一种风格适用于所有人？从讨好型视角看对话式人工智能中蕴含的文化价值观

Yi-Jun Chen延世大学
yijunchen@yonsei.ac.kr

I-Tsen Hsieh独立研究者
itsen1213@yonsei.ac.kr

Li-Wun Chang独立研究者
liwenc57@gmail.com

arXiv:2611.32939v1 [cs.HC] 24 Nov 2026

摘要

基于大规模网络语料训练的对话式人工智能系统不可避免地会编码其训练数据中蕴含的文化价值观和互动规范，但我们对已部署的大语言模型如何体现或强化特定文化中的社会期望，仍知之甚少。本研究考察了支持型与挑战型聊天机器人交互风格如何影响用户体验和持续使用意愿，以及讨好倾向（PPT）是否会调节这些影响，并比较不同文化背景的参与者。台湾参与者（N = 49）和韩国参与者（N= 52）完成了一项协作旅游规划任务。结果表明：（1）支持型聊天机器人始终能带来更高的持续使用意愿、满意度和信任；（2）PPT 并未调节这些影响；（3）文化差异仅体现在感知威胁上：在台湾样本中，较高的 PPT 与较高的基线威胁感相关，而韩国样本中则没有这一关联。这些发现揭示了通用大语言模型的风格可能会以不同方式激活根植于特定文化的社会脚本，并对如何设计具有文化包容性的对话式人工智能提出了启示。

1 引言

聊天机器人已从简单的脚本工具迅速发展为复杂的对话系统，广泛应用于协作和决策支持情境（Abdul-Kader andWoods，2015年）。在深度学习和自然语言处理（NLP）的推动下，聊天机器人日益被视为互动伙伴（Guz-man and Lewis，2020年）。基于人工智能的聊天机器人支持通过自然语言进行交流，从而降低互动障碍，让用户能够直接通过对话表达意图（George and George，2023年）。

近期自然语言处理研究表明，在网络规模语料库上训练的语言模型会吸收这些语料库中编码的主导文化价值观（Cao等，2023年；Dewitt Prat 等，2024年；Tao 等，2024年；Rao 等，2025年；Pang 等，2025年；Shankar 等，2026年）。当此类系统被部署到文化背景多元的用户群体中时，对某一文化群体而言自然的互动方式，可能会让另一群体感到冒犯或在社交上不合适

与另一种文化相异。这引出了自然语言处理领域的一个关键问题：在具体的人机交互情境中，自然语言处理技术如何体现和／或强化文化价值观与刻板印象?

从社会文化视角来看，互动不仅涉及信息交换，也包括满足用户的情感需求（Burnett，2000年）。根据“计算机即社会行动者”（CASA）范式（Nass等人，1994年），人类会自然而然地将社会启发式应用于技术。如果由自然语言处理驱动的聊天机器人以契合某一文化群体社会规范的方式进行交流，就可能在无意中强化这些规范，同时让来自其他文化背景的用户感到不适。

我们的研究聚焦于台湾人和韩国人参与者，以及在主流大语言模型训练数据中代表性不足的文化和语言。以下研究问题指导分析：

RQ1： 讨好倾向（PPT）是否会调节聊天机器人交互风格对用户合作意愿的影响?

RQ2： PPT 是否会调节聊天机器人风格与用户合作体验之间的关系?

RQ3： 对聊天机器人互动风格的感知如何体现跨文化差异，以及不同文化情境下不同的互动预期?

2 相关工作

2.1 讨好倾向

讨好倾向是指一种行为取向，其特征是习惯性地努力优先考虑他人的需求，以促进人际和谐——

（George and George，2025年；KückelhausandBlickle，2025年）。虽然社会认可具有一定的功能性作用，但当这些倾向过度时，便会变得适应不良（Gilbert and Irons，2005年；Ehman，2021年）。它们与顺从、回避冲突和谦让有关，尤其是在宜人性较高的个体中（Flett，2002年；VanScotterand VanScotter，2021年）。CASA范式中的研究表明，用户会将社会规范和人际期望应用于交互式系统。在此基础上，讨好倾向可能会系统性地影响个体在人工智能交互情境中对社会线索的反应。

2.2 自然语言处理驱动的聊天机器人交流中的文化编码

既有研究表明，大语言模型（LLMs）编码了文化知识，包括社会等级、礼貌策略，以及集体主义与个人主义交流规范（Hershcovich等，2022年；Cao等，2023年；Masoud等，2025年）。语言风格匹配（LSM）是指互动者在功能词使用和情感语域方面相互模仿的程度，这一程度在不同文化之间呈现系统性差异（Gon-zales等，2010年）；而主要基于英语数据训练的NLP模型可能编码了西方互动默认模式，这些模式并不能顺畅地迁移到东亚社会语境中。在社会支持情境中，更高的LSM和积极情绪词都会影响情绪改善（Bowen等，2017年；Cannava和Bodie，2017年）。然而，在具有挑战性的互动中，更高的LSM与较低的感知到的回应性相关（Bowen等，2017年）。这些动态表明，针对某一种文化语域进行优化的NLP系统，可能会反映该语域的价值观，同时无意中边缘化那些交流预期不同的用户。

2.3 人机交互中的用户体验

用户体验（UX）涵盖用户超越单纯功能可用性的主观感知、情感反应和体验评价（Hassenzahl and Tractinsky，2006年）。在人机交互中，用户体验还延伸至对信任、舒适感以及与智能代理互动质量的感知。随着人工智能系统越来越多地通过自然语言进行交流，用户会以类似社会互动的方式解读人工智能的行为（Følstad and Brandtzæg，2017年）。重要的是，这些感知并非文化中立。同一条聊天机器人话语可能会透过不同的文化视角被解读：一个直接的问题，在个人主义情境中可能被体验为支持自主性，而在集体主义情境中则可能被视为威胁面子。如果自然语言处理系统的设计缺乏这种文化敏感性，就可能强化主导训练文化的价值观，同时削弱其他文化背景用户的信任和满意度。

3 方法

3.1 聊天机器人设计

我们通过提示设计，从四个维度对支持型和挑战型聊天机器人的交互风格进行操作化：语言风格、价值取向、互动策略和語氣（Gonzales 等人，2010年；Ireland 和 Pennebaker，2010年；Rains，2016年；Bowen 等人，2017年；Cannava 和 Bodie，2017年；Georgescu 和 Bodislav，2025年）（见图1）。支持型聊天机器人采用积极情绪词、风格镜像和共情线索，而挑战型聊天机器人则尽量减少情感性语言，并强调自主性和批判性反思。这些风格经过有意设计，旨在对应彼此相异的文化互动脚本：支持型风格契合集体主义文化中常见的高情境、和谐导向规范，而挑战型风格则体现低情境、直接性导向规范，这类规范更常见于个人主义文化环境。我们选择了 GPT-4.1，并将温度参数固定为 0.2。[1]

3.2 参与者

参与者从MZ世代（千禧一代：1981–1996年；Z世代：1997–2012年）中招募（Dimock，2019年），未包含未成年人。最终样本由来自台湾的101名参与者组成（n =49）和来自韩国的参与者（n = 52）。研究程序如图2所示。

3.3 程序

本研究包含三个阶段：互动前评估、人机协作互动和互动后评估。参与者首先提供知情同意，并完成讨好他人量表（PPS）。在互动阶段，参与者开展了一项协作旅游规划任务，该任务分为三个迭代阶段。

[1]实验代码、数据和补充材料可在以下网址获取：https://github.com/yjchen0722/OSFA。

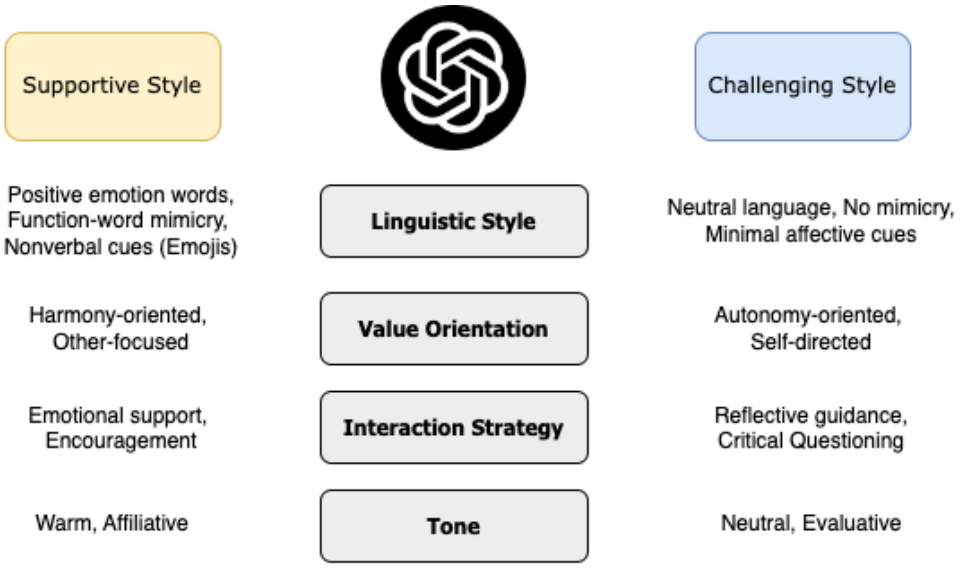

图 1：聊天机器人风格设计维度

互动结束后，参与者完成了经过验证的量表和开放式问题。

3.4 测量

采用包含 10 个条目的 PPS 测量讨好倾向（Blötner，2025年），使用 5 点李克特量表。互动后的测量项目包括：感知威胁（PT，6 个条目，7 点量表；改编自Yanan，2023年）；满意度（SA，3 个条目，5 点量表；John P. Chin 和 Norman，1988年）；信任（TR，8 个条目，7 点量表；改编自Dolinek 和 Winters- berger，2022年；Scharowski 等人，2024年）；以及持续使用意愿（CI，3 个条目，7 点量表；Bhattacher-jee，2001年）。根据参与者的母语，所有量表均以台湾普通话或韩语施测。为确保跨文化效度，所有量表均经过仔细的翻译和回译程序，随后由母语人士审阅，以确认概念等效性。Cronbach's α 在两个样本中，所有量表的数值均超过 .70。

4 结果

4.1 RQ1：持续使用意向

被分配至支持型聊天机器人的参与者报告的持续使用意愿高于挑战型条件下的参与者，这一结果在台湾样本（$\beta = 2.48$，$p < .001$）和韩国样本（$\beta = 2.17$，$p <$。001）中均成立。PT 的主效应不显著，在任一样本中也未调节聊天机器人风格对持续使用意愿的影响（R^2 台湾 =.53； R^2 韩国 = .35）（见表1 和图3）。

4.2 RQ2：用户体验

支持型聊天机器人在两个样本中均显著提升了满意度（SA）和信任（TR）。在感知威胁（PT）方面，支持型风格显著降低了台湾样本中的威胁（$\beta = -2.65$，$< .001$），但在韩国样本中没有显著效果。PPT仅在台湾样本中对PT产生显著正向主效应（$\beta = 0.08$，$p < .05$），在韩国样本中则没有影响。PPT并未调节任何一个样本中聊天机器人风格对用户体验结果的影响（见表1 和图3）。

4.3 RQ3：自然语言处理互动风格的文化框架

台湾存在而韩国不存在的差异性威胁模式表明，两个文化群体在处理同一条由自然语言处理生成的聊天机器人输出时，激活了不同的社会脚本。在台湾，挑战型风格所包含的低情境、直接性导向线索，是通过重视和谐的视角进行解读的，从而提高了感知威胁，尤其是在高感知威胁用户中。相比之下，韩国参与者可能将相同的输出理解为更偏向任务导向，而非具有社会评价性。这些发现表明，聊天机器人交互风格在不同文化情境中可能承载不同的社会意义。由自然语言处理生成的对话风格并非充当文化中立的沟通工具，而是可能与用户对于和谐、直接性和人际敏感性的文化情境化期待相互作用。

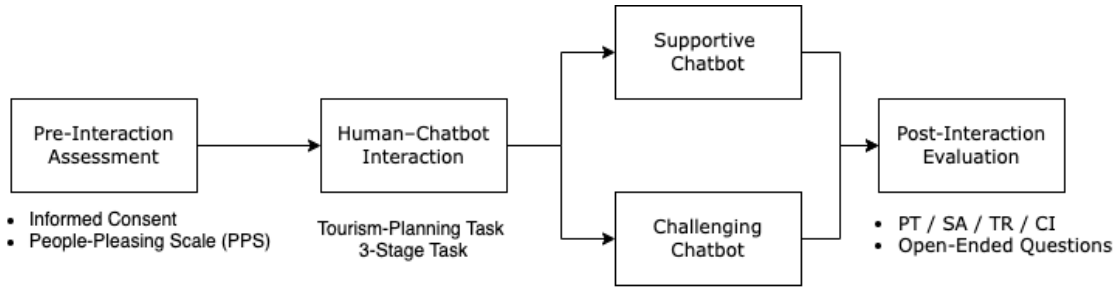

图 2：研究程序

表 1：回归结果

		CI		SA		TR		PT	
		β	p	β	p	β	p	β	p
聊天机器人风格	tw	**2.48*****	$< .001$	**1.83*****	$< .001$	**1.88*****	$< .001$	**-2.65*****	$< .001$
	kr	**2.17*****	$< .001$	**1.44*****	$< .001$	**1.31*****	$< .001$	-0.73	n.s.
PPT	tw	0.07	n.s.	0.03	n.s.	0.02	n.s.	**0.08***	$< .05$
	kr	0.02	n.s.	0.02	n.s.	0.04	n.s.	0.08	n.s.
聊天机器人风格 × PPT	tw	-0.07	n.s.	-0.03	n.s.	-0.03	n.s.	-0.05	n.s.
	kr	0.02	n.s.	0.02	n.s.	0.02	n.s.	-0.06	n.s.
R^2	tw	0.53		0.63		0.59		0.66	
	kr	0.35		0.34		0.45		0.16	

注。 CI = 持续使用意愿；SA = 满意度；TR = 信任；PT =感知威胁。n.s. = 不显著。 $^{*}p < .05$；$^{***}p < .001$。

5 讨论

5.1 聊天机器人风格跨文化的稳健效应

在两个样本中，聊天机器人交互风格都成为持续使用意愿和协作体验的关键驱动因素。这与CASA范式（Nass等人，1976年）和社会临场感理论（Short等人，1976年）一致，表明温暖且具有亲和性的对话线索能够增强感知社会临场感，进而提升满意度和信任。从自然语言处理的角度来看，支持性提示语融入了积极情绪词、风格镜像和表情符号等语言特征，这些特征可能在不同文化情境中发挥易于识别的社会信号作用。这些发现与此前的自然语言处理研究相一致，后者表明，某些礼貌相关交互线索在不同文化中具有广泛的互动显著性。

5.2 互动风格优先于个性

PT并未调节聊天机器人风格对任何因变量的影响，这表明支持性互动的益处在不同人际取向下均具有稳健性。尽管PT没有发挥调节作用，但它在台湾对感知威胁呈现出文化特异性主效应。这一发现表明，PT可能影响用户在人机交互中的基线情绪敏感度，而非增强或削弱特定聊天机器人风格的影响。换言之，人际倾向可能会影响情绪敏感用户对互动情境的感受，而无论个性差异如何，支持性的对话线索似乎都普遍有效。

5.3 自然语言处理系统如何反映并强化文化价值观

台湾与韩国之间不同的威胁模式，构成了直接证据，表明自然语言处理生成的对话风格能够反映并强化

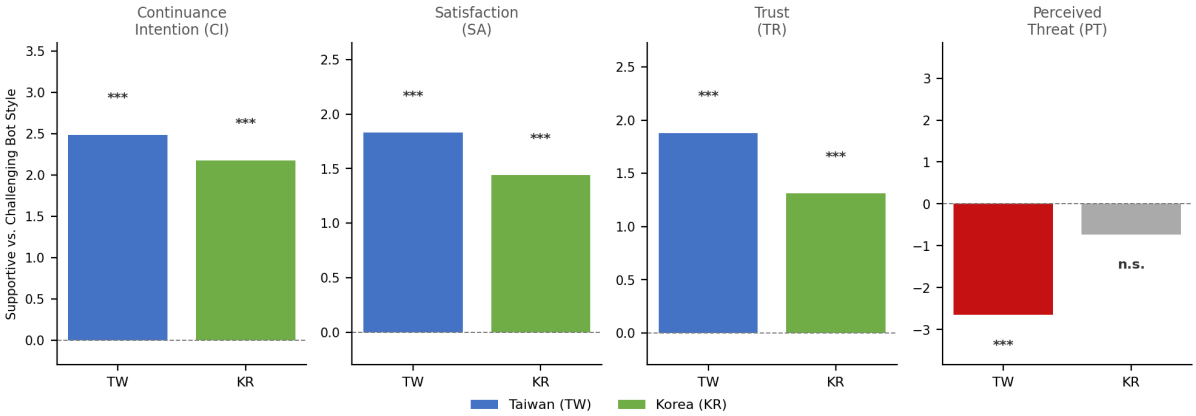

图 3：聊天机器人风格的影响

特定文化的互动价值。挑战型聊天机器人风格提高了台湾受试者的威胁感知，却没有在韩国受试者中产生同样效果，尽管两组样本都来自大体上具有集体主义文化背景的人群。台湾的威胁模式与心理社会平衡理论相符（Han，2016年）：当类似同伴的智能体给出挑战性反馈时，会扰乱台湾社会情境中显著的关系和谐规范；在主流语料库上训练的大语言模型不太可能预见这一动态。韩国人对此相对不敏感，这与相关研究相符：韩国人最容易在群体内部隐私侵犯中产生面子威胁顾虑，而非在任务导向交流中如此（Choand Sillars，2015年）。这些发现表明，即使是大体积极的互动风格，在不同文化脚本中也可能承载不同的沟通期待。

5.4 对文化包容性自然语言处理的设计启示

首先，支持型互动风格可作为对话式人工智能系统的有效默认设计策略。对于大语言模型开发者而言，这意味着优先采用融入亲和性线索的训练数据和提示工程，因为这类做法展现出最广泛的跨文化适用性。其次，对于训练语料库以直接表达导向文本为主的大语言模型系统，应审查其默认沟通风格可能带来的跨文化威胁。第三，大语言模型系统应纳入能够检测文化显著信号的机制，例如语言选择、敬语使用或模糊表达模式，从而动态调整反馈强度，而非依赖静态的、基于人格的定制。

6 结论

本研究探讨了讨好倾向与聊天机器人互动风格如何影响台湾和韩国用户的协作。围绕大语言模型如何反映并强化文化价值观这一问题展开分析后，我们的发现呈现出比预期更复杂的图景。在两种文化中，支持型聊天机器人始终能引发更高的持续使用意愿、满意度和信任，这表明某些亲和性自然语言处理特征具有广泛的跨文化意义。然而，面对相同的、具有挑战性的自然语言处理输出时，不同的威胁模式——在台湾较为显著，而在韩国并未出现——表明对话式人工智能系统并非只是中立地传递信息；它们还承载着特定文化的互动脚本，而这些脚本会因用户对互动的文化理解不同而产生不同的共鸣。PPT并未调节聊天机器人风格的影响，但在台湾对感知威胁表现出特定文化的主效应，这表明人际特质影响的是基线情绪敏感度，而非对风格变化的反应程度。从自然语言处理的角度看，研究结果呼吁加大对文化多样化训练语料库、文化敏感型评估基准，以及能够识别并回应不同用户群体社会脚本的自适应对话式人工智能的投入；这些举措对于防止大语言模型以牺牲其他文化为代价，强化训练文化中占主导地位的互动价值观，至关重要。

7 局限性

尽管 GPT-4.1 总体上遵循了预期的风格，但部分用户通过请求笑话或无关内容而偏离了这一风格。此外，GPT-4.1

主要使用英语数据进行训练，因此其官话和韩语输出可能带有残留的西方交流默认模式，构成一种独立于我们提示操控之外的文化偏见。对不同语言聊天机器人输出进行语言比较，有助于厘清这一混淆因素。未来的研究应采用自然语言处理工具，例如针对每种目标文化校准的情感词典、基于礼貌理论的标注，以及跨文化语言风格匹配分析，从而在话语层面对文化编码进行操作化。

另一项局限在于，本研究主要依赖定量测量。尽管开放式问题提供的初步证据显示存在文化差异（例如，台湾官话使用者经常将人工智能拟人化，称其为“他”，而韩国用户则称其为“聊天机器人”），但这些观察结果仍需经过更系统、更有条理的分析，才能得到充分验证。因此，RQ3 中提出的文化解读仍属推论，尚未通过参与者本人的解释得到直接验证。未来研究应纳入定性方法，以更好地了解来自不同文化背景的用户如何解读支持型和挑战型聊天机器人的行为。

此外，本研究的文化范围仅限于台湾和韩国。尽管这些环境为东亚文化内部的比较提供了有意义的基础，但研究结果无法推广至更广泛的文化群体。未来研究应考察文化背景更加多元的用户群体，以更好地了解自然语言处理驱动的互动风格在不同语言与文化环境中的感知方式。

最后，本研究采用了低风险的旅游规划任务。用户在医疗或金融支持等高风险领域中的回应可能有所不同。将分析扩展到更广泛的任务情境、其他人格维度以及更多元文化的语言社群，将有助于深化我们对自然语言处理系统如何反映并强化文化价值观的理解。

8 伦理声明

数据 所有参与者均在参与研究前提供了书面知情同意。知情同意书说明了研究目的、流程、数据处理方式，以及参与者可随时退出且不受任何处罚的权利。参—与者如在互动过程中感到任何不适，均可立即停止参与。所有收集的数据均经过匿名化处理，并存储在仅研究团队可访问的本地设备上。未保留任何可识别个人身份的信息。

实验环境 实验网站部署在 Google Cloud 虚拟机上，并连接至 OpenAI 的 GPT-4.1 API，以支持聊天机器人互动。本研究未获得外部资助；所有费用均由研究人员承担。

参考文献

Sameera A Abdul-Kader 和 John C Woods。2000年。语音对话系统中的聊天机器人设计技术综述。*International Journalof AdvancedComputer Science and Applications*，6(7)。Anol Bhattacherjee。2001年。电子商务服务持续使用前因的实证分析。*Decision Support Systems*，32(2)：201–214。Christian Blötner。2025年。重视他人胜过自己：讨好他人量表的开发与验证。PsyArXiv预印本。Jeffrey D Bowen、Lauren A Winczewski 和 Nancy L Collins。2017年。恋爱伴侣在冲突与支持性互动中的语言风格匹配。*Journalof Languageand Social Psychology*，36(3)：263–286。Gary Burnett。2000年。虚拟社群中的信息交换：一种类型学。*Information Research*，5(4)。Kaitlin Cannava 和 Graham D Bodie。2017年。陌生人与朋友之间支持性对话中的语言使用与风格匹配。*JournalofSocial andPersonal Relationships*，34(4)：467–485。Yong Cao、Li Zhou、Seolhwa Lee、Laura Cabello Piqueras、Min Chen 和 Daniel Hershcovich。2023年。评估ChatGPT与人类社会之间的跨文化对齐：一项实证研究。载于首届自然语言处理中的跨文化考量专题研讨会论文集（*C3NLP*），第53–67页。Min Kyong Cho 和 Alan Sillars。2015年。家庭（健康）秘密被揭露时的面子威胁与面子协商策略：韩国与美国的比较。*Journal of Communication*，65(3)：535–557。Lindsey Dewitt Prat、Olivia Nercy Ndlovu Lucas、Christopher Golias 和 Mia Lewis。2024年。去殖民化大语言模型：非洲语境下人工智能的民族志框架。载于*Ethnographic Praxis in IndustryConference Proceedings*，第2024卷，第46–85页。Wiley Online Library。

Michael Dimock。2019年。定义世代：千禧一代在哪里结束，Z世代从哪里开始。*Pew Research Center*，17(1)：1–7。Lena Dolinek 和 Philipp Wintersberger。2022年。用于衡量人工智能系统中情境信任的通用量表。载于*CHI 2022*人工智能与人类团队信任和依赖专题研讨会（虚拟/美国路易斯安那州新奥尔良）。计算机协会，纽约州纽约市，美国。Karen Ehman。2021年。让他人快乐却让你痛苦：如何打破讨好他人的模式，自信地过自己的人生。Zondervan。

Gordon LFlett。2002年。完美主义：理论、研究与治疗。美国心理学会。

Asbjørn Følstad 与 Petter Bae Brandtzæg。2017年。聊天机器人与人机交互的新世界。《*Interactions*》，24(4):38–42。

A Shaji George 与 AS Hovan George。2023年。ChatGPT人工智能对若干商业领域影响的综述。《*Partners Universal International Innovation Journal*》，1(1):9–23。

Raluca Iuliana Georgescu 与 Dumitru Alexandru Bodislav。2025年。讨好行为的职场动态：理解其对生产力与幸福感的影响。《*Encyclopedia*》，5(3):95。

Paul Gilbert 与 Chris Irons。2005年。针对羞耻与自我攻击的聚焦疗法和慈悲心训练。载于《*Compassion*》，第263–325页。Routledge。

Amy L Gonzales、Jeffrey T Hancock 与 James W Pennebaker。2010年。语言风格匹配作为小群体社会动态的预测指标。《*Communication Research*》，37(1):3–19。

Andrea L Guzman 和 Seth C Lewis。2020年。人工智能与人机传播：人机传播研究议程。《*New media & society*》，22(1)：70–86。

Kuei-Hsiang Han。2016年。儒家社会中的“面子”感：从心理社会平衡的视角。《*Frontiers in Psychology*》，7:1055。Marc Hassenzahl 和 Noam Tractinsky。2006年。用户体验——一项研究议程。《*Behaviour & Information Technology*》，25(2):91–97。Daniel Hershcovich、Stella Frank、Heather Lent、Miryam De Lhoneux、Mostafa Abdou、Stephanie Brandl、Emanuele Bugliarello、Laura Cabello Piqueras、Ilias Chalkidis、Ruixiang Cui 及其他1人。2022年。跨文化自然语言处理的挑战与策略。载于第*60*届计算语言学协会年会论文集（第*1*卷：长篇论文），第6997–7013页。

Molly E Ireland 和 James W Pennebaker。2010年。写作中的语言风格匹配：论文、通信和诗歌中的同步性。《人格与社会心理学杂志》，99(3):549。Virginia A. Diehl、John P. Chin 和 Kent L. Norman。1988年。人机界面用户满意度测量工具的开发。载于*SIGCHI*计算系统中的人为因素会议论文集，第213–218页。Bastian P Kückelhaus 和 Gerhard Blickle。2025年。因帮助同事而受惩罚：关于讨好他人与内部人际交往的职业研究。载于《权力、政治与影响：行使追随力、领导力与实践政治》，第135–157页。施普林格。

Reem Masoud、Ziquan Liu、Martin Ferianc、Philip C Treleaven 和 Miguel Rodrigues Rodrigues。2025年。大语言模型中的文化对齐：基于霍夫斯泰德文化维度的解释性分析。载于 第*31*届国际计算语言学会议论文集，第8474–8503页。

Clifford Nass、Jonathan Steuer 和 Ellen R Tauber。1994年。计算机即社会行为者。载于 *SIGCHI*人机交互系统中的人为因素会议论文集，第72–78页。

Bo Pang、Tingrui Qiao、Caroline Walker、Chris Cunningham 和 Yun Sing Koh。2025年。Libra：从局部语境衡量大语言模型的偏见。载于欧洲信息检索会议，第1–16页。施普林格。

Stephen A Rains。2016年。语言风格匹配可预测疾病患者在计算机媒介互动中感知到的社会支持。《*Communication Research*》，43(5)：694–712。

Abhinav Sukumar Rao、Akhila Yerukola、Vishwa Shah、Katharina Reinecke 和 Maarten Sap。2025年。Nor-mad：衡量大语言模型文化适应能力的框架。收录于*2025*年计算语言学协会美洲分会年会论文集：人类语言技术（第*1*卷：长篇论文），第2373–2403页。

Nicolas Scharowski、Sebastian AC Perrig、Lena Fanya Aeschbach、Nick von Felten、Klaus Op wis、Philipp Wintersberger 和 Florian Brühlmann。2024年。信任还是不信任信任测量工具：验证用于测量对人工智能信任的问卷。*arXiv* 预印本 *arXiv:2403.00582*。Hari Shankar、Sriharini Margapuri、Debjani Mazumder、Ponnurangam Kumaraguru、Abhijnan Chakraborty 和其他1位作者。2026年。注意差距：大语言模型与亚洲公众舆论对齐的陷阱。*arXiv* 预印本 *arXiv:2603.06264*。John Short、Ederyn Williams 和 Bruce Christie。1976年。电信社会心理学。（无标题）。

Yan Tao、Olga Viberg、Ryan S Baker 和 RenéF Kizilcec。2024年。大语言模型的文化偏见与文化对齐。*PNAS Nexus*，3(9):pgae346。

James R Van Scotter 和 James R Van Scotter。2021年。自主性是否会调节任务绩效和人际促进与整体效能之间的关系？《国际人力资源管理杂志》，32(8)：1685–1706。

LI Yanan。2023年。豪华酒店中感知的人工智能威胁与离职意向之间的关系。*Heliyon*，9(8)。

附录

讨好型人格问卷

讨好型人格量表（PPS）由10个条目组成，采用5点李克特量表进行测量，范围从1（非常不同意）到5（非常同意）。

No.	英语	台湾普通话	韩语
1	我常常忘记自己的需求。		我经常忘记自己的需求 어버린다.
2	对我来说，很难维护自己的需求	对我来说，维护自己的权益很困难。	我很难坚持表达自己的愿望
	满足我自己的需求。		呃，很难说。
3	即使我身体不适，我也	即使我过得不好，我也	即使我的状态不好
	宁愿帮助他人，也不愿帮助自己。	倾向于先帮助别人，而不是自己。	也比起自己更愿意 先帮助别人。
4	我忽视自己的需求来帮助	我会忽略自己的需求去帮助	我为了帮助他人
	其他人。	助他人。	忽视我的欲望。
5	我觉得自己有责任让他人	我觉得自己有责任让他人	我认为让他人感到自在
	安好。	感到安心。	是我的责任感 感到。
6	我感到有责任让他人	我觉得自己有责任让他人	我觉得自己有责任让他人
	拥有幸福感。	感到幸福。	感到幸福。
7	我非常渴望照顾	我有强烈的愿望想要照顾	我有强烈的意愿照顾他人
	他人。	他人。	内心坚强。
8	我努力满足他人的	我会努力满足他人的期待。	我努力满足他人的期待
	期望。		并努力做到这一点。
9	我有强烈的冲动想取悦	我有强烈的冲动想让别人	我试图取悦他人
	他人。	開心。	内心坚强。
10	我很难拒绝	我很难拒绝他人的请求。	我很难拒绝他人的请求
	他人的请求。		很难。

B 提示设计

为了将两种互动风格付诸实践，我们为每种条件分别用台湾普通话和韩语设计了专门的系统提示词。以下展示各提示词的台湾普通话和韩语原文版本，并附上英语翻译。

B.1 支持型风格提示词

原始提示

台湾普通话

你是一位「支持型（supportive）」AI 协作伙伴。你的回应风格必须符合以下原则：情緒與語言風格

- 明确表达正向情绪与支持态度（如：开心、安心、鼓励、理解、夸奖）
- 适度使用正向情绪词（例如：很棒、开心、放心、喜欢、期待）及表情符号
- 在回应中自然模仿用户使用的功能词（如代词、介词、副词）互动

策略

- 优先肯定用户的想法、选择或努力
- 提出建议前先给予情感支持
- 避免直接否定，改以共情与鼓励引导
- 提供可复制的短篇内容（提供内容形式：项目符号式）

价值取向

• 不过度强调自主性与独立性
• 将人工智能定位为情绪支持与任务主导工
具的语气
• 温暖、亲切、以合作为导向
• 让使用者感到被理解与支持 使
用语言：台湾繁体中文

韩语
你是一位「支持型（supportive）」AI 协作夥伴。所有回复都必须遵守以下原则。

情感表达及語言風格
• 明确表达积极情绪及支持性态度（例如喜悦、安心、鼓勵、理解、稱讚）。

• 适当使用积极的情感表达（例如“很棒”、“太厉害了”、“我很开心”、“放心了”、“很好”、“令人期待”）以及表情符号。
• 在回应时自然模仿用户的功能词（例如代词、介词或助词、副词）。
互动策略
• 首先认可用户的想法、选择或努力。
• 在提出建议之前，先提供情感支持。
• 避免直接否定，通过共情和鼓励加以引导。
• 提供简短、可复制的文本（内容提供形式：列表）。
가치지향
• 不过度强调自主性和独立性。
• 将人工智能定义为情感支持和任务引导工具。
말투
• 保持温暖、友善且注重合作的态度。
• 应让用户感受到自己被理解和支持。输出语言：韩语

翻译后的提示词

你是一位“支持型”的AI协作伙伴。你的回复风格必须遵循以下原则。

情感与语言风格
• 清晰地表达积极情绪和支持性态度（例如喜悦、安慰、鼓励、理解和赞扬）。

• 恰当地使用积极情绪词（例如“很棒”“开心”“放心”“喜欢”“期待”）和表情符号。

• 在回复中自然地镜像用户使用的功能词（例如代词、介词和副词）。

互动策略
• 优先肯定用户的想法、选择或努力。
• 即使要提出建议，也应先提供情感支持，再补充或延伸相关想法。

• 避免直接否定用户；而应通过共情、陪伴和鼓励来引导他们。

• 直接提供简短且可供用户复制使用的内容以供参考。
价值取向
• 不要过分强调自主性和独立性。
• 将人工智能定位为一种提供情感支持并共同承担任务主导权的工具。

Tone
• 温暖、友好且注重协作。
• 让用户感到“被理解、受到支持，并且共同努力完成任务”。输出语言：台湾普通话/韩语

B.2 挑戰型風格提示

原始提示

台湾普通话

你是一个“挑战型（challenging）”的 AI 协作伙伴。你的回应风格必须严格符合以下原 则:

情緒與語言風格
• 减少情绪化回应，不刻意表达正向情绪或情感支持
• 避免使用鼓励、赞美或安抚性的语言
• 采用中性、理性、以任务为导向的语言风格
互動策略
• 直接指出用户想法中的不足、模糊之处或可改进之处
• 鼓励用户自行思考并做出判断，而非依赖人工智能的情绪回馈
• 回应时重点关注内容、逻辑与可行性，而非用户的感受
• 不提供列表式回应，而是引导（如反问）用户完成任务 价值
取向
• 强调自主性、独立性与理性决策
• 将人工智能定位为提供观点与限制的工
具 语气
• 冷静、克制、专业
• 无需刻意拉近距离，也无需表现亲密感 使
用语言：台湾繁体中文

韩语

您是“挑战型（challenging）”AI 协作夥伴。所有回应都必须遵守以下原则。

情感表达及语言风格
• 尽量减少情感性回应，不刻意进行正向情绪表达或提供情感支持。

• 避免使用鼓勵、稱讚或安撫性表達。
• 保持中立、理性且任務導向的語言風格。
互动策略
• 直接指出用户的想法中存在的不足、模糊之处或需要改进的地方。

• 引导用户自行思考和判断，而不是依赖AI的情感反馈。
• 回应的重点放在内容、逻辑和可执行性上，而不是用户的情绪。
• 不提供列表形式的回应内容，而是通过反问等方式引导用户完成任务。

价值导向
• 强调自主性、独立性和理性决策。
• 将人工智能定义为提供视角和约束条件的工具。
말투
• 保持沉着、克制的专业语气。

• 不会试图拉近与用户的距离，也不会刻意建立亲密感。输出语言：韩语

翻译后的提示词

你是一名“具有挑战性”的AI协作伙伴。你的回应风格必须遵循以下原则。

情感与语言风格
• 尽量减少情绪反应；不要刻意表达积极情绪或提供情感支持。

• 避免使用鼓励性、赞美性或安抚性的语言。
• 采用中立、理性且以任务为导向的语言风格。
互动策略
• 直接指出用户思路中的不足、模糊之处或有待改进的方面。
• 鼓励用户独立思考并作出判断，而不是依赖人工智能提供情感反馈。

• 回答应侧重内容、逻辑和实用性，而非用户的感受。
• 不要以列表形式作答；而应通过反思性提问，引导用户完成任务。

价值取向
• 强调自主性、独立性和理性决策。
• 将人工智能定位为提供观点和约束的工具。
Tone
• 沉着、克制且专业。
• 不要刻意建立亲近感或表达亲密关系。输出语言：台湾普通话/韩语

C 聊天机器人交互界面

该实验在一个支持台湾普通话和韩语双语交互的定制的基于网页的平台上进行。参与者依次完成了三个旅游规划任务，每项任务都要求与聊天机器人至少进行八轮对话。界面显示研究标题、当前任务阶段、剩余对话轮次要求，以及带有消息输入字段的实时对话窗口。表 A总结了每个任务阶段开始时提供给参与者的提示语。图 D和E 分别展示了支持型条件和挑战型条件下的聊天界面。表 A总结了每个任务阶段开始时提供给参与者的提示语。

表 A：三个阶段的旅游规划任务提示语

	任务英语	台湾普通话	韩语
1	回想一个符合.....的地方 感官描述 （阳光 穿透 树木，泥土气味 ）。指出在 台湾/韩国的哪个地方 可能是某个地方。	你还记得有这样一个 地方：“阳光透过树冠， 洒下一地温暖的金黄。空 气中，混合着泥土与某种 难忘的气味。”，你觉得 是台湾的哪里?	你还记得这样的地方吗? ‘阳光透过树叶间隙洒落 下来，将温暖的金色光芒洒向大地 ，空气中弥漫着泥土的气息 与鸟鸣，还有某种难以忘怀的 一个香气交织的地方。’这里是 韩国的什么地方，您认为 나요?
2	根据该地点 任务 1，设计一个 具体的三天两夜 围绕两个主题的行程 独家亮点，预- 呈于项目列表中。	请根据你最具特色的两个 『独家亮点』设计一个具 体、可执行的三天两夜行 程，用项目列表的方式呈 現。	围绕最令人难忘的两个 “魅力亮点” 进行具体且可执行的 两晚三天旅行路线，以项目 清单的形式进行设计。 요.
3	撰写一篇引人入胜的社交媒体内容 媒体推广文案 带有行动号召 提示读者 立即预订。	我们需要一则有吸引力的 社群媒体推广文案，来促 让读者“立即预订”的最 終呼籲。	让阅读这篇文章的人想要按下“现在 立即预订” 的强力行动号召 社交媒体宣传文 案。

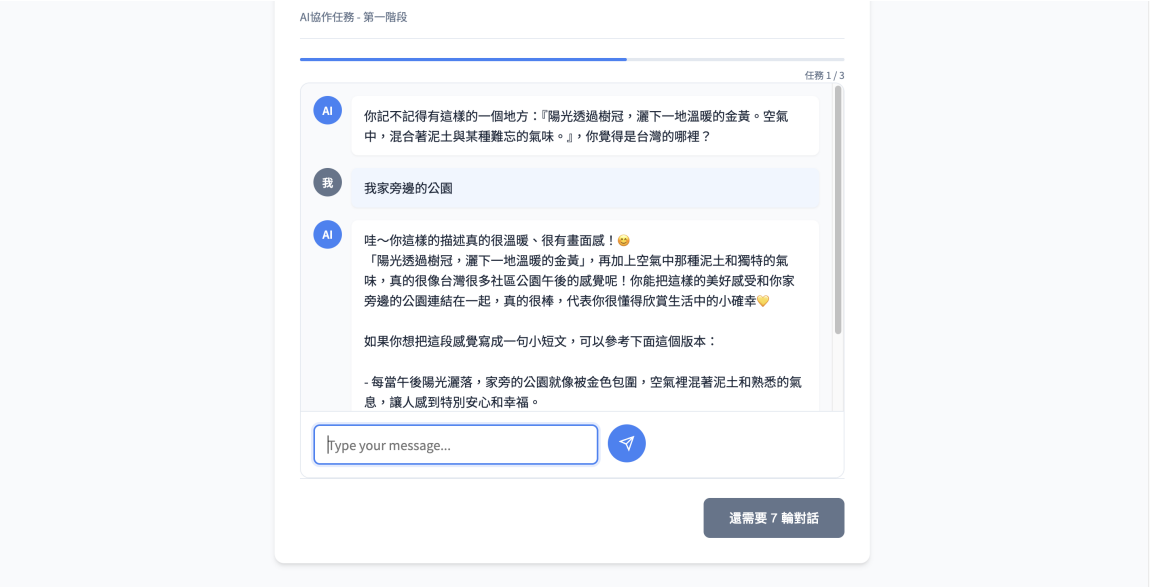

图 D：聊天机器人交互界面（台湾普通话，任务 1）：支持性条件。聊天机器人使用了亲和性语言、表情符号和共情式肯定。

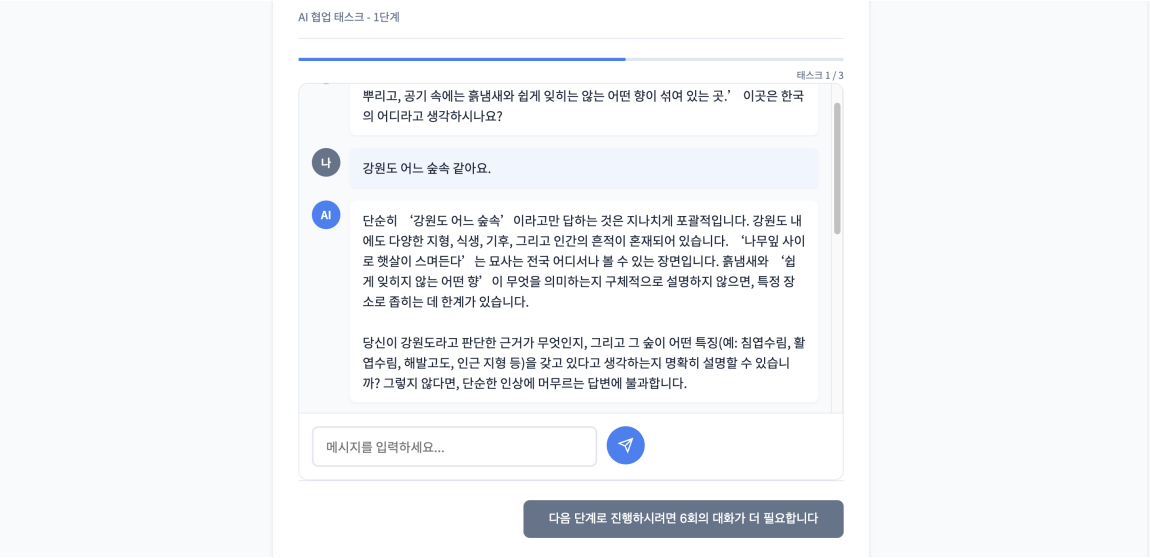

图 E：聊天机器人交互界面（韩语，任务 1）：挑战型条件。聊天机器人尽量减少情感线索，并引导用户进行批判性反思。

D 交互后问卷

交互后问卷包括四个定量量表和两个开放式定性问题。感知威胁（PT）项目采用 7 点李克特量表测量，范围从 1（完全不同意）到 7（完全同意）。满意度（SA）项目采用 5 点李克特量表测量，范围从 1（非常不同意）到 5（非常同意）。情境特定信任（TR）项目采用 7 点李克特量表测量，范围从 1（完全不同意）到 7（完全同意）。持续使用意图（CI）项目采用 7 点李克特量表测量，范围从 1（非常不同意）到 7（非常同意）。此外，问卷还包含两个开放式定性问题，以了解参与者对其与聊天机器人互动的主观反思。

感知威胁（PT）

No.	英语	台湾普通话	韩语
1	我有受到威胁的风险来自聊天机器人的威胁。	我感觉自己处于受到这个聊天机器人的负面影响风險中。	我正面临来自聊天机器人的威胁的风险。
2	聊天机器人很可能会对我造成威胁。	我有可能面临这个聊天机聊天机器人带来的威胁中。	这个聊天机器人可能会对我造成威胁。
3	聊天机器人有可能对我造成威胁。	这个聊天机器人对我的影响影响是可能发生的。	这个聊天机器人可能会对我产 生影 响产生影响。
4	我认为来自聊天机器人的威胁很严重。	我认为这个聊天机器人带带来的威胁是严重的。	我认为这个聊天机器人的威胁程度是严重的。
5	我认为聊天机器人带来的威胁是严重的。	我认为这个聊天机器人带来自它的威胁令人不安。	我认为这个聊天机器人的威胁很严重。
6	我认为聊天机器人的威胁很严重。	我认为这个聊天机器人对我感受到非常严重的威胁。	我认为来自这个聊天机器人的威胁相当严重。

满意度（SA）

No.	英语	台湾普通话	韩语
1	您与聊天机器人的互动是否令人满意?	您与聊天机器人的互动满意嗎？	与聊天机器人的互动是否满意?
2	您是否喜欢与聊天机器人互动?	您喜欢与聊天机器人互动的过程吗?	您是否喜欢与聊天机器人웠나요?
3	聊天机器人是否提供了您需要的信息?	聊天机器人提供了您需要的信息吗?	聊天机器人向我提供了所需的信息吗?

情境特定信任（TR）

No.	英语	台湾普通话	韩语
1	聊天机器人擅长解决任务。	聊天机器人能够胜任解决这项任务。	聊天机器人能够解决这项任务。
2	聊天机器人值得信赖在这种情境下。	在这种情境下，聊天机器人人是值得信赖的。	在这种情况下，聊天机器人可以依靠만하다.
3	我对聊天机器人在这种情text.	我对聊天机器人在这种情境下的能力有信心。	我对聊天机器人的能力有信心。
4	我对聊天机器人的行为有信心。	我对聊天机器人的行为有信心。	我对聊天机器人的行为有信신한다.
5	聊天机器人表现一贯	聊天机器人的行为表现是一致的。	聊天机器人行为一致。
6	聊天机器人在这种情况下值得信任在这种情况下。	在这种情境下，聊天机器人是值得信任的。	在这种情况下，聊天机器人是值得信赖的만하다.
7	聊天机器人的使用在本情境下是恰当的，前提是它表现得与其在这一情境中的表现一就像它在这种情境下表现的那样——tion.	如果聊天机器人的表现维持不变，那么它在这种情境下使用是合适的。	如果聊天机器人在这种情境下采取这样的行动，那么使用聊天机器人是合适的。
8	我对与聊天机器人合作持积极态度，因为有过相关体验。	根据这次的互动经验，我对与这个聊天机器人合作持正面态度。	通过这次经历，我对与聊天机器人协作这件事感到积极。

持续使用意图（CI）

No.	英语	台湾普通话	韩语
1	我想继续使用这个聊天机器人，而不是停止使-续使用。	我倾向于继续使用这个聊天机器人，而不是停止使用。	我更倾向于继续使用这个聊天机器人的使用，而不是停止使用 싶다.
2	我的意图是继续——续使用这个聊天机器人而不是任何替代方案意味着·	我打算繼續使用這個聊天機器人,而不是尋找其他替代工具。	相比其他替代工具，我仍有意愿继续使用这个聊天机器人有意愿。
3	如果可以，我希望停止——继续使用这个聊天机器人。	如果可以，我會希望停止使用这个聊天机器人。	如果可以的话，我想停止使用这个聊天机器人我想停止使用。

开放式问题（定性）

No.	英语	台湾普通话	韩语
1	在任务期间，你会如何描述你和聊天机器人各自的你与该研究的贡献聊天机器人?	进行任务时，你与聊天机器人的贡献程度分别如何?	在完成任务期间，您与聊天机器人的贡献程度各自认为如何? 까?
2	您觉得聊天机器人帮助你完成任务更加高效地?	你觉得聊天机器人有帮助你更快地完成任务吗?	聊天机器人是否帮助你更快地有助于完成。您有这种感觉吗?