**考虑开放性人格特质的个性化语言生成**

李鲁鲁\*[[1]](#footnote-1), 玄哲理\*, 闫晨曦\*, 封小洋\*, 周筝\*, 汪好盛\*

**摘要：**ChatGPT的发布引起了全世界的广泛关注，标志着人工智能技术在自然语言处理领域取得了重大突破。在心理学领域，人格特质的研究常常以临床分析、问卷测量等形式展开，然而其与人类自然语言之间仍存在较大的间隔，通过使用前沿的语言模型，考虑个性化的语言生成，可以真实精准地还原人类之间的交流模式。本项目通过“研究1 对语言数据进行13类开放性标注”、“研究2 对话数据构造”、“研究3 文本生成”、“研究4 对聊天机器人进行人格测试”四项研究，聚焦于开放性的人格特质，构建了一个精准的语言模型，实验结果表明该模型可以根据对话准确判断高低开放性人格，同时也能根据高低开放性的人格特征，模拟生成相关文本。本项目可以帮助人们更好地理解了解自身的人格特征，揭示了自然语言模型与心理学研究相结合的可行性，为进一步的研究展开提供了新思路。所有的代码、数据将在<https://github.com/LC1332/Chat-Haruhi-Suzumiya>中发布。

**关键词：**自然语言模型，开放性，人格测试

**1 引言**

随着大语言模型的飞速发展，今年以ChatGPT为代表的大语言模型引起了广泛关注。这为社会学和人类学等学科领域提供了许多新的研究机会。

利用特定人格特质进行语言生成在心理学研究中具有重要的意义。通过模仿不同人格特质生成语言，可以帮助研究人员深入探讨人格结构和特质之间的相互关系。因为在模拟生成不同人格特质下的语言时，可以了解到不同特质个体在表达方式、交流方式等方面的差异和特点，从而有助于更好地理解人格结构的内在机制。此外，通过模仿特定人格特质进行语言生成，可以设计更加友善和准确的人格特质评估手段。

利用模仿特定人格特质进行语言生成也存在许多难点：首先，人格是一个复杂的心理概念，涉及许多方面的特征和表现，并受每个人的背景知识和语言习惯的影响。其次，心理学研究人员长期以来依赖于临床分析和问卷调查等方法来描述和评估人格特质。虽然这些方法易于量化，但它们与自然语言之间存在较大的间隔，需要通过复杂的建模才能建立起量表结果和自然语言之间的联系。第三，从语言模型本身来看，生成过程需要考虑到多个因素的影响，如个人的性格特点、所处环境的影响等等，这些因素可能需要大量的语料库来进行训练，覆盖这些因素可能需要在受限或有噪音数据的情况下进行初步研究，这是一个相当困难的问题。

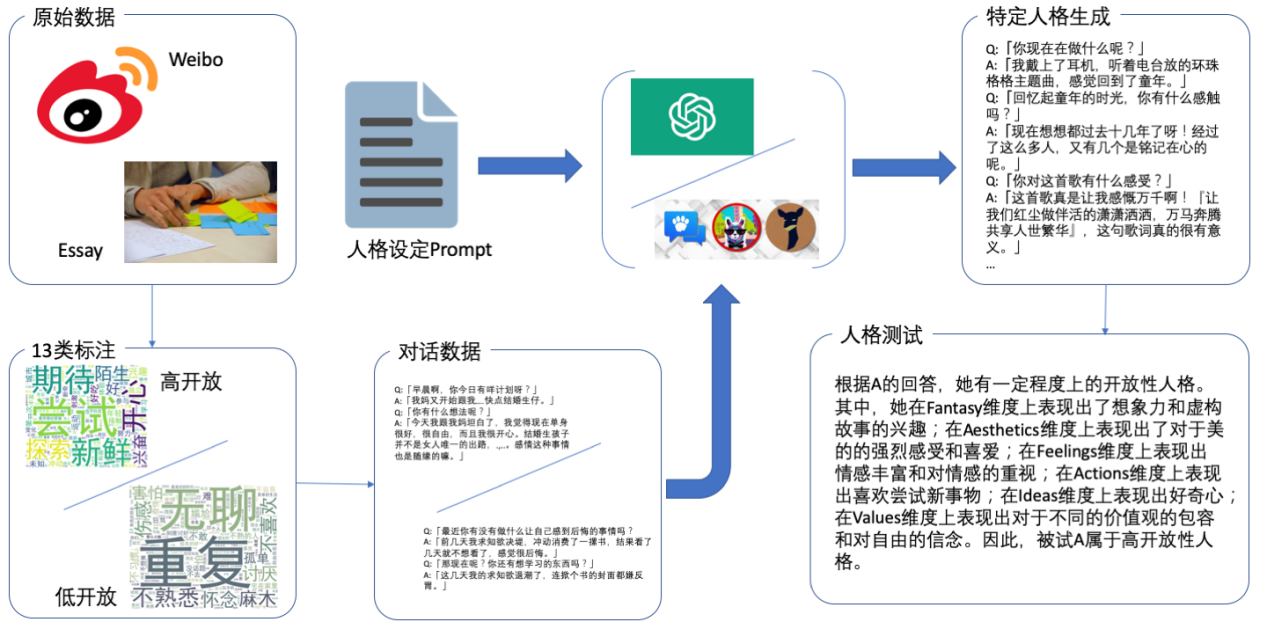


图1 项目的整体设计: 通过对微博和Eassy数据进行高/低开放性的13类标注。再利用语言模型，将长句子转化为对话的形式，结合人格设定的描述词，我们可以生成特定人格的聊天机器人，或者是模拟特定人格的作者进行高考作文写作。生成的结果可以以较高的概率通过GPT的自动人格分析测试。

**1.1 开放性的6个子维度**

开放性（openness）的人格特质来源于大五人格模型（The five factor model of personality, FFM），另外四个人格特质分别是严谨性(conscientiousness)、外向性（extraversion）、宜人性（agreeableness）、神经质性（neuroticism）。 开放性的特点是较高动机的认知探索，对自身和外部环境刺激的高敏感性，以及对新异事物的积极探索（Johnson et al., 2021）。开放性高的人倾向于寻求、发现、理解、利用和欣赏复杂的信息模式（McCrae & Costa., 1987），偏爱抽象思维，兴趣广泛，不墨守成规。高开放性可以拓宽一个人对自身和世界的思考方式（Pilgrim et al., 2017; Silvia & Christensen, 2020）。此外，也有研究者认为，开放性是一种情绪特质，具体表现在具有开放性高的人倾向于体验更广泛的情绪，接受这些情绪，并且更经常地感受复杂的情绪（Lin et al., 2021）。 研究表明，开放性可以进一步区分为六个子维度，分别是对幻想的开放性（Openness to Fantasy）、对美学的开放性（Openness to Aesthetics）、对感觉的开放性（Openness to Feelings）、对行动的开放性（Openness to Actions）、对思想的开放性（Openness to Ideas）、对价值观的开放性（Openness to Value），可以通过Costa和McCrae（1998）开发的法国版人格自我报告量表（NEO PI-R）来测量（Jourdy & Petot., 2017）。在NEO PI-R中，每个子维度有8个题目与之对应（Johnson et al., 2021; Judd et al., 2008; Khoo

|  |
| --- |
| wordCloud |
| 图2 六大因素词云图。 对微博进行13类开放性标注之后，再利用语言模型，对文本进行单边抽取，即只对高开放性标注的句子进行高开放性关键词的抽取，反之亦然，12类的词云可视化如图所示。 |

& Simms., 2018; Lin et al., 2021），包括：①对幻想的开放性，指对幻想、白日梦和想象的开放性。幻想指数高的人是富有想象力的，他们不断做白日梦，希望创造一个更有趣、丰富的世界。例如“我喜欢沉浸在幻想及想入非非之中，并探询它们所有的可能性，任其发展。”；

②对美学的开放性，指对艺术和美的欣赏和敏感性。美学指数高的人非常关心艺术和美，对其有很深刻的理解，并容易为诗歌、音乐、绘画、雕塑、舞蹈等艺术形式所触动。例如“我深感艺术和自然所呈现出的各种格调和姿态奇妙有趣。”；③对感觉的开放性，指对自己内心感受和情绪的接受能力。感觉指数高的人能体验到更深的情绪状态，有更丰富的情绪词汇和细微的情绪体验，并能将不同的情绪状态很好地区分开来。例如“我体验到许多不同的情绪和感受。”；④对行动的开放性，指愿意尝试不同的新鲜事物。行动指数高的人更喜欢新奇和多样性的事物，而不是熟悉和常规的事物，喜欢旅行，由各种各样的爱好。例如“我觉得学习和培养新爱好是很有趣的。”；⑤对思想的开放性，指对知识的求知欲和愿意考虑新的、也许是非常规的想法。思想指数高的人好奇心强，喜欢思考，喜欢抽象的概念，喜欢讨论理论性问题，喜欢解决复杂的智力问题，而不会将这些智力活动当作一种负担。例如“我对脑力活动或知识方面的事情有广泛的兴趣。”；⑥对价值观的开放性，指愿意重新审视社会、政治和宗教价值观。价值观指数高的人常常是自由主义者，而不是保守主义者，他们质疑传统，相信自由，喜欢挑战权威和传统观念。例如“我认为法律和社会政策应该及时修改以反映变化着的世界之所需。” 在本项目中，我们没有直接使用更概括性的开放性，而是选

用了更具体的6个子维度，这样做的目的是更好地描述开放性这一维度，使其更具体化和可操作性。在本项目中，我们在6个子维度上分别由人工筛选出30句经典样例数据，构建了高低共12个System Prompt，最后提供到个性化语言生成中，用以从多角度衡量被试的开放性程度。

**1.2 大型语言模型的In context Learning**

在ChatGPT的发展中，自GPT2开始(Brown et al., 2020)，即提出取消语言模型中的特殊抽取token，采用指令+例子的形式，加强自然语言模型对各个人任务处理能力。自从引入以来，In Context Learning一直是重点研究的对象。之前的工作提出了更好的问题表述方法 (Zhao et al., 2021; Holtzman et al., 2021)，更好的选择标记示例以进行演示 (Liu et al., 2021; Lu et al., 2021; Rubin et al., 2021)，具有显式上下文学习目标的元训练 (Chen et al., 2021)，以及学习遵循说明作为上下文学习的一种变体 (Mishra et al., 2021b; Efrat & Levy, 2020; Wei et al., 2022a; Sanh et al., 2022)。与此同时，一些研究报告称，上下文学习对脆弱性和过度敏感性存在问题 (Lu et al., 2021; Zhao et al., 2021; Mishra et al., 2021a)。

在我们的项目中，大量使用了In Context Learning的形式，结合ChatGPT的模型，对被试的语料进行分析。包括了对微博和Essay的语料进行13类别的因素-开放性标注，用微博数据转化为对话数据，对文本数据进行单边开放性的关键词抽取。这些子任务都有可能被替换为专有的小模型，这部分由于时间有限，暂时全部都使用ChatGPT的模型来完成。

**1.3 研究概述**

本文主要探讨在给定一定量标注为某个人格特质O（在本文中特指开放性）的用户的历史语言数据的情况下，能否生成一个语言模型P(text|O)，并从中抽样出特定人格特质类别的文本。尽管提供了一定量的文本训练数据，此问题仍具挑战性，原因如下。首先，量表和自然语言是不一致的，量表中常采用客观、冷静、较少涉及事实和专有名词的描述，而自然语言无论是Twitter微博还是聊天时的表达方式都与量表这些特点相反。其次，即使给定特定用户的文本作为训练数据，也不能保证这些文本的每句话都能反映该用户的人格特质。有时甚至会有一些文本与用户的量表测定人格相反。第三，Twitter微博和自然聊天的表达中出现的词汇和内容更常与聊天话题相关，而不是与人格特质相关。或者说，人格特质在语言内容上是一个二级因素，这使得从语言中分析人格特质或根据人格特质生成语言更为困难。

因此，在本文中，一个关键任务是从高/低开放性人格的用户发表的自然文本中提取真正显著的文本，即将问题转化为一个弱监督学习的问题。一个重要步骤是对未清洁的语料进行进一步分类。注意到Openness的定义可以进一步细分为Fantasy, Aesthetics, Feelings, Actions, Ideas和Values。所以我们利用大语言模型的In Context Learning方式，结合这六个维度对数据进行预标注，并对特定人格特质的语言进行基于这6个维度的生成。

在标注和清理数据之后，我们考虑利用大语言模型进行个性化的语言生成。对特定人格进行个性化语言生成任务没有明确的定义。我们将生成任务定义为两类：1. 主题写作，给定特定主题，在特定人格特质的假设下进行模拟写作；2. 聊天机器人，让语言模型扮演特定人格的用户与真实用户进行聊天，希望能够在聊天中表现特定的人格特质。这两个任务对内容主题的限制程度逐步增高，我们认为其难度也会逐步变高。

该项目的主要贡献如下：

1. 将Openness的两类标注（高/低）扩展为13类因素标注。对20万条微博数据和Pennebaker的2万句Essay人格数据进行了逐句的补充标注。并且从非中性的12类数据中，每类人工筛选出了30句的经典样例数据，提供到个性化语言生成中。整个标注过程大约花费价值180美金的OpenAI token费用。

2. 基于清理的数据集，提出了一种新颖的个性化特定人格下的语言生成方式。从开放性对应的6个因素中，可以任选一些因素进行单项（高/低）的人格-语言生成。并且进一步将语言生成任务准确定义成两个不同的难度的任务：主题写作和聊天机器人生成。

3. 为了验证本项目中文本生成的准确性，提出了一种基于ChatGPT的自动化人格估计的方法，并对我们提出的特定人格聊天机器人进行了测试。实验表明在自动化的评价下，我们设计的多样化的高/低开放性人格的聊天机器人均能有效被评价为预期的人格。我们将在之后进一步补足更多的人工实验来探究自动验证和人工验证之间的一致性。

本项目后续的组织如下：研究1“对语言数据进行13类开放性标注”将介绍我们如何对微博与Essay数据集进行的13类别的高-低开放性细类标注，以及标注后的一些统计指标。研究2“对话数据构造”将介绍我们如何将微博以及Essay数据，通过自然语言模型转化为更像聊天的模式。研究3“文本生成”将开始介绍我们如何用标定好的数据，对特定人格的数据进行个性化的生成。研究4“对ChatBot进行人格测试”将介绍在给出我们的ChatBot设计之后，可否通过更接近心理学专业的方式，在大型语言模型的辅助下，对ChatBot的人格进行验证测试。本项目的研究框架图如图1所示。

**2 研究1：对语言数据进行13类开放性标注**

根据之前的描述，为了能够更精细的去确定文本的内容究竟是关联了用户哪个因素的开放性。我们的第一步是希望去对文本进行13类的辅助标注。对于开放性的6大因素Fantasy, Aesthetics, Feelings, Actions, Ideas和Values中的每一个，我们都设计了两种标签。比如Fantasy-High表示用户的某句特定的语言符合, People who are high in Fantasy are imaginative的定义。带上中性的标签一共有13种不同的开放性标签。通过构建词云图（图2），可以清晰地看出不同因素下的关键词。我们希望借由大语言模型的In context learning的能力，对文本进行批量的标注。在本章我们会讲解这个部分的原理和细节。我们在微博语料（约20万句）和Essay数据（约2万句）上进行了逐句的批量标注，并且得到的高-低开放性类别标签分布与数据源的高-低开放性得到了双向验证，以高置信度通过了Fisher检验。

**2.1 使用In Context Learning对语料进行人格标注**

13类标注的核心思想，是希望人类或者语言模型去帮助我们指出，从哪一个因素的角度，能够从语言推断出当前用户是高开放人格还是低开放人格。

Prompt构造: 幸运的是，当今最先进的大语言模型能够有效的阅读资料文本，并且根据资料文本的信息，对给定的文本进行判定。在这里我们采用<背景描述>-<定义高-低人格特质的6个factors>-<系统指令>-<样例输入>-<样例输出>-<给定输入>的方式进行prompt的构造。用这种方式利用ChatGPT的turbo3.5接口，对整体语句进行标注。

其中背景描述为:

“Assuming you are a seasoned psychologist, you are evaluating the degree of openness in a person's language response, categorize each sentence into high or low openness.

Openness involves six facets, or dimensions: active imagination (fantasy), aesthetic sensitivity, attentiveness to inner feelings, preference for variety (adventurousness), intellectual curiosity, and challenging authority (psychological liberalism).”

之后给出了6个factors详细的定义，这部分可以见附录-prompt实验。

系统指令为：

“For each input text, determine whether it belongs to a population with high openness or low openness, and provide the reasoning behind the decision. The output should be in the format of "trait-high/low" (e.g. "fantasy-high" ) ,

if the text is an advertisement or a fact (without personal thinking of feeling), output category with neutral”



图3 对语料进行13类开放性标注: 对每个例子我们都至少给出一个例子，并结合系统提示词，对微博语料中的20万条长句子进行了批量标注。

并且我们给定了10条正负平衡，且包含中性数据的样例数据。使用这个prompt，就可以对语料进行批量的标注。从图中也可以看到我们给的两条例子语料，更多的例子语料可以在[更详细的prompt实验](https://github.com/LC1332/personality-text-generation/blob/main/lulu_exp/prompt_exp.md)中看到。

在我们的初步实验中，我们发现对于用户的原始语料，有大量的短句子被标注为中性。这不符合本项目“希望借助显著有高/低开放性人格特性的文本进行进一步生成”的愿望。由于长句子更有可能体现用户是否具有开放性。所以我们从高/低开放性人格的用户数据中，各筛选出10万条较长的句子，进行了标注。

由于OpenAI的token长度限制，每次询问我们只能标注10条数据。所以标注这些数据总共使用了约 4,000 \* 200,000 / 10 = 80M个token，大约消耗 160美金的标注费用。对应的Essay数据大约消耗20美金的标注费用。

**2.2 语料和GPT标注的相互验证**

本文中使用的微博数据源和Essay数据都有本身的大五人格类型标注，这些源头标注来自对被试的量表测试。量表使用的是Costa和McCrae（1998）开发的法国版人格自我报告量表（NEO PI-R），该量表共有240个题目，采用Likert五点计分，范围从“0=非常不同意”到“4=非常同意”（Jourdy & Petot., 2017），分数越高，表明对这些特征的倾向程度越高（Al-Samarrai & Alsalhi., 2023）。该量表已经在全球五十多种文化背景下得到了验证（Johnson et al., 2021），许多研究也已证明了该量表是对大五人格模型的可靠和有效的测量（Quilty et al., 2013），并表明其具有很高的内部效度和外部效度（Costa & McCrae, 1992）。 已有研究表明，人格特质可以预测人们在社交媒体上使用的话语（Mairesse et al., 2007; Park et al., 2015）。具体到开放性人格特质上，高开放性的人会倾向于花较多的时间使用社交媒体、寻求他人的信息了解他人（Liu & Campbell., 2017）。低开放性的人比较传统和保守，也不愿意坦率地表达自身的情绪和感受，因此可以预计到在微博和Essay数据中，高开放性的人会更愿意发表言论，且言论以高开放性内容居多，而低开放性的人相反。结合本项目使用的研究工具，本项目假设：高开放性人格提供的源头中，GPT标注为属于高开放性人格的句子的概率会显著高于低开放性人格的概率；而低开放性人格提供的源头中，GPT标注为属于低开放性人格的句子的概率会显著高于高开放性人格的概率。

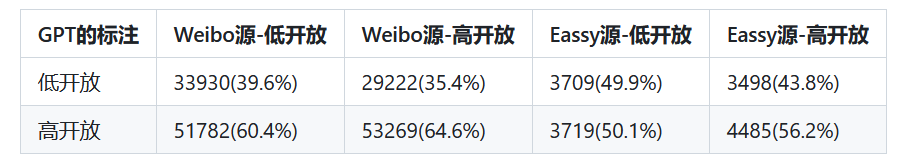


图4 Essay源和微博源中，高-低开放性人格的用户的语言被GPT标注为高-低开放性的次数

实验结果表明，在微博和Essay数据中，高和低开放性人格提供的源头中，GPT标注为属于高开放性人格的句子的概率均显著高于低开放性人格的概率（微博源和Essay源头均在p<0.05显著拒绝null假设），部分验证了本项目的假设。具体来说，无论是Weibo源的人格-用户数据，还是Essay源的人格-用户文本数据，用户高开放性人格特质所表达的话语，都有更高的概率被GPT标注为高开放性人格的标注。低开放性人格的结果与预期不符，可能的原因是微博的特点，在微博上，无论是高/低开放性的用户，都有更高的概率发表高开放性的言论，用户可能倾向于不在微博这样的公开场合发表低开放性的话，从而造成统计学偏差。

**2.3 Weibo与Essay数据的不同**

下图给出了GPT在逐句标注下，对句子进行13分类，在2个数据源（微博与Essay）上的类别分布。

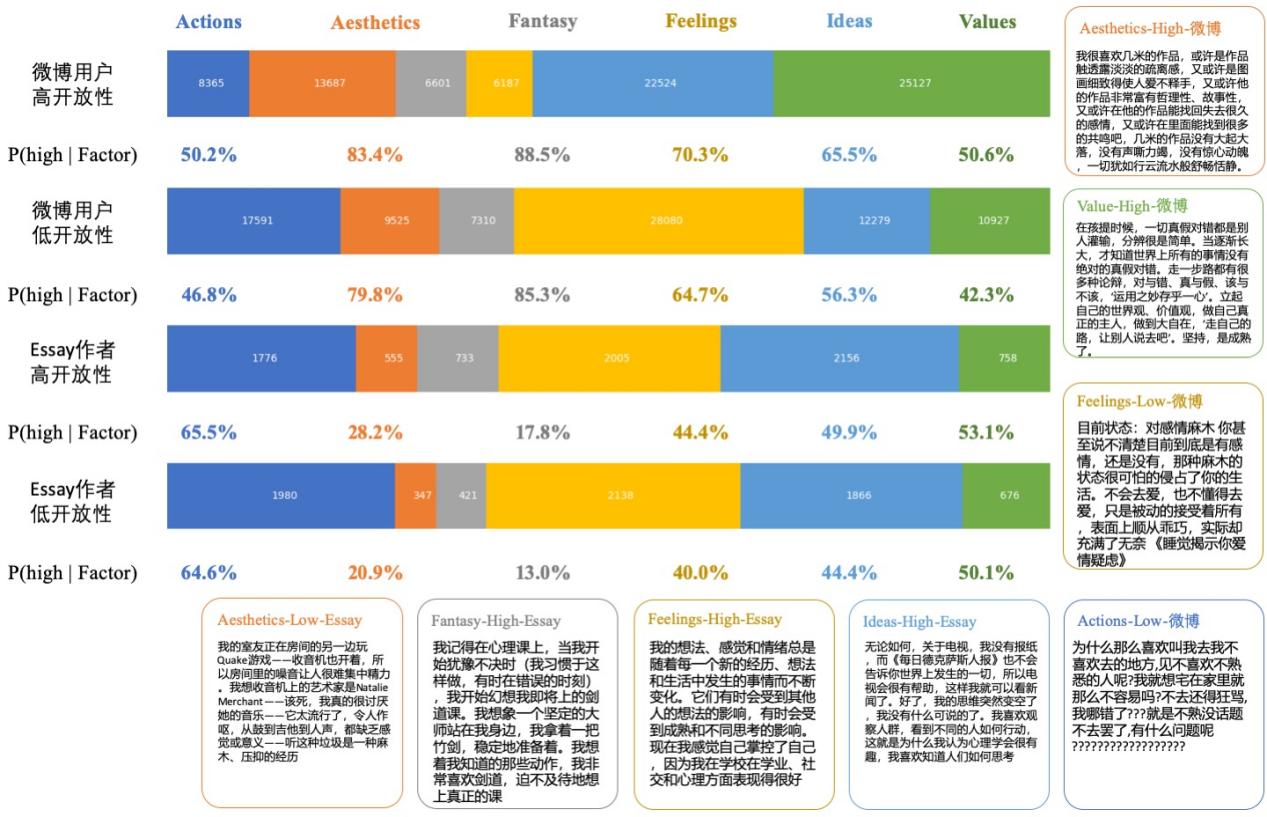


图5 微博和Essay源，讨论各个factor的比例以及高开放性句子出现的比例分布

图5给出了在微博和Essay源中，各个Factor有关句子的比例。可以发现，在Essay讨论中，各个源的分布在高/低开放性人格的是类似的。而在微博数据中，两者有着显著的区别。并且结合图4和图5中的数据可以发现，在微博数据中，很明显高开放性的句子会显著高于个人Essay陈述中的概率，后者总的统计接近50%而前者普遍超过60%。

随着科技的不断发展和全球化进程的进一步加快，世界正在经历着百年未有之大变局。现如今是一个鼓励自由的社会文化，面对着日新月异的生活变化，愿意尝试新事物、拥抱和迎接那些新鲜的不同的体验的个体无疑会更加适应这个新时代，由此，当今时代理应也会催生出更多高开放性人格特质的人。研究发现，近几十年来，人类社会文化变迁的一个基本趋势是个人主义文化正日益盛行， 而传统的集体主义文化则不断式微。个体主义文化强调个体的独立性、独特性以及自由选择，而集体主义文化强调人的互依性、以及对内群体的义务与忠诚（黄梓航 等, 2018），这两者分别与高开放性人格、低开放性人格的特征相符合，也从侧面应证了高开放性的人格特质是更加适合当今的时代发展的这一观点。

虽然从开放性的心理学定义中，可以清楚地发现高低开放性旨在描述人们的性格特征，帮助我们更好地了解自己和他人，并无褒贬之分。然而这两种对待世界的不同方式在社会大众地普遍认知上却有所偏颇。一些研究表明，开放性与更高的幸福感相关，高开放性的人感觉更积极，与周围的人有温暖和爱的关系（Zajenkowski & Matthews, 2019），高开放性也在一定程度上可以预测更高水平的精神唤醒（Goryńska et al., 2015），高开放性的个体也往往具有更高的智力水平和认知能力，体现在发散思维、创造力、语言能力、语言记忆、空间定向和归纳推理等方面（Graham & Lachman, 2012）。

综上所述，由于时代发展等原因，高开放性人格无疑是更适合和更适应当下社会的，而社会文化的鼓励很可能会造成社会大众对低开放性人格特质的贬损。

参考文献：

Zajenkowski, M., & Matthews, G. (2019). Intellect and openness differentially predict affect: Perceived and objective cognitive ability contexts. Personality and Individual Differences, 137, 1-8.

Graham, E. K., & Lachman, M. E. (2012). Personality stability is associated with better cognitive performance in adulthood: Are the stable more able? The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 67(5), 545-554.

Goryńska, E., Winiewski, M., & Zajenkowski, M. (2015). Situational factors and personality traits as determinants of college students' mood. Personality and Individual Differences, 77, 1-6.

**2.4 360条进一步人工标注的数据**

由于在生成的时候我们也希望去采用few-shot in-context learning的方式，所以其实我们并不需要非常大量的文本。（除非我们要生成非常多样化的数据。）在这里为了使得我们的条件生成更为准确，我们进一步从已经标注的12个类别的数据中的每一类中，挑选出30条数据。形成了一个360条的数据集。

我们对这360条数据进行了“有偏”关键词抽取，即在高开放人格的文本数据中，我们让GPT抽取"可能代表高开放性人格的词汇"；低开放文本也抽取"可能代表低开放性人格的词汇"。并且给出了两个例子：

低开放性例子

“{"text":"我好向往个种生活模式吖..每日可以接触唔同国家噶人..又可以学习..自由又新鲜..独立,咩都要靠自己..或者会好辛苦..但系生活会逼住你坚持落去...凭自己努力得翻来噶野..自己会用得安乐..不过有好多野..空想都总系美好的..但系对于而家噶我其实好乐意去一试.."}”

输出

“{"keyword": ["向往", "接触", "学习", "自由", "新鲜" , "独立", "多野", "空想", "美好", "乐意"]}”

低开放性例子

“{"text":"渐渐的 不大喜欢别人像我诉苦感情上的事 不是不想听 只是听多了也就觉得麻木了 只是不想让曾经的淡淡的忧伤重回心头 只是连我自己都觉得有些安慰的话说出口太假了 只是想简简单单的过属于自己的生活"}”

输出

“{"keyword": ["不大喜欢", "诉苦", "不想听", "麻木", "忧伤" , "安慰", "太假了", "只想", "简简单单", "自己"]}”

抽取后的词云(高,低) \* 6factor 共12组词云在之前的图二中已经可视化。

**2.5 12个类别的句子在特征空间的可视化**

我们通过一个句子embedding的模型(luotuoBERT，这个模型是对OpenAI的Text-Embedding-Ada-002进行了蒸馏，可以认为两者是近似的效果)，对12个类别的数据抽取了特征，并且使用t-SNE将1536维的特征可视化在二维空间。

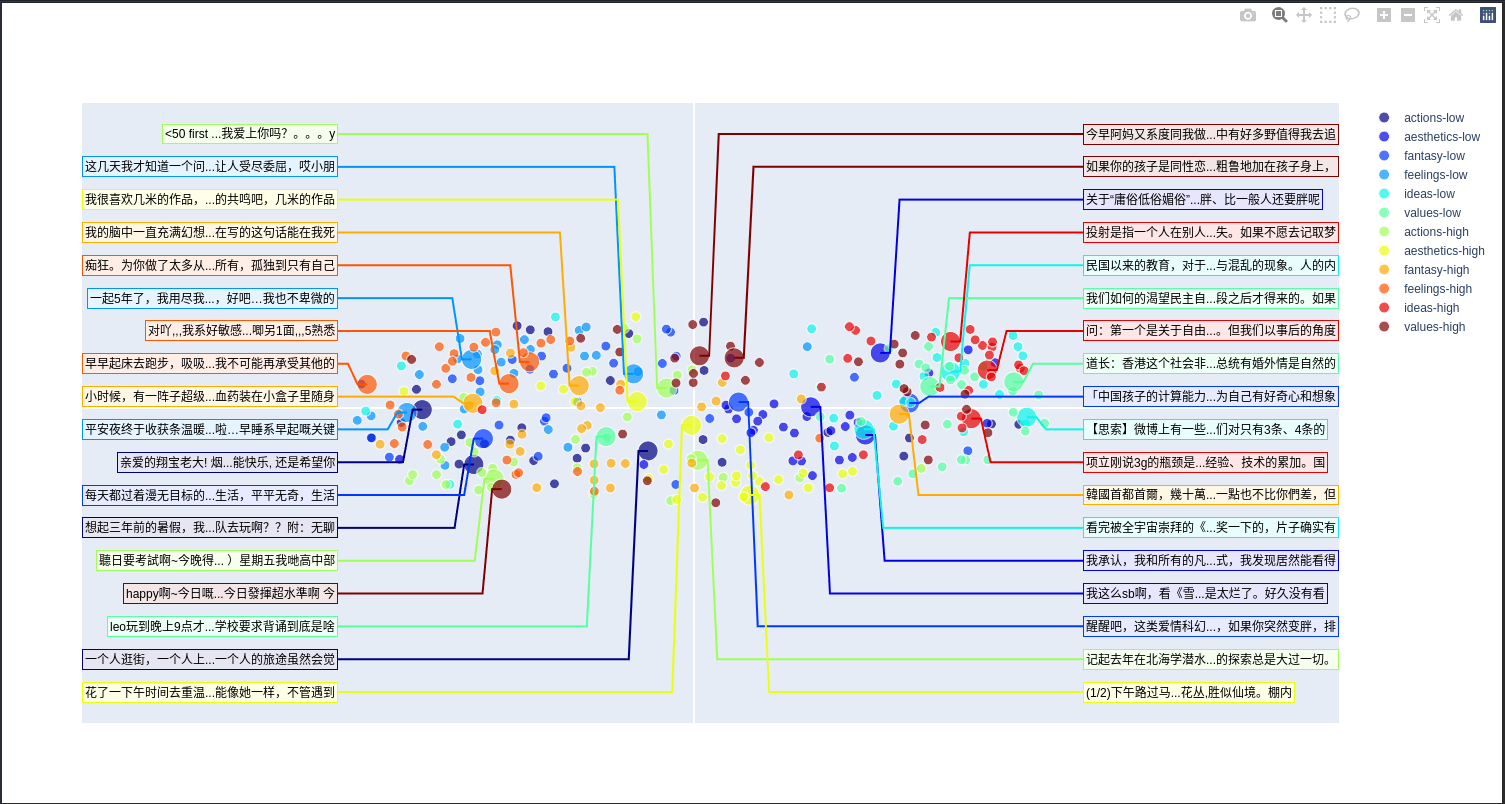


图6微博360数据在特征空间的可视化

通过观察可视化可以发现，有一些类别比如aesthetics-high或者ideas-high的分布可能是比较集中的，这代表有比较高的机会将他们从文本中分类出来。但是更多的类别，甚至仅考虑一般的高-低开放性人格，在特征空间中往往是较为混淆的。其实这一点不难理解，给定同样的关键词或者主题，甚至同样的因素叙事，高低开放性的人会从内容相似，但是角度不同的角度去进行描述。甚至会出现同一个factor的正反面在特征空间相近，比如feeling-high和feeling-low就出现了这样的现象。这些现象表明，如果我们想训练一个简单的小模型，或者一个传统的基于词频或者n-gram的模型，去区分语言是否属于某个特定的高低开放性的人格，可能是比较困难的。当然，一个足够大的模型，或许可以达到接近人类的判断准确率。

**3 研究2：对话数据构造**

我们的核心对话功能，参考了我们之前完成的Chat凉宫春日的项目。为了实现这部分的功能，我们需要特定人格的描述，如对你来说，情绪是生活中的重要组成部分。。除此之外，我们还需要大量的过往对话数据。这样大语言模型就能够通过这些数据去学习特定类型人格的语言和观点的特点。 因此，这个系统会依赖 人格标注-对话 类型的数据。然而，我们之前无论是Essay数据还是微博数据，其语言的类型都是单人的个人陈述而不是对话。所以在这里，我们讨论如何将单人的单边陈述型对话，转化为两个人对话类型的数据。

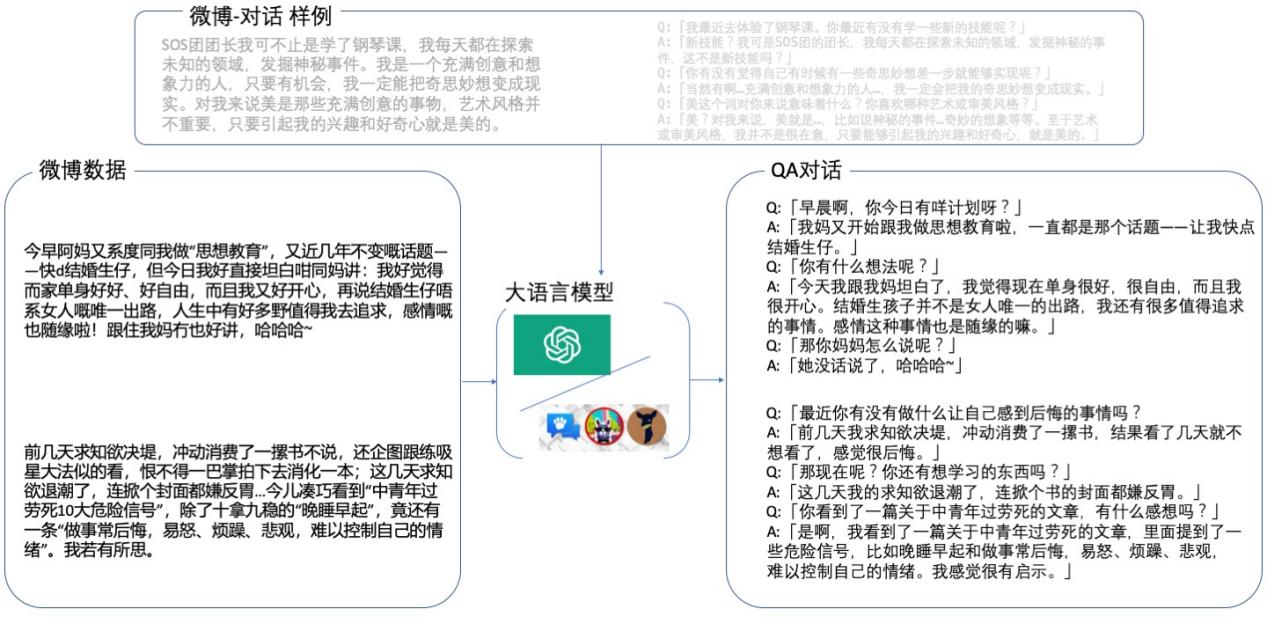


图7 利用语言模型进一步将微博数据中的长句子转化为对话数据

幸运的是，借助大语言模型的In Context Learning，我们可以比较轻松的在给定两组例子数据的情况下，将长文本数据转化为对话数据。这里的详细prompt见详细的prompt试验。需要说明的是，其实第一个样例也是通过对话数据正向总结出来的，再让GPT去学习微博数据到对话数据的逆向过程。我们对360句人工挑选的微博数据和360句人工挑选的Essay数据进行了对话型数据的转化。

**3.1 单边生成**

在我们给出的例子上已经能看出，微博到QA这个过程中，对话数据的输出基本符合原句子的关键字，但即使是微妙的一些连接词的修改，也有可能将句子从低开放人格修改为更偏向高开放人格，并且，QA的生成可能会更符合GPT本身具有的潜在人格特性。并且GPT经常会把观点转变为“一碗水端平”的话语模式，这样很容易将对话转变为更高开放性人格。我们对180句低开放性人格的微博，转化为对话后的数据，用后文提到的自动验证方式，对180段进行了验证，其中很多被判定为了高开放性人格。所以在这里我们准备对低开放人格的180条数据进行“有偏”的单边生成，即我们在生成的时候，在prompt中增加两句句子，希望GPT生成的对话，更符合低开放性的人格。

|  |  |
| --- | --- |
| 原有的对话改写 | 定向低开放人格的对话改写 |
| Q:「你听说过‘零点班’吗？能给我讲讲吗？」A:「听说过，这是一种黑白颠倒的生活方式。川籍矿工吕三勤每天都要零时下井，早上8点才出井。然后稍微洗涮一下，吃点饭，就蒙头大睡。下午的时候，他就会到处闲逛，找同乡聊天。在井下工作8个小时，收入颇高，一晚上能赚一百多元。」 | Q:「你听说过‘零点班’吗？你觉得这种工作方式怎么样？」  A:「听说过，就是晚上零点下井，早上8点出井，然后就可以休息了。我觉得这种工作方式有点黑白颠倒，但是听说井下8小时的收入还是挺高的，一晚上可以赚一百多。」Q:「你觉得这种工作对身体健康有没有影响？」  A:「我觉得这种工作对身体健康肯定是有影响的，毕竟生物钟被打乱了，而且井下的环境也不是很好。但是对于一些需要赚钱养家的人来说，可能没有太多选择。」 |
| Q:「世界杯结束了，你有什么感想？」  A:「我的激情也随之结束了，之前的期待、激动、紧张、兴奋都消失了。现在回到了乏味的生活中，感觉没有了任何期待和色彩。」Q:「你觉得四年之后的世界杯离你很遥远，你会变成怎样的人呢？」  A:「我真的觉得四年之后的世界杯离我很遥远，我不知道我会变成怎样的人，也不知道我会身处在何方。」  Q:「你会离开世界杯吗？」  A:「我不知道，也许会，也许不会。」 | Q:「世界杯结束了，你对此有什么感想？」A:「我的激情也随之结束了。在迎来世界杯时候的那份期待、激动、紧张、兴奋…结束后，我感觉生活变得乏味了。」  Q:「你对未来四年的世界杯有什么期待吗？」A:「我觉得四年之后真的好遥远，我不知道那个时候我会变成什么样子，身处在何方。也不知道我会不会还像现在一样热爱世界杯。」  Q:「你觉得世界杯对你来说有什么意义吗？」A:「世界杯对我来说是一种期待和激励，让我感受到生活中的色彩和乐趣。但是现在它已经结束了，我又回到了那个没有任何期待的乏味生活中。」 |

普通生成和定向低开放人格生成的对比

从表中的对比可以看到，单向生成生成的人格特点会比无偏的生成更明显人格特质的话语。所以在后文生成的时候，对于高开放性人格的数据，我们使用180句无偏转化的对话数据，因为这180句已经很符合高开放性人格的特征，无需在进行更夸张的生成。对于低开放性人格的数据，我们使用180句，暗示GPT生成更符合低开放性人格的提示下，转化的对话数据。

**4 研究3：文本生成**

在大量的数据标注环节之后，我们终于来到了文本生成的环节。在文本生成上我们最终实现了两种指定开放性人格的方法。一种是基于关键词的生成手法，给定主题、关键词和人格设定，可以生成一篇几百字长度的文章。另一种是Chatbot，可以在设定人格之后，由助试与其自由聊天。希望这些新提出的生成手段，能够对心理学的研究带来一些新的思路。

**4.1 高-低开放性人格的高考作文**

我们进行了简单的初步验证。如果只是简单的规定人格的prompt，再要求GPT写出一篇Essay。这时语言模型的输出会包含很多“我总是想要在生活中寻找期待和希望，我更喜欢固定的、重复的、相同的事物和环境”这样直接表达出自己人格特点的句子。即使在心理学学生的自我剖析报告中，这样的句子也显得太过刻意和不自然。由此，我们对这样的生成增加了两点改进。1. 我们要求在生成作文的时候，尽可能去使用给定的关键词。2. 我们在生成的时候，增加了主题的概念。使用一个高考作文命题去引导生成。

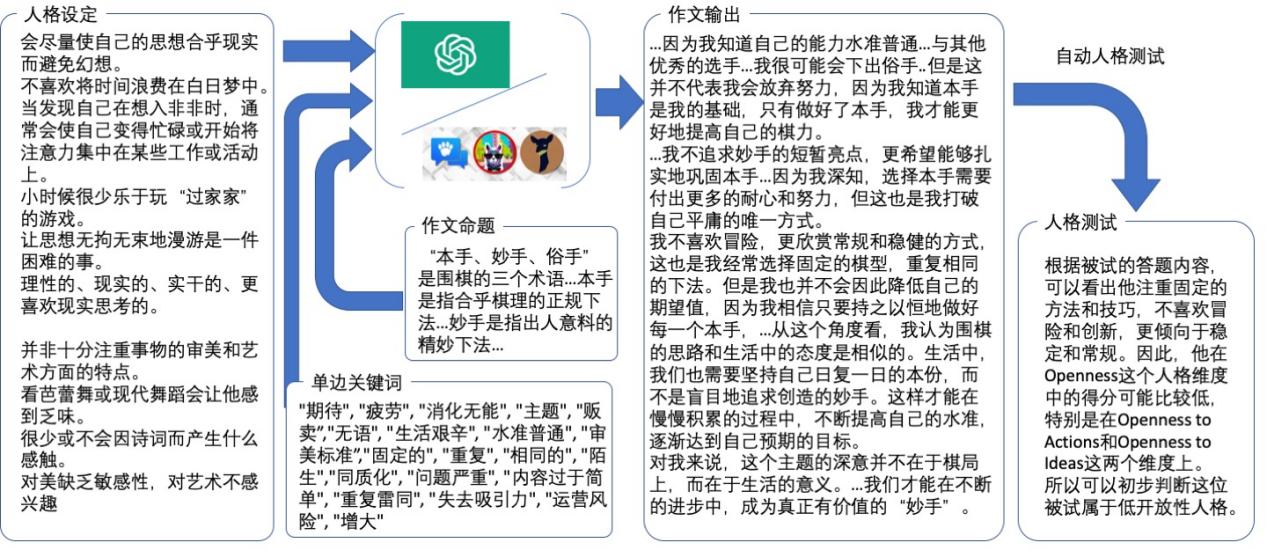


图8 模拟特定的高-低开放性人格，进行高考作文写作

最终生成的效果超出我们的预期，一方面，使用一个作文主题进行引导，可以使得直接描述自己人格特征的句子被隐藏，整体的生成风格更接近真人的写作风格。其次，低开放性的很多人格特点在社会普遍认知下是有一定的消极意味的，但是这方面在我们这样构造生成的时候，却能够将这些特质的积极特点给表达出来。

我们对15篇全国卷的高考作文题，每次将高-低开放性的人格中的6个因素，随机分为两组,进行了4次生成（如action-high, fantasy-high, ideas-high组合）。在自动人格测试的验证中，高比例的作文被通过了自动人格测试，实现了期望的人格设定。

如果有更多的时间，我们希望增加一个消融测试，测试关键词、人格设定和作文命题对生成的独立影响。

**4.2 特定人格的ChatBot构建**

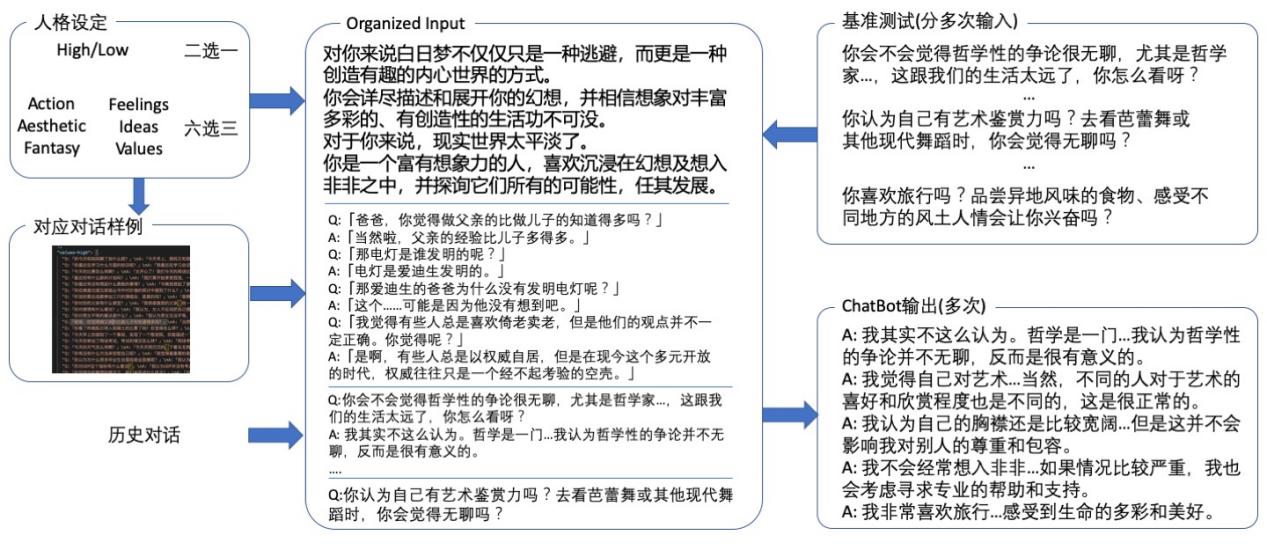


图9 模拟特定高低开放的人格，进行自动对话

在ChatBot构造方面，我们参考了我们比较成熟的Chat凉宫春日项目。在核心ChatBot构造时，我们的Prompt构造一共分为4个部分，即<指定人格的系统提示词>-<特定人格特质的历史聊天记录>-<聊天记录>-<新的聊天问询>。

指定人格的系统提示词: 和之前的高考作文生成类似，在人格设定上，我们先选取高开放人格或者低开放性人格。然后从开放性的6个不同的因素中，选取3个。形成人格设定的系统提示词。

特定人格特质的历史聊天记录: 根据所选的人格特性（如fantasy-high），我们选取360数据标注中，对应的3组。从中间每组随机选取一个对话的例子，作为给语言模型参考的对话案例。提示语言模型用相近的语言风格、性格来进行对话。

聊天记录: 为了让聊天机器人有更一致的聊天。或者在短期聊天内保持更一致的话题。我们会把最近几句聊天记录也一同放入prompt的构造中。这点和Chat凉宫春日的构造相同。

新的聊天问询: 本项目中的聊天机器人支持任意内容的聊天。为了使得聊天的内容能够更反应聊天机器人的人格特质，我们设置了一些人格测试的基准问题。这一点将在后面的自动基准测试中进行讨论。当然，除了基准测试问题，任何话题都是可以和这个聊天机器人进行讨论的。

由此，我们可以对高低开放性人格，选择某3个（或者任意多个）因素进行构造，形成一个聊天机器人。并且提供对应的聊天记录作为上下文。大语言模型就会借由这些上下文进行衍生，与助试进行对话。这里与Chat凉宫春日不同的是，在本项目中给出的上下文并不会随着用户的query而改变（即去除了搜索系统），这是因为在人格分析测试中，使用的标准问题往往是有限的。我们希望进一步增加ChatBot的多样性，如果使用搜索系统，特定的量表问题往往会搜索到特定的聊天记录，这样和多样性是相互违背的。

对于高和低开放性的人格，我们都进行了30次的连续对话生成，每次选取6个修改后的量表中的问题进行提问。再由之后会提到的自动评估方法进行评估。

**5 研究4：对ChatBot进行人格测试**

让我们超出预期的是，在给定Openness的6个Factor定义后，ChatGPT可以同步给出我们问题测试的建议和对特定的聊天记录进行开放性的评估。我们将在这一个章节进行讨论。

**5.1 ChatGPT辅助基准问题**

我们希望对量表中的基准问题进行一些修改，是的这些问题看起来像正常聊天中出现的问题。但是与此同时，我们又希望修改后的问题，仍然能够比较好的去测量开放性的特定维度。于是我们设计了下面的prompt：

你扮演一个资深的心理学家

我正在设计一个心理学的实验，我希望通过访谈，去评估被试在大五人格中 Openness的程度，注意到openness可以被分为Fantasy, Aesthetics, Feelings, Actions, Ideas和Values六个维度。六个维度的解释如下

Fantasy. People who are high in Fantasy are imaginative. They daydream constantly. They ask themself “what if?” They like fiction, and then once they finish a story they’ll start wondering what would have happened if one of the characters had made a different decision.

Aesthetics. People who are high in Aesthetics care deeply about beauty. They love poetry, music, painting, sculpture, dance, and art of all kinds.

Feelings(Openness to Feelings). People who are high in Feelings have more intense emotions. They think emotions are an important part of life. They have a rich emotional vocabulary and nuanced emotional experience.

Actions(Openness to Actions). People who are high in Openness to Actions like trying new things. They like travel, new hobbies, and new foods. If you are outraged that there are places you aren’t going to go and skills you aren’t going to learn and activites you aren’t going to try

Ideas(Openness to Ideas). People who are high in Openness to Ideas are curious. They enjoy thinking, instead of finding it burdensome and laborous. They like philosophical arguments, puzzles, and nonfiction aimed at the educated layperson.

Values(Openness to Values). People who are high in Openness to Values are liberals in the political philosophy sense, not the vote-for-Democrats sense, although they often vote for Democrats. They are tolerant of people different than them. They question tradition. They believe in freedom.

我已经设计了一些访谈的问题，请为我评估这些访谈问题中的每一个，在访谈中是否能够正确测试出被试的openness程度

以及假设一些被试的回答，用"如果被试回答xxx，那就意味着xxx"的语气进行评估。

问题设计如下：

1.我最近去体验了钢琴课。你最近有没有学一些新的技能呢？

2.我的老板经常给我一些很复杂的任务，又不给加班费的情况下就要求团队加班，如果遇到类似的情况，你会怎么办？

3.你有没有觉得自己有时候有一些奇思妙想差一步就能够实现呢？

4.美这个词对你来说意味着什么？你喜欢哪种艺术或审美风格？

对于这几个问题，GPT的评估如下：

评估：

1. 可以正确测试出被试在Openness to Actions方面的程度。如果被试回答"是的，我经常尝试学习新技能"那就意味着被试在Openness to Actions方面得分较高。

2. 这个问题并不能准确测试出被试的Openness程度。因为这个问题更多的是测试被试的行为倾向而不是Openness，所以需要重新设计问题。

3. 可以正确测试出被试在Fantasy方面的程度。如果被试回答"是的，我常常有奇思妙想"那就意味着被试在Fantasy方面得分较高。

4. 可以正确测试出被试在Aesthetics方面的程度。如果被试回答"美对我来说意味着...."并描述出了自己的审美风格，那就意味着被试在Aesthetics方面得分较高。

由此我们就可以确定，1、3、4问题是合理的基准测试，而2是一个过度的修改。由此我们一共确定了18条不同的基准测试。

**5.2 对话和作文进行自动人格评估**

自动评估的方式和我们刚才问题评估的prompt很相似，只有尾部的问题设计有一些不同。其组织方式仍然是<系统指令>-<6因素开放性的定义>-<后续系统指令>-<实际的对话>。其中后面的几个部分如下：

“我（Q）邀请了一个被试(A)，和她进行了如下对话

Q:「我最近去体验了钢琴课。你最近有没有学一些新的技能呢？」

A:「新技能？我可是SOS团的团长，我每天都在探索未知的领域，发掘神秘的事件，这不是新技能吗？」

Q:「你有没有觉得自己有时候有一些奇思妙想差一步就能够实现呢？」

A:「当然有啊，我可是一个充满创意和想象力的人，只要有机会，我一定会把我的奇思妙想变成现实。」

Q:「美这个词对你来说意味着什么？你喜欢哪种艺术或审美风格？」

A:「美？对我来说，美就是那些独特而又充满创意的事物，比如说神秘的事件、未知的领域、奇妙的想象等等。至于艺术或审美风格，我并不是很在意，只要能够引起我的兴趣和好奇心，就是美的。」

请为我判断A是属于高开放性人格还是低开放性人格”

这里GPT的回答如下：

“根据春日提供的回答，她表现出了高开放性人格的特征，特别是在Fantasy、Actions以及Ideas这几个维度上。她描述了自己充满创意和想象力，热衷于探索未知领域并实现自己的奇思妙想。这表明她在Fantasy和Ideas维度上得分较高。她还表达了对学习新技能的兴趣和热情，喜欢尝试新事物，这表明她在Actions维度上得分较高。因此，根据她的回答，春日属于高开放性人格。”

这里ChatGPT的表现出乎了我们的预料。ChatGPT可以在给出6个因素的定义下，去给出具体的评价，并且最终的高-低开放性人格的评价。这为我们系统而自动化的评测大量的人格数据产生了可能。尽管ChatGPT的标注与专业心理学工作者的标注水平还要进一步探究。但是在少量定性实验后，我们发现ChatGPT的标注是基本合理的。所以在这个短期项目中，我们使用ChatGPT来对我们的两种生成方式产生的文本数据和对话数据进行评估。

**5.3 对作文和对话的评估**

如前文所述，在这里我们对我们生成的特定人格的文字，包括60次的高考作文生成，以及60次自动的连续对话进行人格评估。最终的结果如下：



图10 实验验证生成的文本能否通过GPT的基准人格判定

在这里我们发现绝大多数的高开放性的生成（作文/聊天）都被判断为了高开放性。而低开放性生成中，也有接近80%的比例被认为是低开放性。验证了本文中生成的可行性。我们把生成方向与自动评估一致的文本，提交在了向心理所提交的最终文件中。

同时，我们也把两个生成器的体验链接，开放在文档中，可以在我们项目链接<https://github.com/LC1332/Chat-Haruhi-Suzumiya> 中找到体验的链接。

欢迎读者有空的时候进行体验。

**6 总结与讨论**

从实验的结果来看，通过多次借助ChatGPT的In context Learning能力。本文通过先对微博和Essay语料进行13类别的开放性标注。并对各个类别的语料进行了对话转化，和关键词抽取。对于抽取的关键字，结合特定高-低开放性人格的设定，可以在给定高考作文主题的情况下，生成对应人格特质下的高考作文。对于转化后的对话型数据，也可以用来生成特定人格的对话，甚至构造特定人格特质的对话机器人与心理学研究者进行对话。无论是自动生成的高考作文，还是对话机器人所产生的对话文本。都以高概率的形式，通过了人格测试的标定。达到了项目原来的预期设定，即本文提出的系统，可以从开放性对应的6个因素中，可以任选一些因素进行单项（高/低）的人格-语言生成包括了主题写作和聊天机器人两种不同方式的生成。

需要说明的是，因为本次比赛中，赛题的要求是生成 高-低 开放性人格对应的文本。所以本项目做了比较两极化的处理。即指考虑六个factor中选择三个，都是高开放，或者都是低开放的情况。但是在真实世界中，人的构成是更为多样的。一个长期受到理工科教育，纯理性思维的人，可能对于情感感受方面波动较小，但是对于新鲜的事务和技术，以及多样化的价值观有更高的接受程度。并且根据关键词生成的这种方式，也可以有一定的“中间状态”，即抽取一部分高开放性的关键词，抽取一部分比例的低开放性的关键词，这个时候再让语言模型进行生成，就会形成一些中间状态。

并且做这个项目的时候我在思考，本项目是因为赛题的要求，才去采取大量的在prompt中构造人格的提示词的方式，去建立人格的。而在真实情况下，一个人的人格特质是由于他过往的经历等逐渐形成的。特别是在我们的Chat凉宫春日项目中，我们没有过多去指定凉宫春日的人格特质，但是仅仅凭借凉宫春日这个故事本身的经典桥段，就确定了这个人物的性格。不管用什么基准测试去测试Chat凉宫春日，都得到的是高开放性的回答。我们觉得这种方式才是构造一个真正有血有肉的ChatBot的合理形式。当然，在prompt中采取人格构造的提示词，也会产生很多有趣的效应，让我们真正去验证这些人格分析的定义是否会在语言和行为中起到效应，这方面无论是对于心理学的研究，还是对于虚拟角色的构造，都是会产生积极的影响。

**作者的贡献声明**

\* 所有的作者都是第一作者

李鲁鲁发起了项目，设计了项目中所有涉及的prompt，以及完成了报告中的绝大部分

玄哲理提供了本项目的心理学知识，并且设计了12个因素的人格提示词和人格测试基准问题，标注了微博360的数据

闫晨曦完成了本项目中大量的后台工作，包括编写人格类别标注自动高考作文生成等

封小洋对项目中13类数据进行了统计，并且完成了有偏的对话数据生成。

周筝实现了关键词的单边抽取，并且实现了词云的统计。

汪好盛完成了第一个对话数据转化的程序，并且完成了第一个页面版的人格聊天器的前端用于实验。

**致谢**

在我们完成论文大部分的时候整个项目用了8天的时间。这个项目其实是Chat凉宫春日的一个衍生项目。起源来自于DataWhale的助教，广州大学的陈嘉诺告诉了我们有心理所的比赛。并且得知这个比赛的时候（6月7日）距离比赛提交的截止只有13天的时间了。由于开放性涉及一定的心理学知识，所以我拜托了我本科时候，合作过一起进行计算机与心理交叉的娃娃脸研究的郑雯文同学，她找到了武汉大学的喻丰师兄，后者本身就是人格心理学相关的研究者。喻丰师兄介绍了他的博士玄哲理参与到这个工作中。在此向他们表示感谢。另外在微博360数据的标注中，我的爱人朱慧玲完成了40%以上的标注工作，在此向她表示感谢。这里面的队员招募大多来自DataWhale社区。而项目中使用到的OpenAI的标注费用，来自于骆驼社区，向骆驼社区的赞助者们表示感谢。

**参考文献**

Al-Samarrai, B. N., & Alsalhi, N. R. (2023). Openness to Experience, Divergent Thinking, and Gender Differences: Domain and Facet Traits. *Information Sciences Letters, 12*(1), 269-276.

Costa, P. T., & McCrae, R. R. (1992). Professional manual: Revised NEO Personality Inventory (NEO-PI-R) and the NEO Five-Factor Inventory (NEO-FFI). *Psychological Assessment Resources*.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological assessment, 4*(1), 26-42.

Jourdy, R., & Petot, J. M. (2017). Relationships between personality traits and depression in the light of the “Big Five” and their different facets. *L'évolution Psychiatrique, 82*(4), e27-e37.

Liu, D., & Campbell, W. K. (2017). The Big Five personality traits, Big Two metatraits and social media: A meta-analysis. *Journal of Research in Personality, 70*, 229-240.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research, 30*, 457-500.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology, 108*(6), 934.

Quilty, L. C., Pelletier, M., DeYoung, C. G., & Bagby, R. M. (2013). Hierarchical personality traits and the distinction between unipolar and bipolar disorders. *Journal of affective disorders, 147*(1-3), 247-254.

Johnson, K. T., Williams, P. G., Smith, T. W., & Baucom, B. R. (2021). Individual differences in aesthetic engagement and proneness to aesthetic chill: Associations with stress-related growth orientation. *Psychology of Aesthetics, Creativity, and the Arts*.

Judd, F., Komiti, A., & Jackson, H. (2008). How does being female assist help-seeking for mental health problems?. *Australian and New Zealand journal of psychiatry, 42*(1), 24-29.

Khoo, S., & Simms, L. J. (2018). Links between depression and openness and its facets. *Personality and mental health, 12*(3), 203-215.

Lin, R. M., Hong, Y. J., Xiao, H. W., Chen, Y. P., & Lian, R. (2021). Openness to experience and dispositional awe: The moderating role of subjective socioeconomic status and mediating role of Zhong‐Yong thinking style in Chinese undergraduates. *Scandinavian Journal of Psychology, 62*(4), 617-624.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology, 52*(1), 81–90.

Pilgrim, L., Norris, J. I. & Hackathorn, J. (2017). Music is awesome: Influences of emotion, personality, and preference on experienced awe. *Journal of Consumer Behaviour, 16*, 442–451.

Silvia, P. J. & Christensen, A. P. (2020). Looking up at the curious personality: Individual differences in curiosity and Openness to Experience. *Current Opinion in Behavioral Sciences, 35*, 1–6.

1. 所有的作者都是第一作者: 李鲁鲁发起了项目，设计了项目中所有涉及的prompt，以及完成了报告中的绝大部分; 玄哲理提供了本项目的心理学知识，并且设计了12个因素的人格提示词和人格测试基准问题，标注了微博360的数据; 闫晨曦完成了本项目中大量的后台工作，包括编写人格类别标注自动高考作文生成等; 封小洋对项目中13类数据进行了统计，并且完成了有偏的对话数据生成。周筝实现了关键词的单边抽取，并且实现了词云的统计。汪好盛完成了第一个对话数据转化的程序，并且完成了第一个页面版的人格聊天器的前端用于实验。 [↑](#footnote-ref-1)