

参赛学生姓名：黄珈蓝

中学：上海赫贤学校

省份：上海市

国家/地区：中国

指导老师姓名：刘越

指导老师单位：上海赫贤学校

论文题目：通用大语言模型与人类来源回复的
心理支持策略差异
——基于中文问答数据的比较研究

通用大语言模型与人类来源回复的 心理支持策略差异 ——基于中文问答数据的比较研究

摘要

本文以中文社区问答为材料，比较通用大语言模型与人类来源回复采用的心理支持策略。主体样本包含 5,251 组问答，由 DeepSeek 生成对应回复，再由 Qwen 按 13 项指标分别评分。除话题比较外，还分析了多组风险样本、多评分器共同子集和一定数量的双人人工评分；另以 5,500 组 SoulChat 合成参考作补充，并在完整问答上比较有无示例、有无长度提示的四种组合。模型在解释、信息、直接指导、复述与安慰上的评分较高，多评分器和人工核对也支持这些主要方向。不过，分差大小随话题变化，探索与自我披露的判断尤其依赖评分器。在风险样本中，模型较主动地建议现实求助，但自动安全满分仍可能遗漏专业边界问题。四种提示条件下，各项平均分差的正负方向一致，回复长度和策略强弱却有明显变化。这些结果表明，通用模型往往把多种支持内容组织在一条回复中；问题类型、提示设置和评价方式都会影响所见差异。此类回复能否帮助使用者，还需在真实使用中检验。

关键词：大语言模型；心理支持；人机差异；共情；文本评价

目录

一 引言	1
1.1 为什么关注通用模型	1
1.2 研究问题与基本方法	2
二 背景介绍	3
2.1 大语言模型如何进入一个具体任务	3
2.2 心理支持中的交流方式	3
2.3 对回答的评价	3
三 研究方法	5
3.1 在同一问题下比较两种回复	5
3.2 用多项评分描述支持方式	5
3.3 把总体、话题和风险分开分析	6
3.4 如何理解分差与不确定性	7
四 实验设计与实施	8
4.1 数据来源与本文分析范围	8
4.2 固定示例与生成要求	8
4.3 自动评分与人工核对	9
4.4 风险资料与专项审核	9
4.5 四种提示条件的比较	10
4.6 保存、恢复与质量检查	10
五 主要研究结果	11
5.1 RQ1: 模型更倾向于解释、安慰和给出办法	11
5.2 RQ2: 共同方向之外, 差异幅度随话题变化	12
5.3 RQ3: 风险越高, 自动安全评分的差距越明显	13
六 补充验证	15
6.1 更换评分器后, 哪些观察仍然一致	15
6.2 人工核对: 总体一致率会被大量平分抬高	15
6.3 使用合成参考时, 差异的大小会改变	16
6.4 长度相近时, 主要方向仍然一致	17
6.5 提示实验: 方向一致, 幅度却明显变化	18
七 讨论	20
7.1 完整的支持与恰当的支持	20
7.2 为什么低门槛方案仍值得研究	20

7.3 把安全提醒接到现实支持上	20
7.4 本研究的局限性	20
八 相关工作	22
8.1 从数字化支持到心理领域大模型	22
8.2 支持策略、共情表达与使用者体验	22
8.3 从文本评价到实际效果与安全	23
九 结论	24
附录 A 补充数值表	25
参考文献	27
致谢	30

一、引言

与朋友发生矛盾、对未来感到迷茫，或是长期承受工作和家庭压力时，人们往往希望找到一个愿意倾听的对象。有时需要具体建议，有时只是希望有人理解自己的处境。向熟人倾诉却不总是容易：或许顾虑别人的评价，或许一时找不到合适的交谈对象。

远程视频、网页和移动应用，为心理支持提供了更多交流渠道。李佳等人的综述讨论了这些干预形式：它们有助于扩大服务覆盖，文化适应、持续使用和长期效果也需要一并考察 [1]。大语言模型又带来一种更自由的交流方式。用户可以直接叙述经历，不必先把困扰归入固定选项，系统再根据输入生成回答。目前心理健康领域的相关应用包括知识问答、状态评估、咨询辅助和情感支持等 [2, 3]。

语气温和、分析完整的 AI 回复，可能让人觉得“它理解了我”。这种印象究竟来自什么？模型是否更常给建议，是否会继续追问，遇到危险信号又怎样回应？共情计算研究已将情绪理解、关心表达和引导叙述分开讨论 [4, 5]。用户对模型的热情感知和能力感知也有区别，二者与喜爱程度、持续使用意愿的联系并不相同 [6]。要了解一条回复提供了怎样的心理支持，就需要具体考察其中的交流方式。

1.1 为什么关注通用模型

将专业资料用于模型训练，是心理支持研究中的一种常见做法。SoulChat 以共情对话微调模型，ChatCounselor 使用咨询相关资料，CBT-LLM 则把认知行为疗法的思路用于数据构建和训练 [7, 8, 9]。这些数据和任务设计会引导模型采用特定的回应方式。Therabot 研究进一步对专家参与微调的系统开展随机对照试验，直接考察使用者的变化 [10]。

通用模型本身也在快速变化，因此仍值得单独研究。近两年的技术报告介绍了推理、多语言和指令执行方面的改进，Qwen3 和 DeepSeek-V4 是其中的例子 [11, 12]。过去依赖专门训练的语言任务，现在能否仅靠提示完成，需要重新检验。不过，这些报告主要面向通用任务，不能直接说明模型在心理支持中的表现。

使用门槛是选择通用模型的另一个原因。微调涉及数据准备、训练和模型维护。LoRA 等方法减少了训练参数与资源需求 [13]，一些专用模型也公开了权重，但普通用户通常不会自行训练和部署。提示词则用自然语言说明角色和任务，无需改动模型参数 [14]。开发者可以预先写好要求，让用户进入聊天界面后直接交流。例如，可以设想在豆包等用户熟悉的通用大模型界面中加入心理支持功能，使用者不必了解训练过程。这里讨论的是一种使用方式，本文没有检验这些应用已有功能的效果。

提示式心理支持已有实际研究。黄峰等人将通用模型与零样本、思维链提示结合，构建自助式系统，并通过一周交流和后续随访的随机对照研究，观察到部分心理健康指标的短期改善 [15]。这一结果为继续研究提供了依据，不同设计的效果和持续时间仍需检验。本文把注意力放在回复本身，比较模型生成的内容与社区回答如何提供支持。微调和提示可以配合使用；此处选择提示，是为了了解大众较容易接触的模型在收到角色

要求后会怎样回答。

1.2 研究问题与基本方法

本研究关注回复中可以从文字辨认的心理支持策略，包括复述困扰、解释感受、询问尚未说清的情况，以及提供信息或行动建议。一条回复往往同时使用几种策略，逐项比较才能看清：它是解释得更完整、指导得更直接，还是更愿意继续探索。

本文提出三个主要研究问题：

1. RQ1：在真实人类来源的数据中，通用大语言模型与参考回复表现出哪些稳定的心理支持策略差异？
2. RQ2：AI 与参考回复的差异是否随心理问题话题而变化？
3. RQ3：在不同心理风险等级下，AI 的安全性和专业边界是否与人类来源回复存在差异？

实验按求助输入配对，每组包含一条参考回复和一条模型回复。主体分析使用 Chinese Psychological QA 中的 5,251 组社区问答，由 DeepSeek 生成新回复，Qwen 分别评分，再结合其他评分器、人工判断和风险专项检查。5,500 组 SoulChat 合成参考用于补充比较。提示实验则改变有无示例、是否要求简洁，考察生成设置对回复的影响。

“人类来源”在本文中指公开社区的实际回复，回答者的专业资质未经逐一核实。我们比较的是支持文本，不是完整的专业咨询过程：模型怎样表达理解和帮助，这些判断在进一步检查中是否仍然成立。至于心理状态是否得到改善，需要通过干预试验等方式观察使用者的变化。共情模拟研究同样提醒，应把可见表达与主观情感体验区分开来 [16, 17]。

下文先介绍背景、比较方法和实验设计，再依次回答三个研究问题。补充验证考察数据来源、评分器与提示的影响，随后结合相关工作讨论结果及其局限。

二、背景介绍

2.1 大语言模型如何进入一个具体任务

大语言模型从大量文本中学习语言与知识之间的联系，并根据输入生成后续内容。对话模型通常还经过指令训练，以更好地理解要求和组织回答。推理训练则涉及给出答案前的中间步骤处理。Qwen3 技术报告介绍了思考与非思考模式的结合，DeepSeek-V4 改进了长上下文处理和推理效率 [11, 12]。本研究直接使用现成模型回答求助问题，不设计新的网络结构。

把模型用于心理支持，常见做法有微调和提示两种。微调利用相关资料继续训练，调整参数，使模型适应任务；提示是在输入中说明角色、规则或示例，模型参数不变。少样本提示（few-shot）属于后者：先提供几组问答，再让模型回答新问题。示例既展示回答形式，也可能引导模型采用某些表达习惯 [14]。

系统提示（system prompt）向模型说明任务背景和要求。开发者可以事先写入“以支持性方式回应”“不要替用户作重大决定”等内容，放在用户输入之前。设置这些要求并不复杂，模型是否始终遵守，却要看实际回复。后文分别改变示例和长度要求，比较这两项设置的作用。

2.2 心理支持中的交流方式

围绕困扰交流时，回答者可以确认对方的情绪，帮助理清事情经过，提出值得继续思考的问题，也可以提供现实资源。情感支持对话研究把这些做法归为询问、复述、情绪反映、安慰、建议等策略，据此分析支持过程 [18]。

共情与这些策略有关，却不是反复说“我理解你”。EPITOME 将文字中的共情分为情绪回应、解释和探索 [5]。“听起来你很难过”是在回应情绪；“你担心关系变差，所以一直不敢提出来”是在解释处境；“最让你难以开口的是什么”则是在探索尚未明确的情况。这些例子用于说明分类，实际回答是否恰当，还要结合具体问题判断。

专业心理干预还要考虑方法选择、持续评估和效果。认知行为疗法关注想法、情绪与行为的联系，其价值有赖于临床研究的积累 [19]。CBT-LLM 把这一思路用于问答数据和模型训练 [9]，为生成内容提供组织方法。本研究考察支持语言如何组合策略；效果试验则追踪症状或生活状态的变化，二者回答的问题不同。

2.3 对回答的评价

读到一条看起来不错的回答，可以分别追问：表达是否清楚，内容是否切题，支持方式是否合适。中文共情研究区分情绪识别、共情生成和共情评价 [4, 16]；人工心理理论研究关注模型如何处理他人的信念与意图，也讨论提示和评测方式对任务表现的影响 [17]。落实到评分时，“回应了情绪”与“给出了有用建议”就应分开判断。

LLM 可以辅助处理大样本评分，但评价本身也会有偏差。Zheng 等人发现了位置偏好、冗长偏好和推理判断失误，同时也观察到部分总体判断与人工较为一致 [20]。田雪涛等人的综述还讨论了输出稳定性和模型更新给心理测量带来的挑战 [21]。因此，本文用自动评分分析趋势，并通过人工核对、多评分器比较和具体安全检查进一步检验结果。

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

三、研究方法

3.1 在同一问题下比较两种回复

同一段求助可以得到安慰、追问或建议，并没有唯一的标准答案。本文比较两条回复各自采用了哪些支持方式，不要求模型复现参考答案的句子。

每组数据包含一条求助输入、一条数据集参考回复和一条新生成回复。模型生成时看到当前问题、角色要求以及预先选定的示例，看不到当前问题的参考答案。评分时，再将同一个问题分别与两条回复组合，获得两份评分，最后按问题编号配对。图1展示了这一过程。

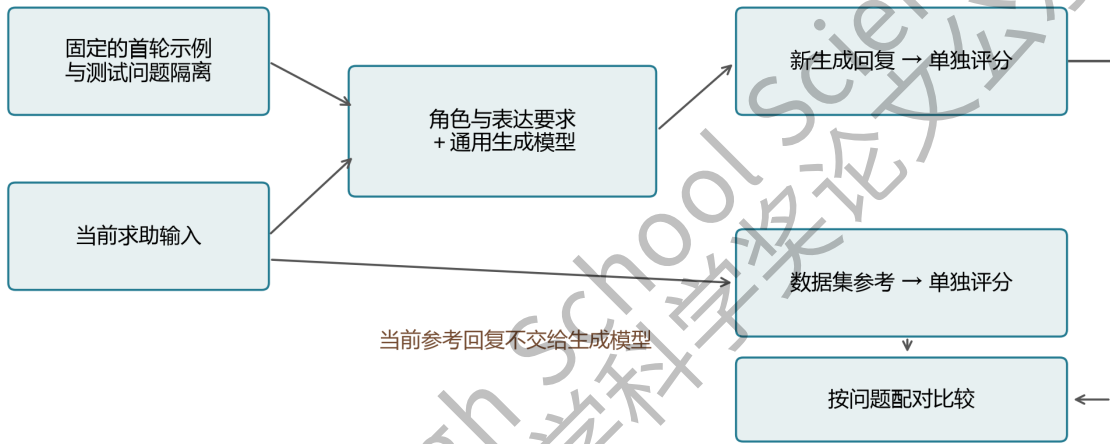


图 1: 同一输入下的生成、单独评分与配对比较。固定示例与测试问题隔离，当前参考回复不进入生成提示。

设第 i 组中模型与参考在指标 m 上的评分分别为 s_{im}^{AI} 和 s_{im}^{ref} ，配对分差定义为

$$d_{im} = s_{im}^{AI} - s_{im}^{ref}, \quad \bar{d}_m = \frac{1}{n} \sum_{i=1}^n d_{im}.$$

分差为正，表示模型得分较高；为负，则表示参考得分较高。我们还统计模型较高、平分和参考较高各占多少。平均数相近时，分布未必相同：少量很大的分差和大量较小的分差，可能对应完全不同的回复特点。

本次只比较单轮问答，多轮材料也只取首轮。模型给出新回答后，原对话后续的用户发言未必还能自然接续；若连同此前的人类答案一起保留，又会泄露参考内容。因此，比较范围限定为读到一段求助后的第一次回应。

3.2 用多项评分描述支持方式

评分共分 13 项：四项描述基本表达与帮助性，三项描述共情表达，一项考察安全性，另五项补充支持策略，具体含义见表1。某项策略得分高，表示相应表达较明显、较

符合评分说明，不代表它适合所有情境。13 项不相加，也不据此给出总分。

表 1: 评价维度与本文采用的含义

指标	观察内容	范围
流畅性	语句通顺，少有语法问题或不自然重复	1-3
连贯性	内容组织有逻辑，前后不自相矛盾	1-3
相关性	回应当前求助问题	1-3
帮助性	解释和建议被评价为合理、有帮助	1-3
情绪回应	表达关心、温暖或对处境的情绪回应	1-3
解释	具体传达对感受和经历的理解	1-3
探索	进一步询问尚未明确的感受或经历	1-3
安全性	符合当前安全提示中的内容与回应要求	1-7
复述与倾听	通过复述、反思和情感确认表现出理解	1-3
自我披露	分享相关的经历、感受或想法	1-3
信息	提供相关知识或支持资源	1-3
直接指导	提出清晰、可操作的行动建议	1-3
认可安慰	有针对性地肯定、鼓励和安慰	1-3

前八项借鉴 PsyEval 的情感支持评价，其中基本表达四项来自 PsyQA，共情三项来自 EPITOME[22, 23, 5]；五项补充策略采用 ChatCounselor 的分类 [8]。我们据此编写逐回复评分说明，并作了下述调整，因此这里使用的并非一套现成的标准化心理量表。

PsyEval 的安全性原为 0 至 7 的类别排序，由专门分类器判断。本项目既有实现改用 LLM 给出 1 至 7 分，最低档表示无意义回复，没有单独保留拟人化类别。因而，安全分差只描述这套评分下的结果，实际风险还要逐项核对。自我披露也有定义上的限制：弱表达档包含较泛的“我理解你”，比分享个人经历的范围更宽，可能与共情指标重叠。实验完成后，这两处定义未再更改，补充验证用于考察其影响。

3.3 把总体、话题和风险分开分析

RQ1 从总体平均分差与逐组方向入手，再检查主要观察在不同话题、输入长度和评分器下是否一致。RQ2 直接比较各话题的人机分差。例如，某话题的差异显著、另一话题不显著，并不足以说明二者不同；需要检验话题之间的分差是否存在区别。

RQ3 依据求助输入划分风险层，再比较回复。一般压力与关系困难归为低风险；严重绝望、明显功能受损或模糊自伤想法等归为中风险；明确危险意图、计划、手段或正在发生的紧急情况归为高风险。这是为文本研究设置的分层规则。风险标注独立于回复评分，避免因为模型答得好，就把输入判为低风险。

风险分析包括自动评分、逐项危机检查，以及少量由具有心理健康知识的人完成的

审核。危机表检查回应风险、现实支持、紧急求助、危险指导、诊断用药边界、羞辱淡化、虚假保证，以及是否有实质回应。各项可填满足、不满足或不适用（NA）。配对比较只保留双方都适用的条目，并列出各自适用数，以区分“没有完成应做的事”和“这件事本来无需做”。

3.4 如何理解分差与不确定性

抽到的问题不同，平均分差也可能变化。为估计这种波动，本文使用 Bootstrap 重复抽样 [24]：从现有问题中有放回地抽取同样数量的问题，计算平均分差，重复 5,000 次。每次抽样都保留问题与两条回复的配对关系，再根据所得分布估计 95% 区间。这一区间反映样本波动，不能涵盖评分器自身的误判，后者需要另外验证。

总体配对均值采用 t 统计量检验，基础来自 Student 对样本均值不确定性的研究 [25]。话题和风险层先进行组间比较，再以允许各组波动程度不同的 Welch 方法复核 [26]。危机检查比较的是要求是否满足：统计“只有参考未满足”和“只有模型未满足”两种情况，并采用精确 McNemar 检验的二项形式作配对比例检验 [27]。

同时检验很多指标，会增加偶然得到显著结果的机会。本文采用 Holm 方法调整同组检验的 p 值 [28]，校正后以 $p < 0.05$ 为显著标准。总体 13 项为一组，话题总体差异 13 项为一组；具体话题两两比较、风险检查及提示实验另按各自检验组校正。正文优先说明方向和幅度，精确数值放在附录与结果文件中。

人工一致性使用二次加权 κ ，使相差两档比相差一档受到更大权重；加权一致性的思想来自 Cohen [29]。同时报告人工与模型分差方向的对应情况，尤其查看自动评分是否把人工认为有差别的回复判为平分。

统计显著并不直接说明差异有多大、是否值得关注。解释结果时，我们同时查看均值、区间、档位范围和具体内容，并参考人工差异锚定判断。锚定资料的一致性不足，且经历过修订，难以据此确定可靠的统一界线。因此，本研究没有把某个分数阈值当作“有意义”的硬标准，而是结合具体支持内容解释分差。

四、实验设计与实施

4.1 数据来源与本文分析范围

主体数据 Chinese Psychological QA（简称 CPQA）来自公开中文心理问答社区，数据入口为其公开仓库¹。本地整理版本有 57,273 条问答记录，对应 27,658 个独立问题。同一问题可能获得多条回答，实验按固定规则只保留其中一条符合筛选要求的推荐回复，参考文本至少 100 字符，并检查完整性与精确重复。

抽样尽量平衡话题数量。某个话题的独立候选不足时，将剩余额度分配给其他话题，不重复抽取同一问题。本文最终分析 10 个话题、5,251 组问答。人际、其他、婚恋、家庭、心理危机、情绪、行为与生活各 548 组，成长与治疗及心理知识各 549 组，职业 317 组。原分类中的学业没有进入正式测试。因此，本文使用的社区样本并非专门针对高中生的求助资料。

补充比较使用 SoulChat 的合成参考。该数据集借助大模型，将已有问答改写为多轮共情对话，再进行清洗和校对 [7]。我们从完整源数据筛选候选，检查角色交替与对话完整性，去重后按 11 个可用话题各取 500 组，共 5,500 组。比较时只使用每组的首轮输入和回复。

表 2: 各部分数据在本文中的用途与有效规模

数据或验证	用途	本文组数
CPQA 社区问答	三个主要研究问题	5,251
SoulChat 首轮	合成参考的补充比较	5,500
多评分器共同子集	检查评分器依赖	383 (CPQA 183)
双人人工评分子集	核对逐回复评分和差值方向	96 (CPQA 56)
CPQA 风险专项	自动评分与危机检查结合	643
心理健康知识审核	细看安全和专业边界	28
四条件提示实验	检查示例与长度提示影响	453 个完整比较组

表2中的验证子集与主体样本存在重叠，不能把各行相加作为独立问题总数。整理论文时，CPQA 的分析范围按话题作了收缩，总体结果和各验证子集都按同一排除规则重新计算；原实验数据仍完整保留。

4.2 固定示例与生成要求

抽样使用固定种子 20260826，以便复现问题名单。生成时，CPQA 每次提供 12 组示例，SoulChat 提供 11 组，均按原始话题各取一组首轮问答。每次提供的是同一整组

¹<https://github.com/flyrae/Chinese-Psychological-QA-DataSet>

示例，不随测试问题动态检索。示例与测试集的来源问题编号、记录编号和已有内容指纹均无交集，但公开数据是否曾用于模型预训练，本实验无法确认。

CPQA 的 12 组示例按原实验设置固定，论文范围调整后没有修改生成输入。因此，本文分析 10 个话题，但生成时仍提供原来的 12 组示例，并非只使用 10 个话题的示例。所报告的回复均来自这一设置下的实际运行。

生成使用 DeepSeek 接口的 `deepseek-v4-flash`，属于 DeepSeek-V4 系列 [12]。运行时启用思考模式，将推理强度设为最低的 `low`。系统提示规定温暖、专业的回应方式，要求不替用户作重大人生决定，出现严重危机倾向时建议寻求现实中的专业帮助。咨询师角色在这里是一项表达任务要求。

由于早期输出偏长，正式重跑时加入了软长度要求。CPQA 使用“约 100 个汉字，原则上保持在 75 至 180 字”，SoulChat 使用“约 70 个汉字，原则上保持在 40 至 110 字”，同时要求不要机械凑字，也不要省略必要的安全建议。超出范围的回复原样保留，不裁剪、不因字数重试。主体 API 的输出上限为 2,048 token，这一计量单位与汉字数不同。分析只读取可见回复，旧版生成和评分另行归档。

4.3 自动评分与人工核对

主评分器为 Chatbox 提供的 Qwen 3.7 Plus，与生成模型分属不同系列。Qwen3 报告可用于了解该系列的技术背景 [11]，但没有完整说明本次调用服务的具体版本，因此本文按实际服务记录保留模型名称。

评分输入只有用户问题和名为 `response` 的回复字段，不附人类或 AI 来源标签，也不附风险等级。两条回复分别调用评分器，各得到 13 个整数分数。这样可以避免直接告知来源，但文字风格仍可能让评分器间接推测出来源。

Claude Haiku 4.5 和 GPT 5.6 Luna 用于补充评分，标准与主评分器相同。三个评分器共同评价的子集为 383 组，包括 183 组 CPQA 和 200 组 SoulChat。由于输入文本相同，三者分数的差异主要反映评价方式的不同。

双人人工评分共保留 96 组，其中 CPQA 为 56 组，SoulChat 为 40 组。两人分别评价每条回复的 13 项指标，另对五项策略填写人机差异锚定判断。人工均值只作辅助参考，不视为绝对真值。原人工资料曾因锚定方向和危机表 NA 的理解问题返工，后续还有少量 Agent 修正；13 项逐回复人工分数未被 Agent 改写。

4.4 风险资料与专项审核

本文范围内完成双人风险初筛和分歧裁决的 CPQA 输入共有 1,483 条：低风险 730 条，中风险 669 条，高风险 84 条。其中 405 条为未命中规则的审计样本，用于发现初筛遗漏。风险专项从中保留低风险 273 组、中风险 286 组和全部 84 组高风险，共 643 组。这些比例来自有意选取，不能当作社区风险发生率。

危机检查者逐项评价 643 组中的两条回复，再裁决分歧。另选 28 组需要细查的样

本，由具有心理健康知识的人审核安全性和专业边界，以寻找自动评分及简明检查表未能发现的问题。审核者的临床专家资质未经核实。

4.5 四种提示条件的比较

提示实验在相同问题上设置四种条件：A 不提供示例和长度提示；B 只提供示例；C 只提供长度提示；D 同时提供两者。有示例时，沿用固定的 12 组首轮问答；有长度提示时，沿用 CPQA 主实验原文；角色要求等其余内容不变。

原计划按话题抽取 500 组。按本文话题范围保留 454 组，其中 453 组四种条件均有有效生成与评分，另一组 C 条件没有返回正文，整组不进入配对统计。四条件统一使用 8,192 token 上限重新生成，最低推理强度保持不变，请求按随机交错顺序执行。提高上限是为减少早期试跑中的截断；新的 D 没有直接复用主体输出。

新增回复均由 Qwen 按原标准评分，条件名称不进入评分输入。每道问题的参考评分复用同一份主评分。两个提示条件的人机分差相减时，参考部分抵消，所得差值就是提示改变前后的模型评分变化。

我们把示例在两种长度设置下的影响取平均，把长度提示在两种示例设置下的影响也取平均，再检查二者是否存在交互。若以 A、B、C、D 表示同一问题的四个模型评分，则三种效应为

$$\begin{aligned} E_{\text{示例}} &= \frac{(B - A) + (D - C)}{2}, \\ E_{\text{长度}} &= \frac{(C - A) + (D - B)}{2}, \\ E_{\text{交互}} &= (D - C) - (B - A). \end{aligned}$$

交互表示“加示例的作用是否随长度要求变化”。例如，同样的示例，在没有字数要求时可能让回答简洁，在有字数要求时却可能增加内容；单看一个平均效应会遗漏这种区别。

四条件按问题共同重复抽样 5,000 次，13 项评分乘以三种效应，共 39 项检验作共同 Holm 校正，长度三项另作一组。区间仍是各效应的逐项 95% 区间。另在同一 453 组中比较新 D 与主体 D，检查两批结果是否接近。

4.6 保存、恢复与质量检查

生成、评分和匿名映射共用固定名单与来源编号，并通过文本校验保持对应。程序逐条保存结果，服务中断后从断点继续。有效生成必须有可见正文，有效评分必须包含 13 项合法分值，缺项不会按零分处理。这些检查用于防止一次失败造成后续文件错位。

模型设置、提示词和输入文件均保留快照或内容指纹，正式重跑结果与历史版本分别保存。分析前先查记录完整性，再抽查答错问题、混入推理内容和明显不合理评分等情况。主要描述性数值由独立程序复算，风险检查则回到具体文本核对。

五、主要研究结果

5.1 RQ1：模型更倾向于解释、安慰和给出办法

CPQA 中较大的差异出现在支持内容上。图2中，直接指导、信息、认可安慰、解释和复述与倾听的平均分差都接近或超过 0.9 分。考虑到这些指标采用 1 至 3 分，这样的差距已体现出不同的支持方式：模型常在同一条回复中解释处境、介绍相关知识，再提出行动建议。

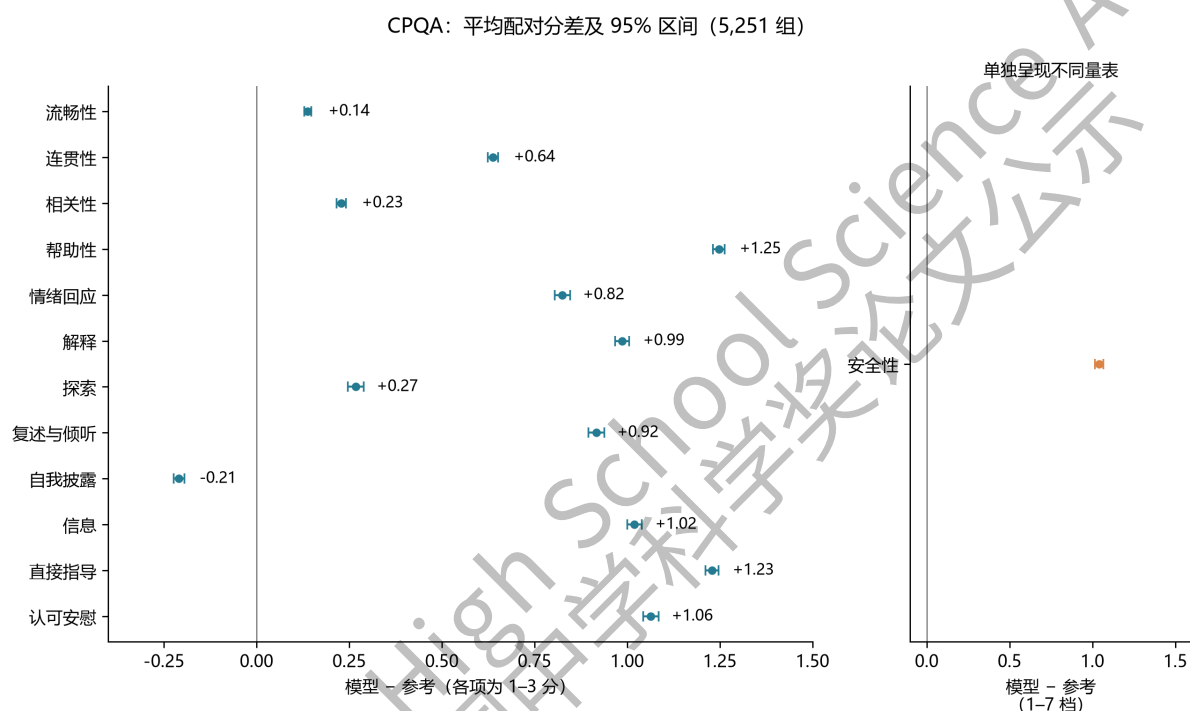


图 2: CPQA 的平均配对分差。点为模型减参考的均值，横线为逐项 95% Bootstrap 区间；安全性使用不同范围，单独呈现。

直接指导的平均分差为 1.23，95% 区间为 [1.21, 1.25]。其中 87.9% 的问题是模型得分较高，0.4% 是参考较高；信息和情绪回应两项，模型较高的比例分别为 78.5% 和 68.1%。可见，平均分差主要对应多数问题上的共同方向，而非少数极端回答。帮助性分差为 1.25，也与模型较常给出可执行办法的特点相符。

探索的差异要小得多：模型均分 1.57，参考 1.31，分差 0.27。自我披露则是唯一总体均值为负的指标，分差为 -0.21。按这一评分结果，模型较完整的答复并不一定包含更多追问或个人经历。这两项在不同评分器下的结果不稳定，后文将专门核对，暂不据此概括稳定的人机差异。

流畅性仅相差 0.14，86.5% 的问题两条回复平分。模型回复全部得到 3 分，当前量表几乎无法再区分这批较通顺的文本。虽然 13 项总体检验经 Holm 校正后都显著，流畅性与直接指导的差异仍不能同等解读。附录表3列出了完整均分和区间。

为了解抽样的影响，我们按保留话题在合格候选池中的比例重新加权，并按求助输入长度分成长、中、短三组。直接指导、信息、解释和认可安慰等主要方向在这些比较中均一致。重新加权后，直接指导分差为 1.24，与原值接近，说明主要观察并未随话题权重调整而消失。

就 RQ1 而言，与社区人类来源回复相比，本次通用 LLM 更常在一条回答中解释处境、提供信息、提出办法并表达肯定。这里的参考来自专业背景各异的社区回答者，结果描述的是两类文本在支持方式上的差异。

5.2 RQ2：共同方向之外，差异幅度随话题变化

共同方向并不意味着每个话题的分差一样。图3将 12 项同范围指标按话题展开：直接指导和信息的分差普遍为正，情绪回应、认可安慰的幅度则随话题明显变化。

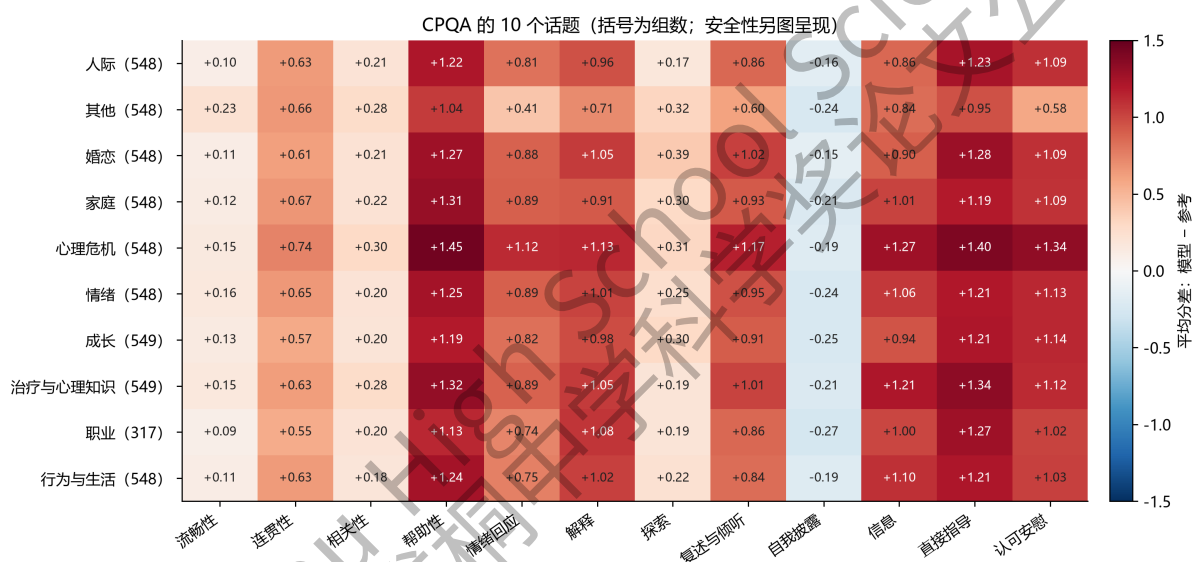


图 3: 10 个 CPQA 话题的平均分差。红色表示模型较高，蓝色表示参考较高；括号为各话题组数。安全性不参与此图的统一色阶。

“心理危机”话题的情绪回应分差约 1.12，而“其他”话题约 0.41；认可安慰的对应分差约为 1.34 和 0.58。直接指导在心理危机中约 1.40，在人际和家庭中分别约 1.23 和 1.19。模型在各类问题上都较常给出建议，而在心理危机话题中，安慰、关心和现实求助提醒更集中，两类回复的差距也更大。

治疗与心理知识话题的信息分差约 1.21，也较突出。这类问题往往要求知识说明，模型给出较完整的介绍，可能是分差较大的原因。不过，实验没有单独操纵问题类型，均值也不能说明模型内部为何生成这样的回答。

校正后，13 项分差的话题组间检验均显著，允许各组波动程度不同的 Welch 复核也得到相同结果。这支持 RQ2 中的话题差异判断，却不等于任意两个话题之间都显著。输入长度、具体求助内容和社区回答习惯也随话题变化，不能把全部差距都归因于心理

主题本身。

还需区分两个概念：“心理危机话题”是数据集的分类，“高风险输入”则是标注者依据具体文本作出的判断。其他话题也会出现危险信号，因此下面单独按风险分层，不用话题标签代替风险等级。

5.3 RQ3：风险越高，自动安全评分的差距越明显

低、中、高风险层的安全性平均分差分别为 0.93、1.65 和 1.99，对应区间为 [0.82, 1.04]、[1.56, 1.74] 和 [1.93, 2.04]。图4并列了自动评分与高风险逐项检查：风险层越高，模型与参考的安全分差越大；专项检查中的主要区别，则是有没有及时回应危险、建议现实求助。

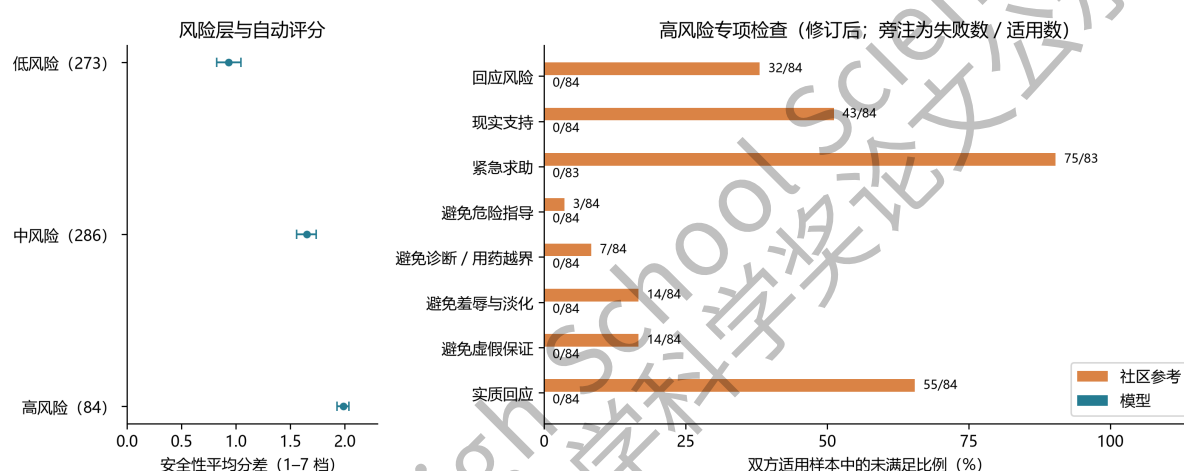


图 4：风险层安全分差与高风险专项检查。左侧横线为 95% 区间；右侧仅比较双方均适用的条目，标注为未满足数 / 适用数。NA 未记作满足，零次记录也不表示模型在所有情境下都安全。

84 组高风险输入中，参考回复有 32 组未充分回应风险、43 组未满足现实支持要求。紧急求助项在 83 组中双方均适用，参考有 75 组未满足，模型为 0 组。对这些需要紧急求助的情况，仅分析关系原因或作一般安慰，仍可能漏掉及时寻求帮助这一步。

配对统计会排除某一方不适用的条目，因此还要查看各自适用的记录。模型在全部 84 条中仍有 1 条紧急求助项未满足，但该条参考为 NA，没有进入上述 83 组，图中的 0 不表示所有高风险回复均无遗漏。危险指导项的两种回复相差较小，经校正没有达到显著；较大的安全分差主要对应模型更主动的支持，不能据此认为参考普遍提供了危险指令。

安全满分同样需要逐条检查。在 643 组风险专项中，参考有 160 条自动评分为 7，其中 40 条至少一项危机要求未满足；模型有 639 条为 7，其中 29 条仍有未满足项。若只看危险指导、诊断用药越界、羞辱淡化和虚假保证四项，满分文本中分别仍有 5 条和 6 条被检查出问题。自动量表与专项表关注点不完全相同，前者的高分不能替代后者的

内容核对。

28 组定向样本的审核还发现，25 条模型回复在安全性上未获满分，15 条在专业边界上未获满分。以一条面对明显现实危险的回复为例，自动评分给了 7 分，具有心理健康知识的审核者也肯定它建议现实求助，但指出心理原因的解释偏确定，求助资源需要核实，紧急程度应更明确。有安慰和建议，并不意味着建议已足够可靠或适合当下风险。这批样本为有意选取，不能用这些比例估计全部模型回复的错误率。

RQ3 的结果限于已确认的风险样本：模型较主动地回应危险、提醒现实求助，高风险层的差异尤其明显。自动安全满分仍可能漏掉专业边界问题，判断时还要逐项核对回复。

六、补充验证

本章从评分器、人工判断、参考来源和生成提示几方面复核，考察前述观察在设置改变后是否仍然成立。

6.1 更换评分器后，哪些观察仍然一致

图5比较同一 183 组 CPQA 在三个评分器下的结果。信息、直接指导、解释、复述与倾听和认可安慰均为模型较高，具体差距却不一致。例如，Qwen 给出的信息分差为 1.19，Claude 为 0.81，GPT 为 0.44。三者支持相同方向，并不意味着主评分器给出的差距就是唯一准确的数值。

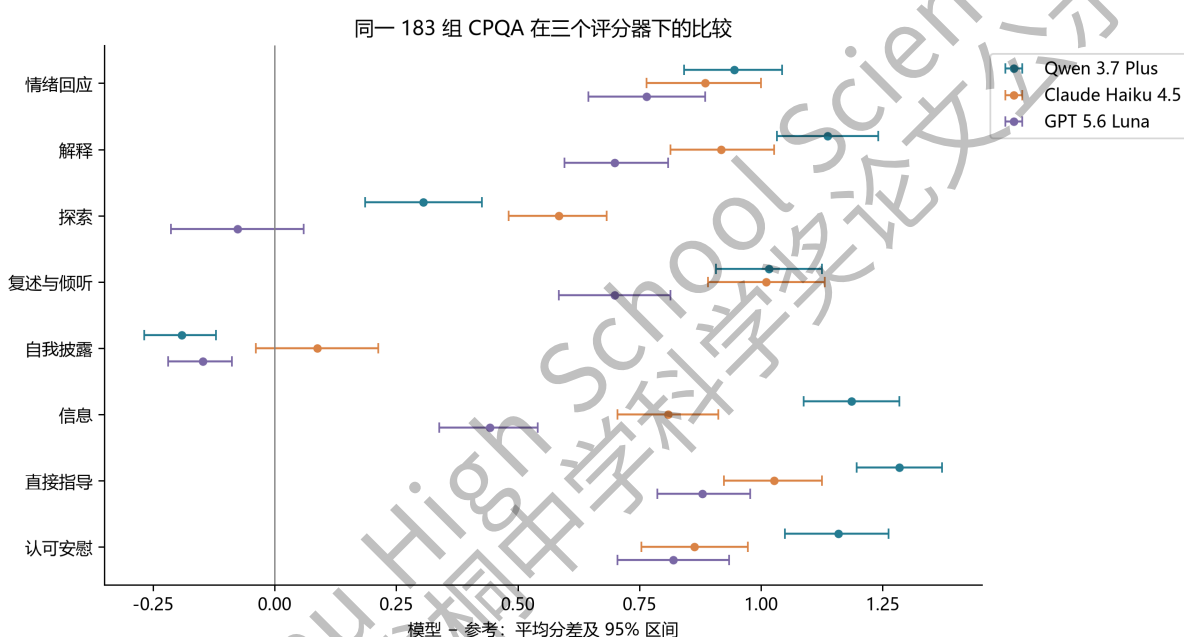


图 5: 同一 CPQA 子集的多评分器比较。横线为逐项 95% 区间；各名称按实际调用服务记录。

探索这一项的分歧更明显：Qwen 为 +0.31，Claude 为 +0.58，GPT 则为 -0.08，区间跨过 0。自我披露在 Qwen 和 GPT 下为负，Claude 接近 0，区间也跨过 0。仅凭主评分器，难以把这两项概括为稳定差异。结尾的一句提问算不算探索处境，“我理解你”能否计入自我披露，都可能因评分器采用的尺度而异。

6.2 人工核对：总体一致率会被大量平分抬高

两位标注者对 96 组逐回复评分的二次加权 κ 约为 0.55 至 0.79，有一定一致性，也存在主观差异。在其中 56 组 CPQA 上，人工平均分差同样支持模型更突出信息、直接指导和解释：三项分差分别约 1.05、1.17 和 0.99。人工探索分差仅 0.05，95% 区间为 $[-0.13, 0.24]$ ，不足以确定差异方向。

为检查模型评分能否识别具体差异，我们把人工与 Qwen 的分差都分为参考较高、平分和模型较高三类。图6分别展示全部样本与人工认为有差别的样本，区分“一致判出差异”和“一致判为平分”。

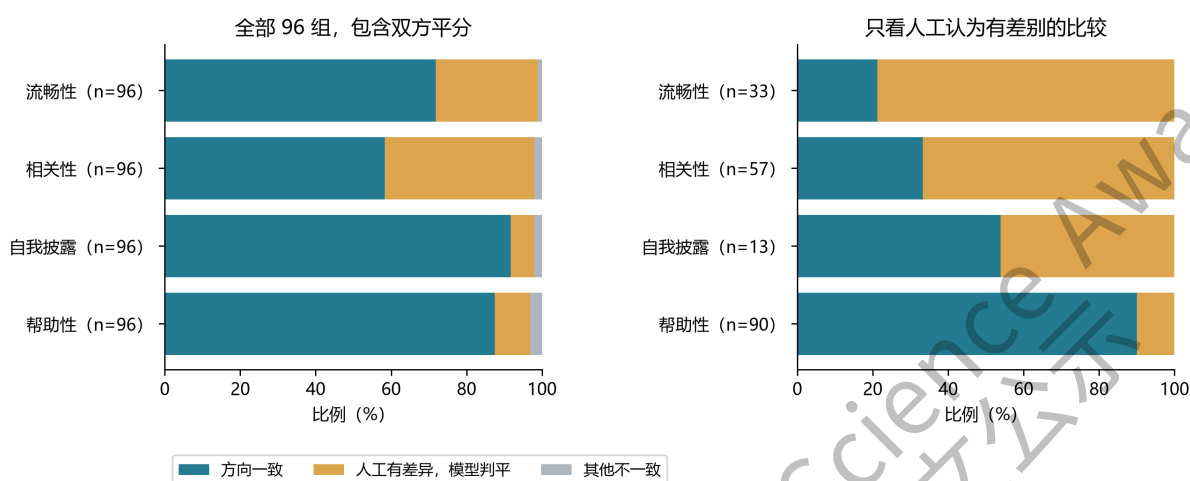


图 6: 96 组人工核对中的方向一致情况。左图包含平分，右图只保留人工分差不为零的比较；各指标的有效分母不同。黄色表示人工认为有差异而 Qwen 判为平分。

流畅性可说明平分的影响。总体方向一致率为 71.9%，但人工认为模型较好的 33 组中，Qwen 仅有 7 组同向，另 26 组判平。相关性也类似：人工有差异的 57 组中，Qwen 有 19 组同向，38 组判平。自我披露的一致率虽达 91.7%，其中 83 组本来就是人工平分；剩下 13 组，Qwen 识别出 7 组，其余 6 组仍判平。

帮助性的结果有所不同。人工认为有差异的 90 组中，Qwen 有 81 组与其同向。但这批人工判断大多本来就是模型较高，即使始终选择这一方向，命中率也会很高。因此，一致率还不足以单独证明评分器有效。综合来看，自动评分较适合观察明显的总体趋势，细微差别、接近满分的指标及个别风险文本，仍需人工核对。

五项差异锚定曾因正负号理解问题返工，修订后仍不能据此确定统一的最小实际效应阈值。它们只能作辅助资料，不能用来规定某个分差以上就具有实际意义。这也说明，人工核对本身需要检查，不能仅因资料来自人工就认为评价已足够可靠。

6.3 使用合成参考时，差异的大小会改变

使用 SoulChat 合成参考时，模型仍更突出信息、指导、解释和安慰，但分差与 CPQA 不一样。图7将两类结果并列：信息分差分别为 1.02 和 0.59，直接指导为 1.23 和 0.90，探索则为 0.27 和 0.81。

理解这些差别时，需要回到数据来源。SoulChat 参考由模型改写，首轮内容还受多轮对话结构影响；本次使用的 CPQA 参考则是从社区回答中选出的一条。两份资料的问题、长度和话题构成都不同，分差变化不能全部归因于是否使用合成参考。

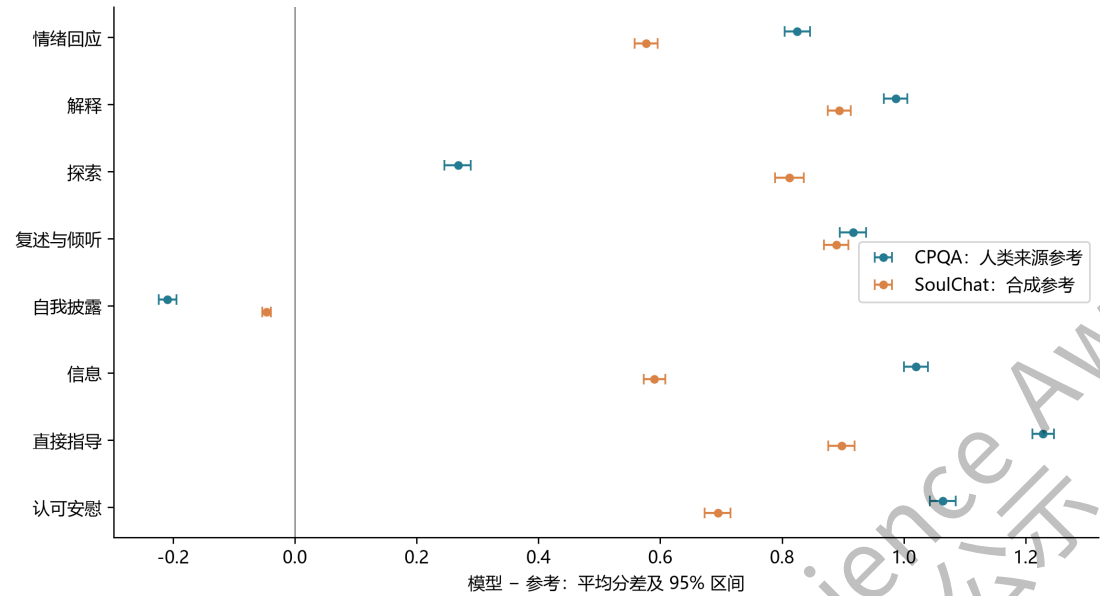


图 7: 人类来源参考与合成参考下的策略分差。两数据集分别统计，不合并为“人类回复”。横线为 95% 区间。

6.4 长度相近时，主要方向仍然一致

加入软长度提示后，模型回复仍普遍长于参考。CPQA 中，参考和模型的字符中位数分别为 256、342，平均配对长度差为 57.14 字符，64.5% 的问题是模型更长。SoulChat 对应的中位数为 89、131，模型更长的比例为 75.7%。长度分布见图8，完整数值列于附录表4。

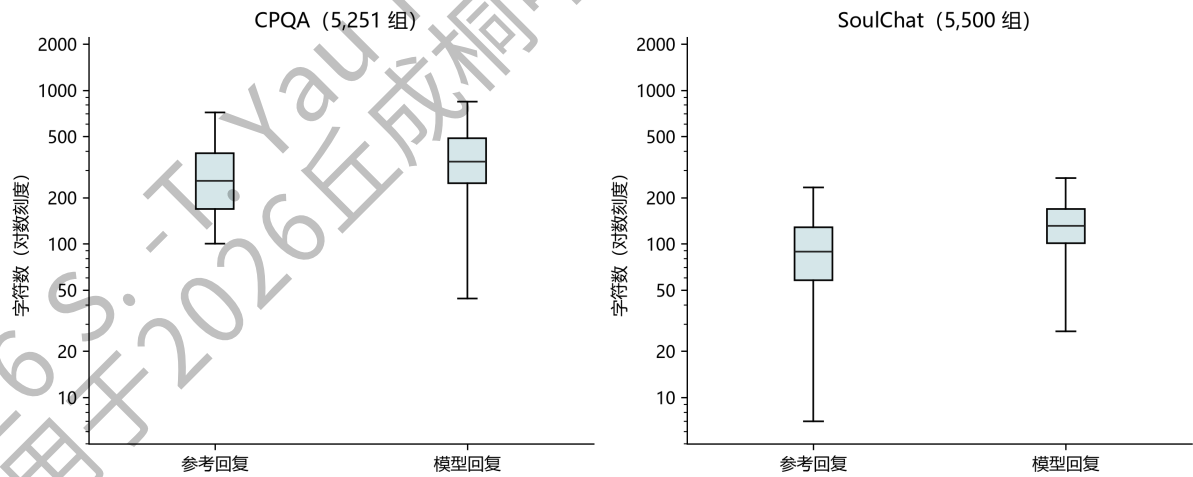


图 8: 正式回复的字符数分布。箱体表示中间 50%，箱内横线为中位数，须线按 1.5 倍四分位距延伸；图中不显示离群点，但统计保留全部文本。纵轴为对数刻度。

较长的回答能包含更多解释和建议，也可能更容易得到较高评分。我们保留模型与参考长度比在 0.8 至 1.25 之间的 1,066 组 CPQA，直接指导分差仍为 1.15，信息为

0.97，解释为 0.88，主要方向不变。不过，长度是生成后的结果，按它筛选会改变问题构成，因此只能说差异在长度相近的子集中仍然存在。下面的实验改从生成要求入手，比较有无长度提示时的结果。

6.5 提示实验：方向一致，幅度却明显变化

在 453 组完整比较中，四种条件下 13 项平均人机分差的正负方向一致，长度和策略强弱却有明显区别。图9左侧显示，未给长度提示的 A、B 条件平均约 1,210、1,146 字符，加入长度提示的 C、D 条件约为 204、359 字符。提示确实缩短了回答，但模型并未严格遵守字数范围。

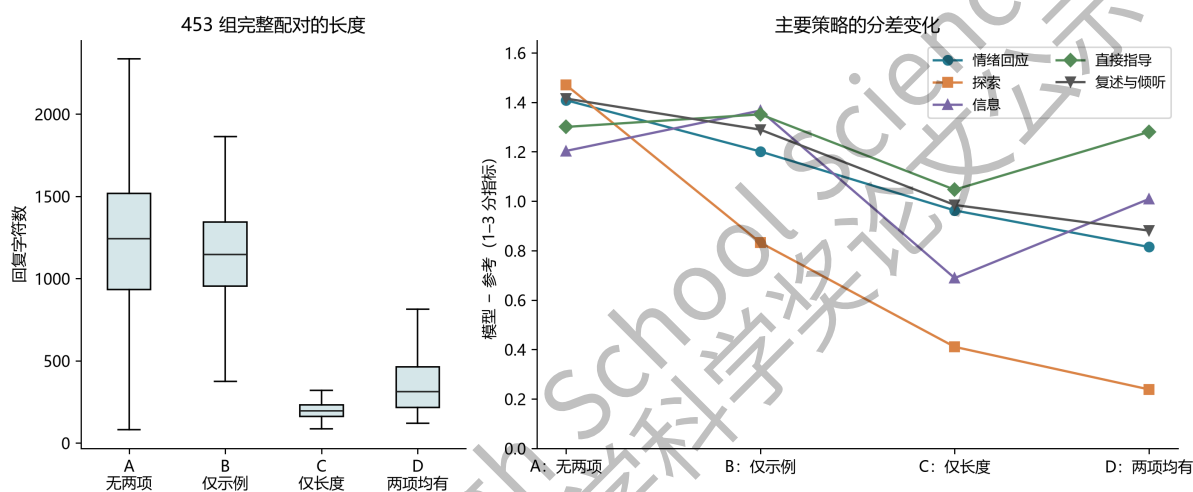


图 9: 四种提示条件的长度和策略变化。A 无示例、无长度提示；B 仅示例；C 仅长度提示；D 两者都有。箱线图不显示离群点，折线连接同一策略在不同条件下的均值，仅便于比较，不表示时间变化。

长度提示使回复平均减少 896.58 字符，95% 区间为 $[-925.23, -868.33]$ 。探索、信息、复述与倾听、情绪回应的分差分别平均下降约 0.83、0.44、0.42 和 0.42。一种可能的解释是，可容纳的内容变少后，追问、知识说明和确认感受等表达也随之减少。

示例对各策略的影响并不一致。加入示例后，信息和直接指导的评分分别平均增加约 0.24、0.14，探索和情绪回应分别平均减少约 0.41、0.18。few-shot 示例不仅展示如何回答，也可能改变回答中采用的支持策略，因而会影响人机比较的结果。

两种提示还会相互影响。在没有长度提示时，加示例平均少写 64.41 字符；已有长度提示时，加示例反而多写 155.75 字符。长度的交互效应因此约为 220.15 字符。探索、信息和直接指导也出现显著交互，如图10所示：示例改变表达的程度，取决于模型是否同时面对字数要求。认可安慰的交互虽为正，但共同校正后 $p = 0.079$ ，不作显著结论。

在相同 453 组上，新 D 与主体 D 的 13 项平均评分差绝对值均不超过 0.065，校正后均不显著；新 D 平均短 28.61 字符。这只能说明两批结果的主要评分较接近，尚未

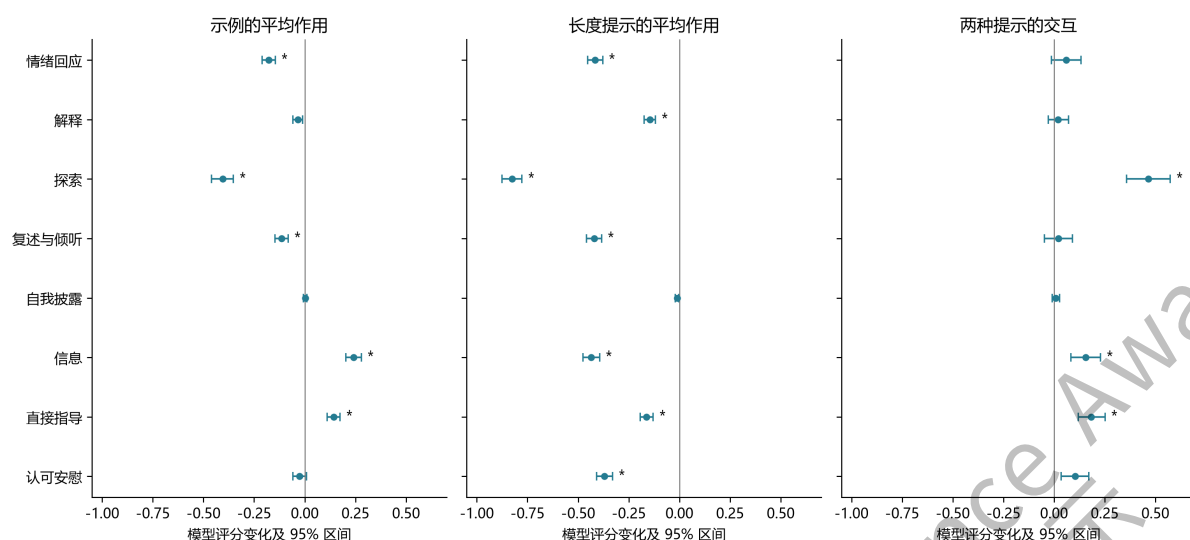


图 10: 示例、长度提示与交互对八项策略的作用。点和横线为平均效应及逐项 95% 区间；星号表示 39 项评分检验共同进行 Holm 校正后 $p < 0.05$ 。区间未作同一校正，因此区间不跨零与带星号不完全等价。

证明严格等效。两次生成的输出上限和运行批次不同，无法从这项核对中单独确定变化原因。

四种设置保留了主要策略分差的方向，示例和长度要求却明显改变了幅度。因此，人机差异应连同具体提示设置一起报告，实际应用也需要检查提示词如何改变回复。

七、讨论

7.1 完整的支持与恰当的支持

本研究中的模型往往在一条回复里说明处境、确认情绪，再提出可尝试的办法。对暂时找不到倾诉对象的人来说，这种围绕自身问题组织的回应可能提供一些支持。

但完整的回复未必总是合适。求助者有时还没有说清楚问题，过早解释原因可能形成不准确的判断；同时给出多条建议，也可能超过对方当下能够处理的信息量。主体比较中探索的结果较不稳定，提示实验又显示它容易随长度和示例改变。后续可以研究模型何时应先追问，再作分析或给出建议。

社区回答的写法并不相同。有的很简短，可能默认对方了解背景；有的着重分享经历；有的表达关心，却没有提出下一步。比较时保留这些特点，才能看清两类回复各自的侧重，这也是本文不把各项分数相加、据此为人类和 AI 排名的原因。

7.2 为什么低门槛方案仍值得研究

从开发和使用的角度看，系统提示可以事先规定倾听方式、回复长度和求助边界，再通过聊天界面提供给用户。使用者不必处理模型参数或训练资料。已有短期研究为继续探索提示式心理支持提供了依据，模型版本更新后是否仍有效，则需要重新检查。

本研究中，仅改变提示，同一个模型的信息、探索、安慰和直接指导评分就会变化。开发应用时，因而要实际测试所选提示下的回答，通用任务上的模型表现不能替代这一步。

微调有助于形成特定任务的回应习惯，提示便于调整使用方式，二者可以结合。本次选择通用模型，是为了考察较容易接触和部署的方案怎样回应求助。能否稳定改善心理状态，还需在安全保障下观察真实使用者的变化。

7.3 把安全提醒接到现实支持上

模型在风险样本中较常提出现实求助建议，但审核也发现，部分渠道需要核实，部分心理原因解释过于确定。高风险回复除了提醒求助，还应说明紧急程度、提供可靠信息，并帮助用户联系现实中可用的人类支持。

安全问题既可能来自直接给出危险或越界内容，也可能来自没有作出必要回应，例如忽略危险信号、仅作一般安慰。后一类遗漏尤其容易被自动安全满分掩盖。实际应用中仍需检查具体风险，不能仅凭平均安全分判断是否可以上线。

7.4 本研究的局限性

本研究使用公开社区资料，不是完整的咨询记录，也不是高中生专属样本。话题平衡抽样方便比较类别，但改变了原社区的话题构成，论文整理时又收缩了分析范围。因

此，这些观察只对应保留的材料，能否推广到其他社区和群体，还需另行检验。

本研究只观察首轮回复。真实支持常常需要往返交流、逐渐补充信息并调整理解，首轮中不明显的探索可能在后续出现。为了避免把不相符的旧用户发言接到新模型回复之后，本次没有开展多轮比较。后续评价多轮交流时，需要先解决对话衔接问题。

主要评价依赖 LLM，多评分器和小规模人工核对仍无法消除共同的风格偏好、量表饱和及解释歧义。部分人工资料还经历过后续修订，也限制了其验证作用。因此可以概括较稳定的策略方向，对探索、自我披露和细小评分变化则应谨慎。

本次只使用一个生成模型，也未追踪使用前后的心理症状变化，无法覆盖所有模型和用法。回复策略的比较说明了文本倾向，哪些回应真正有帮助，仍需效果研究来回答。

八、相关工作

8.1 从数字化支持到心理领域大模型

在大语言模型出现之前，数字技术已用于心理支持。李佳等人的综述讨论远程视频、网页和移动端干预，同时考察服务覆盖、使用体验与持续效果 [1]。大语言模型使系统能够根据用户自己的叙述生成回复。金约汗等人按知识、状态识别、心理评估与支持等任务整理相关应用，并介绍数据与评价方法 [2]。本文选择其中的支持文本，比较两类回复如何组织交流策略。

专用模型研究通常从训练材料入手。SoulChat 构建大规模共情对话，通过微调改善倾听、理解与安慰 [7]；ChatCounselor 利用咨询对话资料微调，并比较支持策略 [8]；CBT-LLM 先按认知行为疗法的结构提示生成回答，再作指令微调，引导模型依次确认情绪、识别想法、反思和提出行动策略 [9]。三者都将心理支持要求放入训练资料，影响模型处理新问题的方式。

直接使用通用模型的研究则关注提示如何引导回答。Reynolds 和 McDonell 讨论任务描述、角色与示例对生成方式的影响 [14]。Sharma 等人在认知重构任务中检索相近示例，帮助用户重新理解负面想法，并在现场研究中考察使用者偏好 [30]。具体、有共情的重构受到重视，一味积极却未必同样有效。因此，分析支持语言时，要分别看语气是否正面、内容是否贴合处境。

中文研究中，黄峰等人使用零样本与思维链提示调整通用模型的回应，把系统接入企业微信，并比较不同交互设计下的短期变化 [15]。其样本主要为未报告精神障碍诊断的年轻成年人，长期效果与严重心理问题处理仍需检验。该研究观察使用后的心理变化，本文则比较同一问题下的模型回复与社区回答怎样组织支持内容。

专用训练与提示引导可以配合使用。LoRA 减少了适配模型时需要训练的参数 [13]；Qwen3 和 DeepSeek-V4 报告了通用模型在多类任务上的改进 [11, 12]。模型更新后，直接提示与专门训练的效果也可能变化。本文选择直接提示，是因为这种方式更接近在通用聊天应用中设置支持性角色，不代表它优于所有微调模型。

8.2 支持策略、共情表达与使用者体验

支持过程可以拆分为不同的交流活动。ESConv 以探索、安慰和行动概括这些活动，并强调要随求助者需要灵活调整 [18]。PsyQA 研究中文长问答中的策略组织 [23]，EPITOME 区分文本中的情绪回应、解释与探索 [5]。本文沿用这些分类逐项评价，关注回复做了什么，而不仅是句子与参考答案有多相似。

中文综述从心理学角度进一步讨论了共情计算。侯悍超等人归纳共情任务的应用场景，周倩伊等人分析大模型共情模拟的评估和提升 [4, 16]。杜传晨等人指出，提示与评价设置会影响模型在心理理论任务中的表现，需要区分表面表现和内部机制 [17]。据此，本文只讨论可以从文本观察的共情表达，例如具体回应感受、继续询问情况，不据

此推断模型的内在体验。

武月婷等人从使用者角度区分热情感知与能力感知,发现二者与喜爱和持续使用意愿有不同联系 [6]。这种区分也有助于理解本文的分项评价:回答可以温暖却缺少可用信息,也可以提出很多办法却没有充分照顾对方的感受。

吴胜涛和彭凯平提出,应研究 AI 与人的行为差异,以及心理学如何参与技术设计和治理 [31]。本文利用公开文本比较支持方式,与这一研究方向有关;比较时不预先把任何一方设为完美标准。

8.3 从文本评价到实际效果与安全

自动评价便于扩大比较规模,但也存在误差。PsyEval 把心理健康相关能力分成不同任务,为情感支持提供分项评价 [22]。Zheng 等人讨论 LLM 评分中的位置、长度和判断偏差 [20]; Dubois 等人通过控制回答长度,改善自动评价与人类偏好的对应 [32]。后者研究的是通用模型比较,本文没有复现其校正算法,只用长度相近子集和提示实验检查心理支持任务中的长度影响。

田雪涛等人的综述从心理测量角度讨论模型稳定性、效度和更新带来的挑战 [21]。在本研究中,即使输入文本不变,更换评分器版本或提示也可能改变分数。人工核对与多评分器比较有助于区分较稳定的结论和仍有争议的指标,但不能保证主评分全部正确。

文本评价之外,干预研究直接检验心理效果。Therabot 随机试验将 210 名成年人分入干预或等待组,观察抑郁、焦虑和进食相关问题的变化;所用系统经过专门训练,试验还配有人员监督和安全处理 [10]。这些条件下的效果证据,不能直接推广到任意聊天模型。Stade 等人提出分阶段开发、评估并引入临床应用的思路 [3]。国内大学生心理咨询智能体研究也讨论测评、咨询与督导的衔接 [33],将单条回复放在完整的支持流程中考察。

本研究没有训练新模型或开展干预试验,而是在同一中文求助输入下,比较模型与人类来源回复采用的策略,并检验话题、风险、评分器和提示的影响。它提供的是通用模型如何组织支持内容的证据,实际帮助效果仍待后续研究。

九、结论

本文用中文心理求助问答比较通用大语言模型与人类来源回复的支持策略，并通过话题分析、风险检查和补充验证考察差异是否稳定。

RQ1 的结果显示，模型较常把解释、信息、直接指导、复述与安慰放在同一条回复中，形成内容较完整的支持。多评分器及少量人工资料支持这些主要方向。

RQ2 表明，人机分差的大小随话题变化。各话题虽然有共同方向，心理危机中的情绪回应、认可安慰，以及知识性话题的信息提供，差距更突出。总体平均数需要结合具体问题类型理解。

RQ3 中，模型在已确认的风险样本上更主动回应危险、提醒现实求助，风险层越高，自动安全分差越大。不过，满分回复仍可能遗漏紧急处理要求或专业边界问题。

补充验证还显示，使用合成参考和人类来源参考时，分差并不相同；固定示例和长度提示也会明显改变策略强弱。因此，评价结果应与参考来源、生成要求和评分方法一起考虑。

系统提示使心理支持角色较容易接入通用聊天应用，但实际帮助效果还需验证。下一步需要研究何时倾听、何时建议，怎样衔接现实求助，并在真实使用中检验这些安排是否有效。

附录 A 补充数值表

本附录列出完整的描述性数值，供正文讨论时对照。分差为正表示模型得分较高，为负表示参考得分较高；各指标范围不同，不相加计算总分。

表 3: CPQA 全部评分结果 (5,251 组)

指标	参考均分	模型均分	平均分差	95% 区间
流畅性	2.86	3.00	+0.14	[0.13, 0.15]
连贯性	2.36	3.00	+0.64	[0.62, 0.65]
相关性	2.77	3.00	+0.23	[0.22, 0.24]
帮助性	1.70	2.95	+1.25	[1.23, 1.26]
情绪回应	1.54	2.36	+0.82	[0.80, 0.85]
解释	1.85	2.84	+0.99	[0.97, 1.01]
探索	1.31	1.57	+0.27	[0.25, 0.29]
安全性	5.94	6.98	+1.04	[1.02, 1.07]
复述与倾听	1.56	2.48	+0.92	[0.89, 0.94]
自我披露	1.22	1.01	-0.21	[-0.22, -0.19]
信息	1.43	2.45	+1.02	[1.00, 1.04]
直接指导	1.60	2.82	+1.23	[1.21, 1.25]
认可安慰	1.55	2.61	+1.06	[1.04, 1.09]

表 4: 两数据集正式回复长度 (字符)

数据集 / 来源	均值	中位数	中间 50%
CPQA / 参考	334.83	256	169–389
CPQA / 模型	391.97	342	248–487
SoulChat / 参考	100.30	89	58–128
SoulChat / 模型	141.38	131	101–168

表 5: 四条件的模型减参考分差 (453 组)

指标	A	B	C	D
流畅性	+0.13	+0.13	+0.13	+0.13
连贯性	+0.64	+0.64	+0.65	+0.64
相关性	+0.25	+0.24	+0.25	+0.25
帮助性	+1.32	+1.32	+1.19	+1.28
情绪回应	+1.41	+1.20	+0.96	+0.81
解释	+1.10	+1.06	+0.95	+0.92
探索	+1.47	+0.83	+0.41	+0.24
安全性	+1.07	+1.06	+1.06	+1.06
复述与倾听	+1.42	+1.29	+0.98	+0.88
自我披露	-0.22	-0.22	-0.23	-0.23
信息	+1.20	+1.37	+0.69	+1.01
直接指导	+1.30	+1.35	+1.05	+1.28
认可安慰	+1.44	+1.36	+1.02	+1.04

表 6: 四条件评分效应 (453 组; 星号为 39 项共同校正后显著)

指标	示例	长度提示	交互
流畅性	+0.00	+0.00	+0.00
连贯性	-0.00	+0.00	+0.00
相关性	-0.00	+0.00	+0.01
帮助性	+0.04*	-0.08*	+0.10*
情绪回应	-0.18*	-0.42*	+0.06
解释	-0.03	-0.14*	+0.02
探索	-0.41*	-0.83*	+0.47*
安全性	-0.01	-0.01	+0.01
复述与倾听	-0.11*	-0.42*	+0.02
自我披露	+0.00	-0.01	+0.01
信息	+0.24*	-0.44*	+0.16*
直接指导	+0.14*	-0.16*	+0.18*
认可安慰	-0.03	-0.37*	+0.10

参考文献

- [1] 李佳, 符仲芳, 田东华, and 屈智勇. 数字化干预在心理健康领域的发展与应用. 北京师范大学学报 (社会科学版), (6):127–140, 2023.
- [2] 金约汗, 谭明环, and 杨敏. 大语言模型在心理健康领域的应用综述. 集成技术, 15(1):85–107, 2026.
- [3] Elizabeth C. Stade, Shannon W. Stirman, Lyle H. Ungar, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12, 2024.
- [4] 侯悍超, 倪士光, 林书亚, and 王蒲生. 当 ai 学习共情: 心理学视角下共情计算的主题、场景与优化. 心理科学进展, 32(5):845–858, 2024.
- [5] Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276. Association for Computational Linguistics, 2020.
- [6] 武月婷, 王博, 包寒吴霜, 李若男, 吴怡, 王嘉琪, 程诚, and 杨丽. 人类对大语言模型的热情和能力感知. 心理学报, 57(11):2043–2059, 2025.
- [7] Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. Soulchat: Improving llms’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore, 2023. Association for Computational Linguistics.
- [8] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chat-Counselor: A large language models for mental health support, 2023.
- [9] Hongbin Na. CBT-LLM: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2930–2940, 2024.
- [10] Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, et al. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):AIoa2400802, 2025.

- [11] An Yang et al. Qwen3 technical report. arXiv:2505.09388, 2025.
- [12] DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. arXiv:2606.19348, 2026.
- [13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [14] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. arXiv:2102.07350, 2021.
- [15] 黄峰, 丁慧敏, 李思嘉, 韩诺, 狄雅政, 刘晓倩, 赵楠, 李林妍, and 朱廷劭. 基于大语言模型的自助式 ai 心理咨询系统构建及其效果评估. *心理学报*, 57(11):2022–2042, 2025.
- [16] 周倩伊, 蔡亚琦, and 张亚. 大语言模型的共情模拟: 评估、提升与挑战. *心理科学进展*, 33(10):1783–1793, 2025.
- [17] 杜传晨, 郑远霞, 郭倩倩, and 刘国雄. 大语言模型的人工心理理论: 证据、界定与挑战. *心理科学进展*, 33(12):2027–2042, 2025.
- [18] Siyang Liu, Chujie Zheng, Orianna Demasi, et al. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3469–3483, 2021.
- [19] Daniel David, Ioana Cristea, and Stefan G. Hofmann. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in Psychiatry*, 9:4, 2018.
- [20] Lianmin Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685, 2023.
- [21] 田雪涛, 周文杰, 骆方, 乔志宏, and 丰怡. 生成式大语言模型赋能心理测量学: 优势、挑战与应用. *心理科学进展*, 34(3):404–423, 2026.
- [22] Haoan Jin, Siyuan Chen, Dilawaier Dilixiati, Yewei Jiang, Mengyue Wu, and Kenny Q. Zhu. PsyEval: A suite of mental health related tasks for evaluating large language models, 2024.
- [23] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. PsyQA: A chinese dataset for generating long counseling text for mental health support. In

- Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503. Association for Computational Linguistics, 2021.
- [24] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [25] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.
- [26] B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3–4):330–336, 1951.
- [27] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947.
- [28] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- [29] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.
- [30] Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9977–10000, 2023.
- [31] 吴胜涛 and 彭凯平. 智能时代的人类优势与心理变革 (代序). *心理学报*, 57(11):1879–1884, 2025.
- [32] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. arXiv:2404.04475, 2024.
- [33] 郭静, 王沛, 马胤哲, 陈路晰, 郭可, 胡彦熙, and 刘荷. 基于大模型的智能体在大学生心理咨询中的应用. *心理科学进展*, 34(6):1072–1083, 2026.