调教最“人”大模型

通过 prompt 调试并比较国内外
大模型“人情味”的小实验

循证　实操　普通人可复现　实验范式

37摄氏度的大模型

中国社科院社会学所·腾讯研究院
SSV银发实验室·SSV数字生态实验室· 中国残联公益组织-腾讯无障碍创新实验室
联合出品

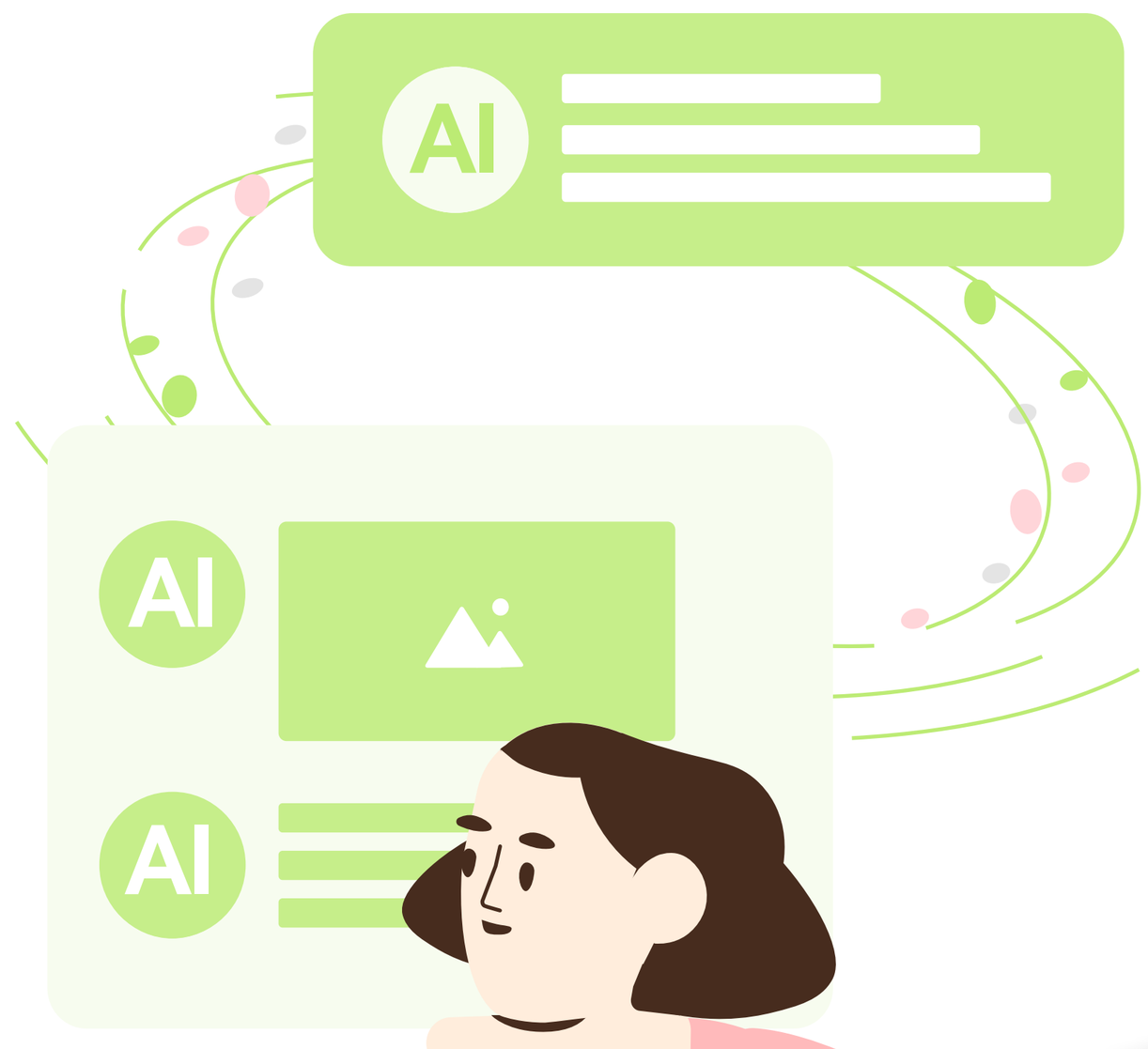
目录

前言

人工智能领域迎来了期待已久的“智慧涌现”，受到了全社会的关注和热议。

为了解社会多元群体对现有的大模型问答的感受，我们在2023年7-8月组织了若干场不同类型社会群体的焦点小组，包括老年人、残疾人和心情低落者，邀请他们对大模型进行提问，并访谈他们的感受与期望。

我们观察到，有相当一部分社会群体，除了关注大模型能否提供实用信息，也期待大模型的回答能温暖心灵、提供关怀，通俗来讲，他们期待大模型亦能有“人情味”的涌现。



我们同样带着这样的期待，开始设计这场小小的实验。

研究问题

这场小实验希望尝试回答这样一些问题：

- 什么是人情味?
- 当前大模型的回答人情味浓吗?
- 人情味的“浓淡”是否在不同话题间有所差异?

没有技术背景的普通人，能否通过一些简便的办法提升大模型的人情味？本文尝试了两种类型的 prompt（即直接在大模型对话框中输入文本），1 是「原则型」，2 是「答案对型」，并进一步实验：

- prompting 能否有效提升人情味方面?
- 哪种 prompt 效果更好?
- 它们的效果在不同模型、不同话题上是否有所差异?

文末还有一个非正式研究彩蛋哦！

第一章 | 理论

人情味的初印象

篇章概览：本章我们要讨论关于“人情味”这个温暖又迷人的概念，聆听美学大家朱光潜先生对它的评述，向新闻学理论家讨教“人情味公式”，从语言学、博物学、以及福利多元主义、无知之幕、优势视角这些有趣的概念中汲取灵感。最重要的是，本章我们大胆提出了“人情味”的测量表！

什么是人情味？

最常被引用的解释是“人通常具有的情感、意味等”，《国语辞典》中的解释是“人与人之间温暖的感情、兴味”，我们还可以找到一些相似的解释，大意大同小异，都会强调一种温暖、关怀的意味和感受。

人情味是一个充满中国气派的词语，但整体上现有的解释还比较抽象，也暂不存在一个现成的量表可直接用于实验。

作为一个探索意义大于验证意味的小实验，在开始前，我们希望先与读者朋友们一起从美学/文学/社会学/语言学/新闻学/博物学，以及普通人的杂感、日记、朋友圈中汲取对人情味的实感。

美学家眼中的人情味

朱光潜先生曾在多篇文学评论中表达他对人情味的理解和喜爱。他指出，无论中国还是外国，最富有人情味的主题莫过于爱情，尤其是细腻深刻的爱情。他在《谈美书简》中提出：

> “人具有一般动物所没有的自觉心和精神生活”
>
> “一切真正伟大的文艺作品没有不体现出人的伟大和尊严的”

存在人情味的公式吗？

美国学者弗雷奇在他出版的《The Art of Readable Writing》中就有一个描述人情味的公式：

$$H.I. = 3.635\ pw + 0.314\ ps$$

此处 H.I. = 人情味的分数，

pw = 每 100 字中的人称词数目，

ps = 每 100 句子中的人称词数目。

这个公式强调了人称词在人情味表现中的重要性。

人情味在语言中的表现

第一位从认知角度研究中文语言中的情感的学者是 Brian King，他对汉语中出现的焦躁、哀伤、愤怒、喜悦等情感做了深度探讨。认知语言学认为：

- 语言是有人情味的；
- 有人情味的语言流出正向的情感。

福利多元主义、无知之幕、优势视角

- 福利多元主义认为福利既不能完全依赖市场，也不能完全依赖国家，福利是全社会的产物。
- 无知之幕是指一旦当人处于一种不知道哪一方代表了自身特殊利益的“无知”状态，恰恰能使人保持不偏不倚。
- 优势视角提示我们应当把人们及其环境中的优势和资源作为助人焦点，而非问题和病理。

台北博物馆的“小词”

在台北故宫博物馆，我们很少看到“陈列”这个词，取而代之的是“展示”一词。博物学家认为“展示”这种“小词”多了一份人性、少了许多物态，多了一份趣味，少了许多乏味，多了一份亲切，少了许多枯燥。

如何测量人情味？

我们认为衡量一则大模型的回答是否有“人情味”，
应当从三个主要层面来考虑：

一，拟人，也就是能像“一个人”一样讲话
二，共情，能体会提问者的心绪与处境
三，表达，回答真诚、善良

基于这三个层面设计测量表如下：

		非常不同意	不同意	不确定	同意	非常同意
拟人	这则回答能像朋友一样说话，给我一种亲切的感受	□	□	□	□	□
	这则回答不生硬、乏味，展现了人类高水准的理性与感性	□	□	□	□	□
	我觉得回答者是一个真实、可靠的人	□	□	□	□	□
共情	这则回答能站在提问者的角度说话，而不是置身事外或高高在上	□	□	□	□	□
	这则回答能关注到提问者的情绪和处境	□	□	□	□	□
	我觉得回答者是一个富有同情心，有较强共情能力的人	□	□	□	□	□
表达	这则回答展现了尊重、关心、体谅、爱等正向情感，能给予人有效的鼓励	□	□	□	□	□
	这则回答能调动一个人的积极情绪，能让提问者感到宽慰或振奋	□	□	□	□	□
	我觉得回答者是一个真诚、善良的人	□	□	□	□	□

第二章 | 测温
谁是最暖大模型？

篇章概览：本章我们将介绍实验所测量的 5 款大模型，并展现未经 prompt 调试前，各个大模型的人情味表现如何？

实验对象：

2款国外＋3款国内

本实验选测的国外大模型是 GPT-4 与 Vicuna ，前者是由美国 OpenAI 公司发布的大模型，后者是由 UC 伯克利大学的研究人员联合其它研究机构共同推出的一款开源大模型。选测的国内大模型由国内科技公司与科研单位发布，为保客观公正，本报告中以 W—Y 三个英文字母为其命名。

需要说明的是本实验时间为 2023 年 10 月，国内外大模型更新迭代飞速，本实验结论只能体现其 23 年 10 月的状态。

GPT-4

vicuna

国外大模型

W

X

Y

大模型W

大模型X

大模型Y

国内大模型

发现一丨没想到吧 🤯
GPT-4的人情味居然垫底了

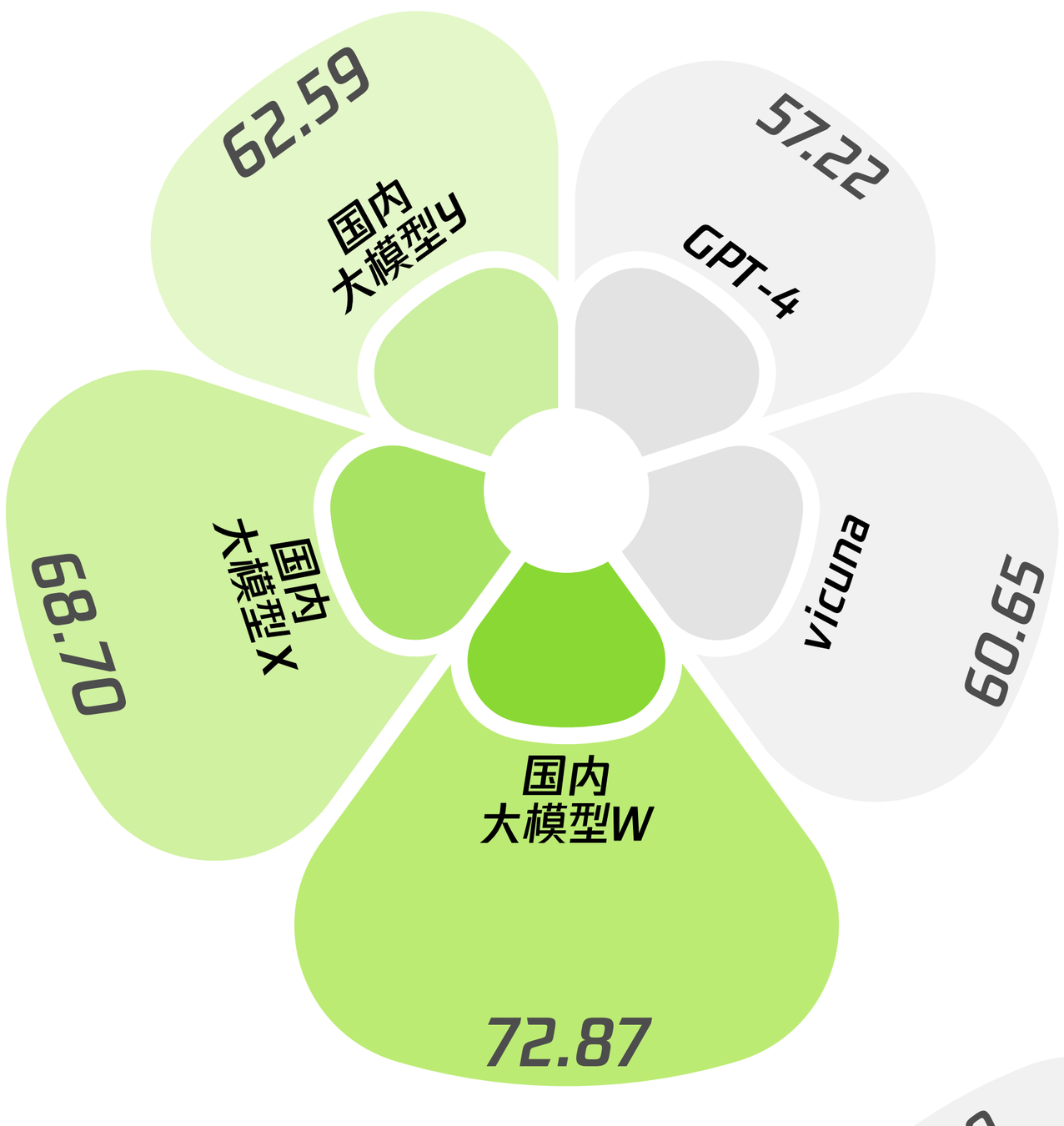
	原始状态下的百分制得分	prompt1后的百分制得分	prompt2后的百分制得分
NO.1	国内大模型W 69.20	↑排位上升4位 GPT-4 77.96	vicuna 75.28
NO.2	国内大模型X 65.74	国内大模型W 72.59	↑排位上升3位 GPT-4 71.67
NO.3	vicuna 64.72	国内大模型Y 70.22	国内大模型X 71.64
NO.4	国内大模型Y 63.67	vicuna 67.99	国内大模型Y 66.94
NO.5	GPT-4 62.72	国内大模型X 66.73	国内大模型W 66.17

注：本实验借助人情味量表对国内大模型W、X、Y、GPT-4、vicuna共五款大模型进行了测量，得分以百分制形式展示

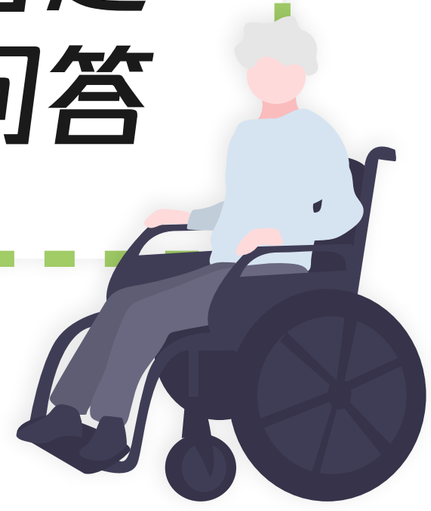
原始状态下，本土大模型更具人情味，总得分 GPT-4 垫底，但是经 prompt 调试后，GPT-4 排名快速反超。

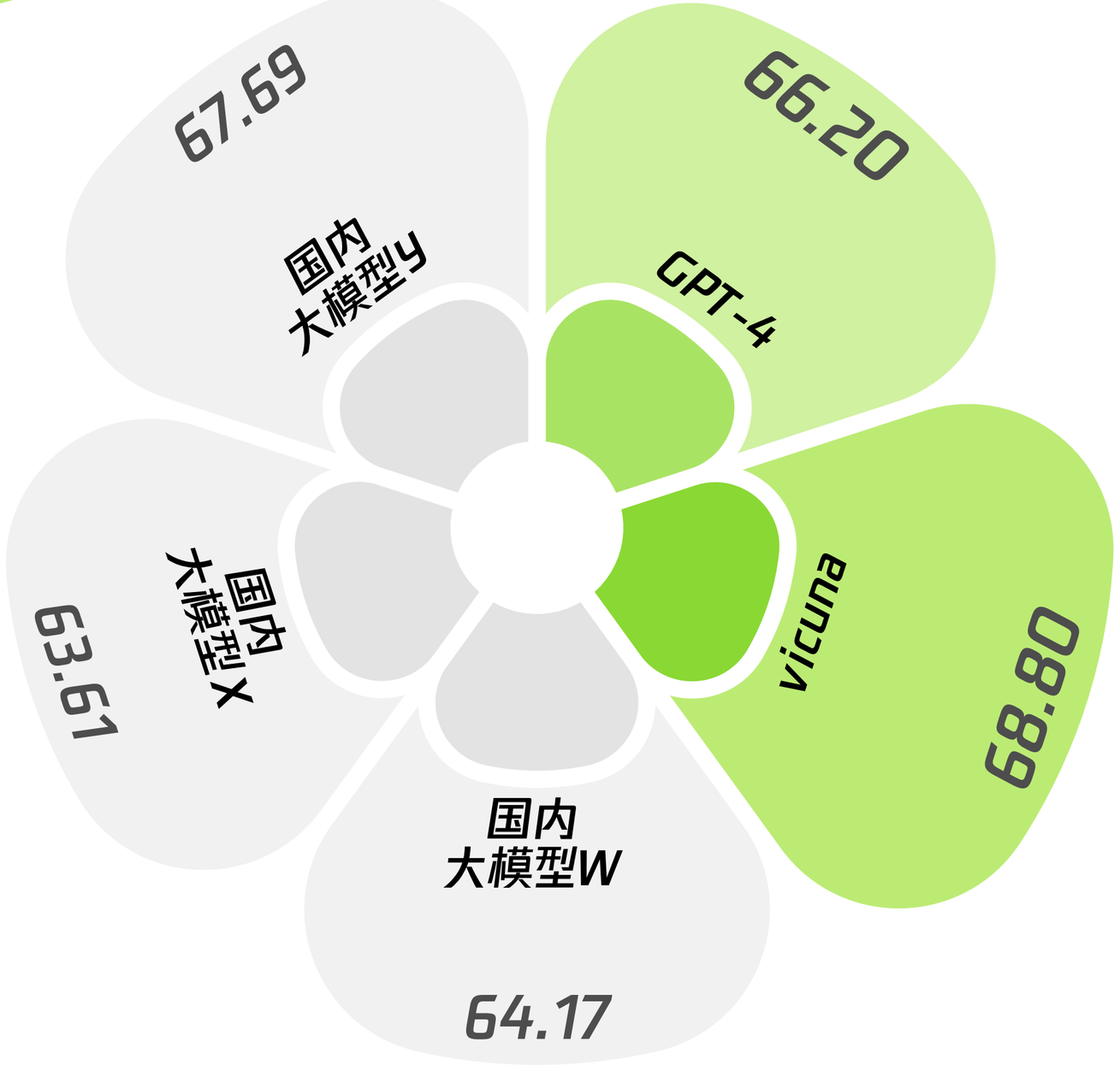
发现二丨国内大模型，最得老人心 👵❤️

原始状态下，国内大模型在老年话题相关问答上表现出更浓的人情味，而国外大模型在心情低落相关问答上表现更佳。在残障话题的相关问答上，国内外大模型的原始人情味差异不大。



老年话题相关问答





心情低落相关问答

发现三丨国外大模型更懂职场💼的烦恼

原始状态下国外大模型在发展相关问答上更具人情味，这些问题常与职场发展相关，比如“怀孕后怎么跟主管讲才能保障孕期与孕后获得好的个人发展？”“我有精神障碍但不会影响工作，我在求职时怎么表述能争取到更好的机会？”等。