺畅进行？

4. 你对在对话间插入人工时长有什么看法？

(a) 你通常在什么情况下使用人工时长？(b) 对话对象不同时，使用人工时长的方式会有所不同吗？如果会，有哪些不同？(c) 插入人工时长后，是否还能继续进行自然对话？

5. 项目中还有其他困难吗？或者有没有哪一刻让你想要结束对话？

（a）如果有，原因是什么？（b）在塑造角色设定的过程中，哪个环节最困难？你是否曾努力维护角色设定？如果个人资料中包含大量关于角色的信息，你能通过对话记住这些信息吗？对话过程中，你对自己的角色设定有多清楚？（c）在整个项目过程中，你是否曾想改变对自己的设定？还有没有其他你想参考的演员？

6. 关于项目中具体对话的其他问题

(a) 哪位用户给你留下的印象最深？原因是什么？这段对话有什么特点？(b) 你最想和哪类用户继续对话？有没有谁让你觉得像“朋友”？如果有，为什么？(c) 你最想和哪类用户结束对话？换句话说，在哪些对话中，你觉得和对方疏远或相处尴尬？

7. 这个项目如何改变了你自己？

(a) 角色设定与你现实中的自我有何相似之处？(b) 项目开始前后，你现实中的自我是否发生了变化？

全文

主持人

1. 我没有参与过这种采用角色设定的大规模对话数据集项目。不过，我在收集和管理自然语言数据集方面有不少经验。2. 最具挑战性的部分，是激励演员并理解他们的情感，尤其是他们在进行对话时遇到的困难。这个过程类似于精神科咨询。因此，我经常询问参与者是否还好，例如每周开始时。有些演员感到不自在或遇到困难时会向我求助，于是我认为其他演员可能也有类似想法，便征求他们的意见。当演员在我的支持和咨询下受到鼓励，并决定继续与用户参与者进行对话时，我觉得最有价值。3. 演员告诉我，当用户参与者以令他们感到厌烦的方式与他们交谈时，他们通常会很挣扎，却又不能表现出任何不快（大多是因为他们的人设）。由于这类难以应对的用户屡见不鲜，随着时间推移，一些小举动也成了压力诱因。远程交谈情境也让他们更难处理自己的情绪，有时甚至会让参与者感到不适。在这种情况下，我会立即表现出共情态度，或在力所能及的范围内解决问题，有时这也能让参与者感到宽慰。问题解决后，演员在之后的对话中谈到相关话题时会自在得多，我也会把这类特殊情况补充到

指南中（尽管并非所有问题都能百分之百解决）。

(a) 与演员沟通的关键在于信任和融洽关系。访谈演员并预先培训演员指南，有助于建立融洽关系；此外，通过即时通讯软件与他们交流，可以让他们感到舒适并熟悉项目本身，而不是面对带有合同的正式情境。我担任了主要角色，因此持续与他们沟通；这或许使他们对我产生了信任，甚至愿意告诉我他们在执行任务时遇到的所有困难和问题。在其他自然语言数据集任务中，尤其是词性标注等语言学任务中，可能不需要这一过程，因为任务已经预先定义，工作人员大多是该话题的专家，目标也很明确，构建流程能够顺利进行。然而，在对话数据集创建过程中，可能会出现各种特殊冲突，因为该任务至少需要两名参与者，而他们的背景、想法和行为可能有所不同。这说明，让参与者对项目承担责任，并让他们体验与陌生人建立融洽关系的重要性。

4. *(a)* 当我无法辨明演员的实际状态时，会感到无所适从。从这个意义上说，当演员详细告诉我他们的困境，或坦诚回答我的问题时，我觉得他们最配合。相反，如果他们只说自己的状态还好，或不在多个候选项中做出选择，就会让我难以真正了解他们的处境。

(b) 当演员没有坦诚地向我讲述自己的困境时，我会感到难以应对；如果他们选择的角色设定与现实中的自我不一致，这种困难就更加明显。例如，如果他们的角色设定是愉快而活跃的，而原有个性却沉静内敛，沟通就无法按预期进行，从而妨碍演员与主持人之间融洽关系的建立。

(c) 演得出色，并展现出与真实个性截然不同的角色设定，无疑是演员值得具备的一项能力。然而，对主持人而言，与这类演员组织沟通（并在访谈期间预测沟通情况）颇具挑战。因此，如果演员在自我披露或分享意见时能保留其个性，并据此设定角色，将有助于避免主持人因演员表里不一的态度而无法与其建立融洽关系。

5. 演员的报酬与用户参与者的报酬采用了不同的方式。奖励有两种发放方式：按固定期限发放和按对话发放。我们最初选择了前一种策略，目标是让每位演员完成 *600* 段对话，这种方式持续了约一个月。之后，在收集演员的反馈后，报酬方式改为按对话发放。演员表示，后一种方式更好，因为按固定期限发放会让他们因未达到规定数量而感到压力，而按对话发放奖励则让他们能够更灵活地安排时间。

6. *(a)* 我认为，主持人在该项目的总体工作量中约占 *10%* 至 *15%*。研究人员发起项目，工作人员负责资料搜集，而主持人则负责激励工作人员、解决工作人员之间或研究人员与工作人员之间的冲突，并促进整体沟通（就像我们血液中的血小板一样！）。为此，主持人应能够体谅工作人员的不便并给予他们安慰；这与此类数据集构建过程相契合，因为这一过程需要与工作人员建立融洽关系。

(b) 如果有一个论坛，让角色扮演者能够表达自己的情感并分享信息，那就更好了，这样也不必每次都亲自向主持人寻求帮助。由于构建工作是远程进行的，角色扮演者之间几乎没有机会彼此熟悉，因此他们很少向其他角色扮演者公开自己所面临的困难。如果能够分享这些处境，并且角色扮演者之间相互产生共情，那么与仅仅依赖主持人的情感共情相比，他们将获得更多样化的观点和解决方案。

受访者 #1

1. 之前没有匿名聊天的体验。2. 觉得有义务继续对话。感觉更像是在和新人交谈，而不是和朋友或家人交谈。3. 采访者（主持人）让人感觉不错，但那些身份不明的众包人群却不是这样。有些人甚至把交流对象当成人工智能。为了应对尴尬的情况，我根据自己的个人资料（例如看电影的经历）准备了一份常见问题及答案清单。4. 当对话变得无聊，或同一个话题持续太久、难以继续聊下去时，就会插入人工时长。几小时的时长主要是为了避免混淆，并非应对像隔了四五天那样的超长间隔。有些用户觉得受到了冒犯，因为他们认为自己的话被突然打断了。5. 作为一个对自己的生活标准颇为严格的人，和许多有时会变得粗鲁或欠考虑的人交谈颇具挑战。我觉得如果交流对象越界，我会变得稍微好斗一些，尤其是那些不顾我温和体贴的角色设定而对我恶语相向的人。6. 最令我印象深刻的一次是，对方告诉我，他/她当时也在读和我一样的书。我们产生了很强的共情，对话流程也非常顺畅。让我想要停止或结束的对话，是对方说话非常主观，或者只讲自己的个人故事而不听我说。这种情况会消耗我的精力。7. 一开始，我试着贴合自己的角色设定并保持礼貌，例如使用敬语。不过，随着对话继续进行，我逐渐开始配合交流对象，而不再严格遵循最初的设定。在现实生活中，我有时会把一些小事也告诉朋友，因为他们感觉就像这个项目里的用户一样。我平时和身边的人聊得很多。不过，对话伙伴的数量突然激增，确实是一种颇具挑战的体验。

甚至会把一些小事也告诉朋友，因为他们感觉就像这个项目里的用户一样。我平时和身边的人聊得很多。不过，对话伙伴的数量突然激增，确实是一种颇具挑战的体验。

受访者 #2

1. 是的，这和我在常玩的游戏里与用户聊天的方式很相似。

2. 这种对话风格更像是我与同事交流，而不是与亲密朋友交谈。我比起自己的表情，更关注对方的表情，因此会仔细斟酌用词。

3. *(a)* 与项目经理练习时，过程更投入，节奏也更快。我觉得我们双方都积极参与其中。不过，与众包工作者的对话和我原先的预期略有不同。有些对话节奏非常缓慢。有几个人两三天都没有回复。我有时甚至会忘记查看对方的回复，这打断了对话流程。

(b) 起初，我花了一些时间才适应自己创建的角色设定，但大多数众包工作者都问了我很多问题。很多人问了类似的问题，比如“你有什么兴趣？或者你最感兴趣的是什么？”。刚开始时，我只是被动地回答；后来，为了让这个过程更有趣，我会编造一些回答，或者深入探讨某个想法，并围绕它继续聊下去。

4. *(a)* 设置人工时间的好处是，我可以自由安排休息、午餐或下班。有些参与者不想在对话中途结束，因此我有时会利用人工时间，平和地结束与他们的对话。不过，当我必须在对话中插入人工时间以达到最低次数要求时，会有点尴尬。我觉得如果最低次数要求能稍微降低一些就好了。

5. *(a)* 最具挑战性的部分是回答一连串有关我兴趣的问题。因为别人不断问我新的兴趣是什么，所以应答并不容易。要不断想出以前没提过的新兴趣很困难。比如，如果对方问：“你对什么感兴趣？”我会回答：“电子设备。”接着对方可能会问：“我喜欢三星，你呢？”但对方却不顺着这个话题继续，而是不断提出新问题，比如：“哦，是吗？那你还有什么别的兴趣？”之后，当谈到各种兴趣时，我会当场查找相关信息，确保自己说得没错。由于我一边做这些事，一边继续对话，对话经常被打断，所花的时间也稍微延长了。

(b) 我的角色设定与真实自我相似，因此在对话时，我并不难记住这一概念。

(c) 在项目期间，我每时每刻都想改变自己的角色设定。我确实查看过其他人设，但不太记得哪一个最令人印象深刻。不过，如果还有下次，最好建议人设不要把概念列表列得太长。

6. *(a)* 我首先想到两个人：其中一个人让我感觉过于亲近（可能只有我有这种感觉）。不过，那是一次很有趣味性的对话。我不知道为什么人们会觉得与我如此亲近，但我为对话不得不那样结束而感到遗憾。

(b) 第一个是与我有共同兴趣的人。第二个是对我所说的话反应很好的人。第三个是在积极参与对话的人。我只要读他们说的话，就能判断一个人是在装腔作势还是真诚。一个人是否有兴趣提出问题并回答问题，似乎显而易见，即使他们其实并不感兴趣。有些人无论说什么都会改变话题，只是为了消磨时间。

(c) 尽管参与者应该已经意识到我们应当使用人工时间并继续交谈，但有些人忽视了这条规则。他们只是说自己想说的话。例如，有些人在人工时间结束后，突然按照当前时间说话，尴尬地继续对话。有些特定的人一天只说一个词。此外，我不喜欢那些让我无法判断他们是想和我聊天还是想提问的人。无论你说什么，都必须一直谈论 *A*，并继续谈论 *A-1*，但 *B* 总是不断出现；而如果你问 *B*，*C* 又会出现。

另一个挑战是，有些人的对话让我感到沮丧。他是个难以应对的对话伙伴。无论这些人说什么，都会陷入低落的情绪。这个人在整个对话过程中并没有说令人沮丧的话，但整体情感是消极的。我不确定，但我猜这是因为他过去在朋友交往中有过不愉快的体验。我觉得这是因为我的角色设定看起来像个所谓的“圈内人”，像是一个对各种事物都感兴趣的外向者。也许这就是他想让我给他一些建议、帮助他维持良好关系的原因。我积极地给出了一些建议和指导，但无论如何，最终得出的结论都很消极。

7. *(a)* 刚开始对话时，我认为真实的自我与这个角色设定大约有 *40%* 至 *50%* 的相似度。经过几次对话后，相似度达到了近 *60%*。这 *10%* 的差异随着我参与得越来越多，逐渐发生了变化。换句话说，我逐渐调整自己的习惯，以符合我设定的角色。例如，我不常出门，但有一天，对话伙伴问我：“你今天去散步了吗？”我并没有去散步，但如果我说去了，就会觉得自己在欺骗对方，所以我特意出门散了步，也对自己的日常生活做了一些小改变。

（*b*）起初，大量的对话以及警报声不断响起让我倍感负担。由于许多交流对象无视规则，或再次不听通知，因此应对这种情况非常困难。
积极的变化是，随着我们继续交谈，如果对方感兴趣，我并不介意自己的爱好范围被迫扩大。例如，我不是美国电视剧的粉丝。不过，

我从对话伙伴那里得到了很多建议，因此学到了很多，也接触到了许多令人兴奋的事物。

总的来说，这个项目趣味性十足，我希望以后还能再次参与。我觉得这是一次意义重大的体验。我这辈子还有机会和想法、品味各异的人交谈吗？这是一次很棒的体验。这样的机会很难得。我现在正在工作，自从有了这次体验，我和身边的人聊了很多。后来，同事们说他们也想试试，甚至问该如何参与。身边的人也对此很感兴趣。

受访者#3

1. 由于访谈形式会根据受访者略有变化，我们没有向受访者*2*提出问题*1*。2. 由于访谈形式会根据受访者略有变化，我们没有向受访者*2*提出问题*2*。3. 起初，大量用户涌入聊天室，情况十分棘手，我无法按照自己的意愿展开对话。但在与*20*至*30*位用户完成对话后，我开始适应这种状况。我也有一些让对话顺畅进行的策略——引入各种有趣的话题，不敷衍地回应或回答，表现得自然一些，就像用户是我的朋友；不断思考和学习新的话题或议题；与交流对象建立联系，例如给予共情、理解和赞美。4. 由于规则要求至少三次设置人工时长，尽管我并不想这么做，还是经常照做。但有时，在对话中遇到一些情况时，例如节食或去艺术展览等，我也不得不设置人工时长。插入人工时长有时会让我忘记对话原本的脉络，所以这样做可能并没有帮助。如果这项规则不是强制性的，那就好多了。5. 项目开始后，我还得修改人设细节，因此很难假装自己是一位真正的画家。为了解决这个问题，我努力获取一些关于艺术的新信息。然而，用户查看我的个人资料并发起对话时，提出的问题都大同小异。甚至有些不真诚地与我对话的用户，也总是问些毫无意义的问题，这让我很恼火。由于我必须同时回复许多用户，我记不住对话细节，有时也会跟不上对话的脉络。6. 真诚的回应最令我难忘，例如“太棒了！你真是一位画家”“你很有才华！为什么不开设一个*YouTube*频道呢？”或者“很高兴认识你并和你交谈！”。一位交流对象告诉我，和我交谈之后，他/她改变了很多。这给了我继续坚持下去的力量，尽管这个项目多少令人疲惫。那些用毫无意义的句子填满对话的人，例如“你现在在做什么？”“你吃午饭了吗？”，让我感到疲惫，我也不想回答这类问题。7. 在对话中，每当用户向我推荐电影或艺术作品时，我都会去观看或欣赏。这或许是项目结束后我的生活发生的一些变化。这个角色设定与我真实身份的相似度约为*60%*。其余*40%*则是因为从某些角度来看，它反映了我的理想自我形象（也可能是一种愿望或梦想）。

受访者 #4

1. 只有一些随机聊天服务的体验。在一些游戏中聊过天，但没有扮演不同的角色设定2. 这是第一次体验，我投射了一部分自我，以免出错，但有时发现自己说的话并非本意。3. 众包人群的背景多元且不明确，因此他们对我的角色设定有不同的理解。我以为他们会根据我的角色设定采取某种行动，但遇到了许多特殊情况，导致回答起来很棘手。为了应对这些情况，我回想起用户参与者中一些与我相似的反应，并在另一段对话中加以运用。4. 由于我做这份兼职工作的同时还要做其他工作，无法立即回复时便会使用人工时长。一些用户反应得当，另一些则在插入人工时长后难以恰当作答，并忽略了这段人工时长。适应人工时长也颇具挑战（例如，人工时长是三天，但现实世界中只过去了半天）。因此，我会查阅对话历史，以便给出恰当的回答，并弥补已经过去的人工时间。我没有利用对话历史来避开不熟悉的话题。在整个项目中，使用人工时长并不是什么大问题；由于工作时有时无法及时回复，我经常使用以“小时”为单位的时长，这与实际处境相符。5. *(a)* 最具挑战性的是参与者数量相当多。要让对话内容与用户相匹配并不容易，因此回复时有必要查阅对话历史。此外，在调查对话情况时，要回答自己是否感受到联结感或趣味性也有些困难，因为我与新人交谈时的感受并没有太多变化。主观问题很难回答。*(b)* 我设想的角色设定与我的原有个性约有 *5%* 至 *60%* 重合（这让我在与用户交谈时相当谨慎）。我认为这有助于延续对话。重合之处在于不喜欢读书，这一设定旨在引起用户的兴趣。人们经常问我，既然不喜欢书，为什么会成为图书管理员。这个设定成功开启了对话，但每次都用同样的回答让我觉得不自在。*(c)* 因为我的角色设定是“图书管理员”，人们会向我寻求书籍推荐；但我设定自己不喜欢阅读，因此很难在这些角色之间找到平衡。此外，由于我说自己喜欢待在家里，所以不得不解释外出的缘由，例如去见朋友，

我参考了另外 *2* 至 *3* 位角色扮演者，发现“*Jade*”能流畅地引导对话，同时又保留角色设定。我觉得这种角色设定对我来说更容易持续下去。

6. *(a)* 最令我印象深刻的是与“厌倦标注的标注员”的对话。在谈论社会时，对方恰当地运用了常识和知识。我们甚至聊到了数字时代图书馆的创新，以及未来的图书馆会是什么样子。我起初好奇这类对话是否符合项目要求，但很快便对这场对话产生了兴趣，问了许多问题，也从中获得了充实的知识。另一位令我印象深刻的参与者似乎是兼职图书馆工作人员。总的来说，这场对话的内容让我记忆犹新。

   (b) 有时我觉得这场对话很有趣。当他们认真倾听我的意见时，我会有这种感觉；友好的闲聊也很重要。谈话内容主要围绕日常生活。遗憾的是，他们达到目标数量后也结束了对话，尽管我还想和他们多聊一会儿。

   (c) 我记得两位参与者。一位非常悲观，比如说：“假期我会独自一人。我好孤独。”我当时只想尽快结束对话。我不知道该如何应对和关照对方的情感。另一位参与者则不断提问。我一回答，对方就换一个话题，让我觉得自己像是在接受盘问。

7. 我以前常常忘记回复即时通讯软件上的消息，一拖就是几个小时，但参与这个项目后，我努力在现实生活中的对话中立即回复。过去我一直为这种习惯感到困扰，但这个项目似乎改变了我的个性。