arXiv:2601.01407v1 [cs.CL] 4 Jan 2026

从情绪分类到情绪推理：提升大语言模型的情商

ArjhunSreedar[1] Rohan Pillay[1] Laukik Patade[1]马萨诸塞大学阿默斯特分校，美国马萨诸塞州阿默斯特{asreedar, rpillay, lpatade}@umass.edu

摘要

本研究探讨合成情绪思维链数据能否提升较小型开放式大语言模型（LLM）的情绪推理能力。我们设计了一种多智能体生成流程，用于生成治疗式对话，并将其转换为带有解释的结构化情绪多项选择题（MCQ）。我们提出，在该数据集上对多种7B模型进行微调，应能显著提升它们在EmoBench 风格评测中的情绪理解和情感意识，这表明无需进行架构变更也能诱发情绪推理能力。结果表明，微调后的Mistral 7B 在情绪理解（EU）方面的得分从10.5提升至20.5，在情绪分析（EA）方面的得分从40.5提升至60.0，验证了合成情绪推理数据在增强模型处理细腻情绪任务的能力方面的有效性。

1 引言

大语言模型在许多语言理解任务中表现出色，但往往将情绪理解视为简单的标签预测问题，忽视了一个人在特定情境下产生某种感受的原因（Sabour 等人，2024年）。这一局限阻碍了共情式对话代理、虚拟心理健康助手和情绪自适应辅导工具等应用的发展，而这些应用都需要稳健、类人的情绪推理（Rashkin 等人，2019年）。

传统的情绪识别方法侧重于分类准确率——预测一个人是感到“快乐”“悲伤”还是“愤怒”——却不要求模型解释背后的原因、预判后果或考虑其他视角。这种浅层处理无法捕捉人类情绪体验的复杂性，因为情绪源于个人经历、情境背景、社会动态和认知评估之间错综复杂的相互作用。例如，一段长期关系结束后，一个人可能同时感到如释重负和内疚。要理解这种混合情绪状态，需要推理相互冲突的价值观、社交期望和预期结果，而不是简单地从预定义分类体系中选取一个标签。

本研究旨在利用合成情绪选择题数据集，开发并微调规模较小的开放式大语言模型，使模型能够通过情绪思维链推断、解释和推理人类情绪，而不只是预测标签。我们通过生成结构化推理轨迹，将情境线索与情绪结果经由中间步骤联系起来，例如识别相关的个人信念、辨析依赖视角的解读，以及预测情绪后果，从而旨在培养一种更接近人类理解方式的情商。这种方法对构建人工智能系统具有实际意义：这类系统能够参与情绪敏感型互动、提供个性化心理健康支持，并根据不同用户的情绪状态和需求调整沟通策略。

本文的主要贡献有三点：

- **情绪推理合成数据**。我们创建了一个受治疗启发的合成数据集，其中包含 3,000+ 个涵盖情绪理解和情绪分析类别的情绪推理选择题实例。
- **多智能体生成流程**。我们实现了两种生成流水线（一种带有质量监督的多智能体对话系统，以及一种简洁的流水线），并将对话转换为带有解释的结构化选择题。
- **7B 微调评估**。我们使用低秩适配对多个 7B 规模的模型进行微调，并在 EmoBench 风格的情绪理解/情绪分析评估中取得了稳定的提升，其中 Mistral-7B 的改进尤为显著。

这项工作关系的代码可在 GitHub 上获取[1]。

2 相关工作

Mohammad 等人（2018年）推出了 SemEval-2018 推文中的情感任务，提供了带有情感和情感极性标注的大规模 Twitter 数据，推动了多标签情绪分类的发展，但并未明确以情绪因果推理为目标。Strapparava 和 Mihalcea（2007）发布了 Affective Text 数据集，用于为新闻标题分配情绪分数，凸显了从简短且缺乏上下文的片段中识别情感的难度，同时仍将该任务设定为标签预测。

为推动对话共情的发展，Rashkin 等人（2019年）提出了 EMPATHETIC-DIALOGUES，这是一个以情绪情境为基础的多轮对话数据集，可用于训练共情回应模型，但不包含明确的推理链。该数据集中的每段对话都由说话者描述个人情绪情境开始，随后倾听者通过共情话语回应，以承认并肯定说话者的感受。尽管该数据集已被证明有助于训练模型生成符合上下文的共情回应，但它并未要求模型阐述将情境细节与情绪状态联系起来的推理步骤，也未要求模型解释为何某些回应比其他回应更具共情性。

Poria 等人 (2019年) 引入了 MELD，这是一个包含情绪和情感标签的多模态、多方对话数据集。该数据集强调说话者之间的互动和多模态线索，但同样侧重于识别，而非逐步解释。MELD 在《老友记》电视剧语料库的基础上扩展而来，为多方对话中的每条话语添加了详细的情绪标注，捕捉情绪如何在社会互动中变化，以及不同说话者如何影响彼此的情绪状态。然而，与早期数据集一样，MELD 将情绪视为一种类别标签，根据上下文和多模态特征进行预测，而不是将其视为模型应明确阐述的推理过程所产生的结果。

在建模方面，Devlin 等人（2019年）表明，BERT 等大型双向 Transformer 能够针对情绪和情感极性任务进行有效微调，建立了强大的基线模型。这些模型依赖强大的上下文表征，但缺乏明确的推理监督。BERT 通过掩码语言建模预训练捕捉双向上下文的能力，使其能够学习丰富的情绪语言表征，但该模型仍然是一个黑箱，只能生成预测结果，却无法解释其推理过程。Liu 等人（2019年）通过 RoBERTa 改进了这一方向，展示了更多数据和优化后的预训练对下游情感任务的益处。凭借更完善的训练流程和更大规模的预训练语料库，RoBERTa 在多个情绪分类基准测试中取得了最先进的结果。

Wei 等人 (2023年) 证明，思维链提示能够引导大型大语言模型进行逐步推理，这表明推理可以通过提示而非架构变更涌现。他们的研究表明，在少样本示例中加入中间推理步骤后，大语言模型能够解决使用标准提示时无法解决的复杂多步问题。这一发现开启了新的探索方向：情绪推理——根据某人的处境、信念和社会关系理解其产生某种感受的原因——是否也能通过结构化提示和监督来引导或增强。

在此基础上，Li 等人（2024年）提出了情感思维链，表明结构化的情感解释能够提升共情能力和情感生成质量。他们的方法会生成中间推理步骤，将情境线索与情绪结果联系起来，并明确建模情感反应背后的认知评估过程。通过训练模型生成这些推理轨迹，他们证明了生成的情感反应在准确性和感知共情能力方面均有所提升。

由Sabour 等人（2024年）提出的 EmoBench 用于评估大语言模型在情感推理任务上的表现，并揭示许多模型在因果关系和基于视角的情感理解方面仍有困难。与以往侧重分类的情感数据集不同，EmoBench 要求模型识别复杂情境中的情绪、解释情感原因、预测情感后果，并推理同一情境中的不同人如何基于各自的视角和背景产生不同感受。对前沿模型的评估显示，其情绪推理能力存在显著不足，这也凸显了对针对性训练数据和方法的需求。

最后，Hendrycks 等人 （2023年）；Zhou 等人

[1] https://github.com/kernelism/EC2ER

（2024年）使用合成数据，引导模型趋向于人类价值观和结构化推理，从而促成了我们的合成情感推理数据集。这些工作表明，经过精心设计的合成数据在教导模型以结构化方式推理复杂现象方面，可以达到与自然产生的数据同等的效果，甚至更胜一筹。我们的工作将这一合成数据范式扩展到情绪推理领域，生成治疗式对话，并提取结构化情绪推理任务，为语言模型中的情商提供训练信号和评估指标。

3 方法论

3.1 任务与数据源

我们的工作以 EmoBench 为基础。EmoBench 是一个人工构建的基准测试，用于评估大语言模型在多个以推理为核心的任务中的情商（Sabour 等人，2024年）。EmoBench 包含多种简短情境，并配有相应问题，要求模型推断情绪、解释原因与反应，并对不同视角进行推理，而非仅仅预测标签。该基准测试主要分为两个维度：情绪理解（EU）和情绪分析（EA）。情绪理解任务评估模型能否正确推断某人的感受及其原因，涵盖多个子类别，例如识别模糊情境中的复杂情绪、察觉较简单的模型可能忽略的细微情绪线索、基于影响情感反应的个人信念与经历进行推理，以及站在可能对同一事件做出不同反应的不同个体的角度思考。情绪分析任务则评估模型能否恰当地回应情绪情境，包括选择能够肯定感受的支持性回应、在情绪激烈的情境中采取有效行动、提出情绪调节策略、缓和冲突，以及设定恰当的界限。

在我们的实验中，我们主要使用多项选择题（MCQ）划分，其中包含数千个实例，涵盖广泛的基础情绪和复杂情绪。每道题都提供一个情境、一小组选项，以及一个适合进行定量评估的标准答案标签。每道多项选择题都由一段简短的情境描述、一道考查情绪推断或情绪觉察的问题，以及四个候选答案组成模型必须从中选出最合适的选项。这种形式既便于自动评估，又要求模型围绕情感动态进行多步推理。

EmoBench 既是我们的基础数据集，也是留出评估集：我们保留其原始测试样本用于基准测试，并设计了合成的治疗风格数据，使其结构与 EmoBench 一致，同时扩展对复杂情绪、细微情绪线索、个人信念与经历以及观点采择等方面的覆盖。通过保持与 EmoBench 结构兼容，我们确保使用合成数据微调的模型能够直接在成熟的情绪推理基准测试上接受评估，从而可以与基线模型和替代方法进行公平比较。

3.2 合成数据生成流水线

我们的数据集由合成的治疗风格情境组成，其中患者 角色设定与治疗师 代理互动，并由编排器 引导，确保内容与角色设定的背景和情绪主题保持一致。提取器代理将这些对话概括为情绪多项选择题，每题包含一个情境、一组候选情绪标签、正确选项，以及一段采用 JSON 格式的情感思维链解释。

我们实现了两条互补的生成流水线，以生成多样化情感推理数据：

用于情绪理解/情绪行动生成的多智能体对话系统（MADS）。MADS流水线（算法1）通过多个专业化智能体之间的协同交互，生成情境丰富的治疗对话。系统首先从包含200万个人物画像的集合中加载人物画像Ge 等人（2025年），每个人物画像都说明了人口统计信息、性格特征、生活状况和情绪倾向。随后，系统会加载相匹配的情感主题（如悲伤、职场压力、家庭冲突或身份探索），并抽取指定数量的人物画像—主题配对，以确保情感类别和人际关系情境的多样性。

对于每个采样对，BackgroundGenera-torAgent都会创建一份详细叙事，将即将展开的对话置于角色设定的生活经历、当前处境以及具体情绪主题之中。这份背景叙事通常

长达 200—400 个词，并包含有关人际关系、近期事件、持续性压力源以及角色设定典型应对模式的相关细节。这份叙事为 ClientAgent（代表患者）和 TherapistAgent 提供共同背景，确保对话连贯且情感真实。

多轮对话随后展开，由 ClientAgent 发起会谈，描述自己的情境；接着，TherapistAgent 通过共情倾听、反思性提问或温和的质疑作出回应，旨在深化情绪探索。SupervisorAgent 每两轮监控一次对话质量，检查对话是否充分确立了清晰的情境、表达了真实的情绪、探索了因果因素，并呈现了真正的困境或决策点。对话至少持续 4 轮，最多持续 14 轮；当所有质量标准均已满足或达到轮数上限时，对话即停止。

生成高质量对话后，专门的提取代理会处理对话，生成多个 EU条目和 EA条目。对于 EU条目，提取代理会识别对话中 3-4 个时刻，在这些时刻，客户的情绪状态可通过语境线索、个人信念或依赖视角的解读推断出来；随后编写选择题，测试模型能否准确识别情绪并解释情绪产生的原因。对于 EA条目，提取代理会识别 3-4 个决策点，在这些节点上，客户或治疗师必须选择如何回应、调节情绪或处理人际互动，并据此编写问题，测试模型能否结合情绪语境选择恰当的行动。

所有提取的条目均写入 JSONL 文件，并附有详细元数据，包括角色 ID、主题、对话长度和类别分布。系统每 50 段对话保存一次检查点，以便在发生中断时恢复，并跟踪类别分布，确保情绪理解和情绪分析各子类别的覆盖均衡。最终输出包括分别存储情绪理解条目、情绪分析条目的 JSONL 文件，以及用于分析和消融研究的综合元数据。

简洁的单轮对话流水线。简洁的流水线（算法2）提供了一种更精简的方法，无需耗费额外资源进行多轮对话，即可生成聚焦的情绪情境。该流水线一次性抽取人物设定及其相关属性，并

算法 1 多智能体对话系统（MADS）情绪理解/情绪行动生成

输入： 角色设定 文件，主题 文件，$num\ dialogues$，大语言模型配置
2：加载人物设定和主题；检查兼容性；采样性；采样 $num\ dialogues$（角色设定、主题）对
对于每个（角色设定、主题）对**执行**
  通过 Back- 生成背景叙述 groundGeneratorAgent
  初始化 ClientAgent、TherapistAgent、SupervisorAgent
  运行多轮对话（最少 4 轮，最多 14 轮）
  SUPERVISOR 每 2 回合
  满足所有标准后停止（情境、情绪、原因、困境）
  提取多个EU条目（每段对话3–4个）通过提取代理
  提取多个EA条目（每段对话3–4个）通过提取代理
  将所有EU/EA条目写入JSONL；跟踪类别分布
  每 50 条对话保存一次检查点
循环结束
输出： EU条目（JSONL），EA条目（JSONL）、元数据（JSONL）、统计数据

生成精简的背景叙述，进行一段较短的 5 轮治疗师与患者对话，然后立即提取 EU/EA条目，而无需单独进行督导评估。

这条流水线的主要优势在于效率：通过生成更短的对话（最多 10 条消息，而 MADS 中最多为 28 条），并在生成过程中直接提取内容、无需进行质量检查点保存，精简版流水线能够以单位计算资源生成更多情境。这使其特别适合快速扩大数据集规模、探索新的属性组合，或针对不同生成参数开展消融研究。

利害程度（低、中、高）、情绪混合类型（愤怒与受伤、宽慰与内疚、嫉妒与钦佩、爱与怨恨、羞耻与反抗、担忧与兴奋）、洞察力水平（低、中、高）、应对方式（回避、问题解决、宣泄、讨好他人、理智化、自责）、沟通方式

（退缩的、歉疚的、讽刺的、直率的、回避冲突的），以及视角类型（自我视角、他人视角、中立第三方）。

这些属性通过概率采样生成，以确保覆盖范围广泛，同时保持组合的现实性。例如，洞察力较低且采用回避型应对风格的角色设定，更可能表现出退缩或转移话题的沟通方式；而洞察力较高且具有解决问题倾向的角色设定，则可能展现出更直接、更具分析性的沟通方式。负责生成背景叙述和对话的大语言模型会收到提示，自然地融入这些属性，营造真实可信而非刻意编造的情境。

生成对话后，流水线会立即创建情绪理解和情绪分析选择题，验证其 JSON 格式，并将其写入 JSONL 输出文件。每处理 10 个角色设定就进行一次检查点保存，以平衡崩溃恢复需求与输入/输出开销。最终输出包括 personas_final.json，其中包含所有采样得到的角色设定及其属性；以及 items_final.jsonl，其中包含所有生成的选择题条目及其解释、正确答案和元数据。

角色设定与属性多样性。为确保广泛覆盖各种情绪情境，我们精心设计了人物设定采样策略和属性空间。人物设定涵盖多种背景，包括年龄范围（18–75 岁）、职业（学生、医护人员、教师、工程师、艺术家、退休人员等）、关系状态（单身、有伴侣、已婚、离异）、文化背景（代表主要世界地区），以及基线情绪倾向（焦虑、乐观、多疑、富有同理心、谨慎、善于表达）。

对以下属性进行采样，为每段对话添加多样化的个人经历和丰富的情境，从而生成情感框架更佳的多项选择题：

- **情绪理解领域**：情绪识别、原因推理、观点采择、混合矛盾情绪、预期情感后果
- **情绪应用领域**：支持性回应、行动选择、情绪调节策略、冲突降级、设定界限
- **关系类型**：恋爱关系、家庭关系、朋友关系、工作关系、社交关系
- **问题聚焦类型**：自我、他人
- **冲突领域**：职业表现、金钱与财务、照护负担、友谊疏远、浪漫关系中的信任与嫉妒、姻亲家庭矛盾、网络沟通误解、跨文化误解
- **利害程度**：低、中、高
- **情绪组合类型**：愤怒与受伤、宽慰与内疚、嫉妒与钦佩、爱与怨恨、羞耻与反抗、担忧与兴奋
- **洞察程度**：低、中、高
- **应对方式**：回避、过度纠结于解决问题、宣泄、讨好他人、理智化、自我责备
- **沟通方式**：含蓄的间接表达，歉疚而柔和，带有——sar-

Alg **算法 2** 简明情绪理解/情绪分析合成数据集 y数据集

第1代：**输入：** 人物档案，$num\ samples$，大语言模型配置 2：从 JSONL 加载人物设定；抽样 n =min($num\ samples$, $|\mathcal{P}|$) 个人物设定 3：**对于**每个抽样到的人物设定 p **执行**4：抽样属性：欧盟/欧洲经济区领域、关系类型、冲突领域、情绪组合、应对方式等。5：生成背景：大语言模型根据人物设定 + 属性创建叙事 6：开展对话：治疗师—患者多轮对话（5 轮，最多 10 条消息）7：生成选择题条目：大语言模型根据背景+ 对话创建情绪理解/情绪分析选择题 8：验证格式：检查 JSON 模式、必填字段和允许使用的情绪标签 9：检查点：每 10 个人物设定，保存中间人物设定（JSON）和条目（JSONL）

10：**结束循环**

11：保存最终输出：personas final.json、items final.jsonl 12：**返回：** 摘要（路径、数量、类别分布）

讽刺性转移话题、直率且爱批评、回避冲突且顺从 –

- **视角类型**：自我视角、他人视角、中立第三方 – – – –

这一丰富的属性空间能够生成涵盖人类情感体验全部复杂程度的情境：从简单明了的单一情绪情境到模糊的混合情绪困境，从低风险的社交尴尬到高风险的关系危机，也从界限分明的视角采择任务到多种解读皆有可能的微妙情形。

所有数据均通过 GPT-OSS-20B 提示和结构化模板生成，而非从真实用户处收集。这样既能控制情绪分布、角色设定变化和文化中立性，又能避免隐私问题。使用合成数据还使我们能够生成自然发生的治疗性对话中可能鲜少出现的情境，例如现有数据集中代表性不足的复杂情绪、文化背景或人际互动的特定组合。

3.3 数据预处理

预处理会将原始多轮对话转换为精简的情境描述和选择题记录。该流水线会移除说话者重复发言且未补充新信息的冗余轮次，将说话者标签规范为统一标识符（CLIENT、THERAPIST、SUPERVISOR），并通过检查必需的 JSON 字段（情境、问题、选项、正确答案、解释）以及核实所有情绪标签均属于预定义分类体系，来确保数据符合架构要求。

其他步骤包括使用设定莱文斯坦距离阈值的模糊字符串匹配，对近乎相同的情境进行去重，以确保每个训练样例都提供独特的信息。我们还采用简单的平衡启发式方法，避免情绪或人物设定过度偏斜：如果某个类别在生成过程中占比过高（超过预期比例的 1.5 倍），就会拒绝并重新生成后续样本，直至恢复平衡。

对于带有自相矛盾的解释（即给出的推理与所选答案相矛盾）、JSON 结构无效（缺少字段、数组格式错误、引号未转义）或情绪标签超出词汇表范围的实例，我们会将其过滤掉，以提高监督质量。此外，我们还会进行基本的语言质量检查，例如确保情境长度在 50 到 500 个词之间、问题语法正确且含义明确，以及干扰项（错误答案选项）具有合理性，但根据情境内容又明显不正确。

完成所有预处理步骤后，最终数据集包含 1,527 个 EU 条目，涵盖复杂情绪（ 382 ）、情绪线索（385）、个人信念与经历（378）以及观点采择（382）；此外还包含 1,486 个 EA条目，涵盖个人-他人（371）、个人-自我（372）、社会-他人（372）和社会-自我（371）。各子类别的分布几乎均衡，细微差异源于自然抽样波动和质量筛选。

3.4 数据标注

标注过程融入生成环节：提示词和模板会指示大语言模型在单个结构化JSON对象中同时输出情境、正确的情绪标签和解释。这种方法从设计上确保了标注的一致性，因为生成情境的同一个模型也会根据对情境的理解，确定情绪标签并生成推理轨迹。

作者对其中50个示例进行了人工检查，以验证标签和推理是否合理，以及解释是否连贯并以情境为依据。在检查过程中，我们发现并纠正了若干系统性问题：（1）解释援引了情境中不存在的信息；（2）所选情绪标签含义模糊，或与某个干扰选项同样相符；（3）推理链跳过了必要的推理步骤。此次人工审查的反馈用于优化提示词和筛选规则，从而提升整体数据质量。

我们通过让两位标注者独立判断每个条目的标签和解释是否符合相应情境，计算了经审核子集的标注者间一致性。结果显示一致性较高（科恩卡帕系数 $=0.78$），大多数分歧都源于真正存在歧义的案例，在这些案例中，多种情绪都有合理依据。我们将这些模糊条目连同置信度分数一并保留在数据集中，从而更真实地反映情绪推理难度的分布。

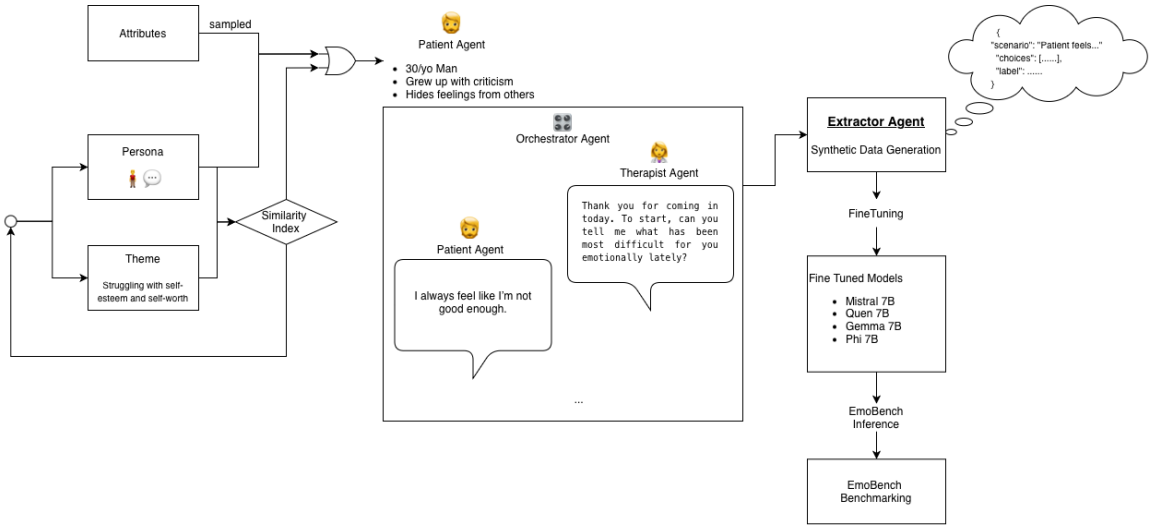

图 1：用于合成情绪推理数据生成、模型微调和 EmoBench 风格评估的整体方法流水线。该流水线从角色设定和主题选择开始，经过带有质量监督的多智能体对话生成，提取结构化的 EU/EA条目，使用参数高效微调对模型进行微调，并通过留出的 EmoBench指标进行评估。

4 基线方法

基线对应的是模型在原始 EmoBench 数据集上的表现，代表各模型在使用我们的合成情绪推理数据进行任何微调之前所达到的准确率。

5 我们的方法

我们的方法将情商重新定义为一个监督式推理问题：给定角色设定和情境，模型必须推断原因及后果，以预测情绪并生成解释。我们不再将情绪识别视为简单的分类任务，而是将其构建为思维链生成任务，要求模型阐述从情境线索推导出情绪结果的推理步骤。

与 EmoBench 数据的比较。
EmoBench 包含由领域专家手动筛选并标注的评测数据，质量高但规模有限（约 400 个测试条目）。相比之下，我们的流水线通过合成方式生成数据，旨在涵盖情绪推理的不同方面，包括复杂情绪（涉及混合或模糊情绪状态的情境）、情绪线索（需要推断的细微情绪迹象）、个人信念和经历（情绪反应取决于个人背景的情况），以及观点采择（不同的人在某些情境下会有不同的感受）。这样，我们便能系统地改变情境结构和角色属性，同时保持与 EmoBench 风格的推理任务兼容。

通过生成数千个具有多样情绪特征的训练样例，我们可以探究情绪推理能否跨类别泛化，以及合成数据能否为模型学习稳健的情绪推断模式提供充足信号。我们的假设是，让模型接触多样的情境及明确的推理链，将教会模型把情绪推理分解为若干组成技能（识别相关线索、调用个人背景、预判后果、考虑其他可能性），并使这些技能迁移到新情境中。

使用 LoRA 进行参数高效微调。我们通过低秩适配（LoRA）实现参数高效微调（PEFT），对 Mistral 7B 、 Qwen2.5-7B-Instruct 和 Gemma-7B-IT 等较小的开源大语言模型进行微调。LoRA 在 Transformer 注意力层中引入可训练的低秩矩阵，使我们只需使用全量微调所需参数的 0.1%–1% 即可调整模型。这种方法具有多项优势：（1）降低内存需求，从而能够在消费级 GPU 上进行训练；（2）由于参数空间更小，收敛速度更快；（3）大多数参数保持冻结，因此能够保留基础模型的能力；（4）通过适配器模块实现便捷部署，这些模块可在推理时进行切换

。

训练目标和指令格式。模型通过指令式提示进行训练，这些提示会呈现情境、提出一个具体问题，并要求模型输出一个 JSON 对象，其中包含预测情绪和情绪思维链解释。指令格式遵循以下模板：

<情境>[情境文本]</情境>

<问题>[探究情绪理解或情绪分析的问题]</问题>

<选项>
A) [第一个选项] B) [第二个选项] C) [第三个选项] D) [第四个选项]</选项>

生成一个 JSON 响应，包含：{ "reasoning": "<推理>", "answer": "A/B/C/D" }

模型首先经过训练，生成一段推理轨迹，说明相关的情境线索；在适用时，结合角色背景或文化背景；考虑其他解释；并通过将证据与情绪结果联系起来，为所选答案提供依据。这种结构化输出格式确保模型学习的不只是预测情绪，还要阐述人类根据上下文推断情绪时所采用的推理过程。

训练损失是针对生成的 JSON 序列采用标准的因果语言建模目标，训练过程中使用教师强制。我们对 JSON 响应中的所有 token 采用统一的损失权重。实现使用 Hugging Face Transformers 和 PyTorch，实验在通过 Google Colab Pro+和大学计算集群提供的 NVIDIA A100 GPU （ 40GB 显存）上运行。

6 结果

我们评估了模型在使用 EmoBench 风格的情绪理解（EU）和情感意识（EA）数据进行微调前后的不同版本。使用合成情绪推理数据进行微调，可提升大多数模型和类别的表现。

6.1 跨模型比较

表1展示了多个模型和微调方案的结果。

情绪理解（EU）。Qwen2.5-7B 和经过微调的 Llama3.1-8B 获得了最高的整体情绪理解得分（0.31 和 0.215）。基础版 Mistral 和 Llama3.1-8B 在情绪线索和复杂情绪方面表现不佳。经过微调的 Mistral 在大多数情绪理解维度上都有所提升（例如，复杂情绪：0.14 $\rightarrow$ 0.22），但在高层次推理方面仍逊于表现最强的基础模型，这表明预训练和任务特定微调都很重要。

情感意识（情绪分析）。Qwen2.5-7B（基础模型和微调模型）的情感意识得分最高，尤其是在个人-自我和社会-自我情境中。微调后的 Mistral 整体情感意识得分超过 Llama3.1-8B（0.625 对 0.605），凸显了治疗风格指令微调的益处。个人-自我情境的得分始终最高，而 Llama3.1-8B 在未经微调时得分仍然较低。

7 错误分析

表现不佳的类别。社会-他人情绪分析和个人信念情绪理解仍然具有挑战性（例如，微调后的 Mistral 在社会-他人情绪分析上的得分为 0.50，在个人信念情绪理解上的得分为 0.11），这反映出相关任务对复杂的心智理论和背景知识有较高要求。增加有针对性的合成数据或可弥补这些不足。

常见失败模式。模型会混淆相似的情绪、依赖文化假设、无法识别模糊情绪，有时还会生成语义不连贯的解释。未来的工作应纳入更细粒度的情绪标签、反刻板印象示例、歧义标记以及解释质量指标。

遵循模式。 微调通过最小长度训练目标和温度控制采样，提高了 JSON 的可靠性，并减少了推理轨迹被截断或过于简短的情况。

模型名称	微调	EU（子类别）	EA（子类别）
Mistral-7B	No	总体：0.105 复杂情绪：0.14 情绪线索：0.18 个人信念：0.05 观点采择：-0.09	总体：0.405 个人-他人：0.38 个人-自我：0.46 社会-他人：0.30 社会-自我：0.48
Mistral-7B	Yes	**总体：0.155** 复杂情绪：0.22 情绪线索：0.18 个人信念：0.11 观点采择：-0.13	**总体：0.625** 个人-他人：0.66 个人-自我：0.74 社会-他人：0.50 社会-自我：0.60
Qwen2 5-7B	No	总体：0.31 复杂情绪：0.43 情绪线索：0.39 个人信念：0.29 观点采择：-0.21	总体：0.68 个人-他人：0.62 个人-自我：0.78 社会-他人：0.60 社会-自我：0.72
Qwen2 5-7B	Yes	总体：0.29 复杂情绪：0.39 情绪线索：0.36 个人信念：0.29 观点采择：-0.19	总体：0.685 个人-他人：0.60 个人-自我：0.80 社会-他人：0.60 社会-自我：0.74
Gemma-7B-IT	No	总体：0.19 复杂情绪：0.24 情绪线索：0.25 个人信念：0.16 观点采择：-0.15	总体：0.565 个人-他人：0.54 个人-自我：0.60 社会-他人：0.50 社会-自我：0.62
gemma-7b-it	Yes	总体：0.18 复杂情绪：0.20 情绪线索：0.21 个人信念：0.18 观点采择：-0.15	总体：0.565 个人-他人：0.54 个人-自我：0.60 社会-他人：0.48 社会-自我：0.64
Llama-3 1-8B	No	总体：0.04 复杂情绪：0.08 情绪线索：0.00 个人信念：0.04 观点采择：-0.03	总体：0.13 个人-他人：0.12 个人-自我：0.12 社会-他人：0.12 社会-自我：0.16
Llama-3 1-8B	Yes	**总体：0.215** 复杂情绪：0.27 情绪线索：0.18 个人信念：0.16 观点采择：-0.24	**总体：0.605** 个人-他人：0.54 个人-自我：0.72 社会-他人：0.54 社会-自我：0.62

表1：各模型的情绪理解（EU）和情绪分析（EA）得分，包括子类别表现。

8 结论

使用合成的情绪链式思维数据进行微调，可显著提升 7B 规模模型的情绪推理能力。Mistral 7B 的情绪理解提升了+47.6%，情绪分析提升了 +54.3%，输出也更加连贯、结构更加清晰。治疗式指令能有效训练情绪推断和响应选择能力。

剩余的挑战包括处理模糊/混合情绪、保持解释的忠实度，以及提升社会-他人情绪分析和个人信念情绪理解的表现。未来的工作将探索更大的模型规模、多任务训练、思维链评估，以及在真实对话数据集上的验证。

参考文献

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。Bert：面向语言理解的深度双向变换器预训练。发表于*NAACL-HLT* 论文集。

Tao Ge、Xin Chan、Xiaoyang Wang、Dian Yu、Haitao Mi 和 Dong Yu。2025年。利用 1,000,000,000 个人物设定扩展合成数据创建。

Dan Hendrycks、Collin Burns、Steven Basart、Andrew Critch、Jerry Li、Dawn Song 和 Jacob Stein- hardt。2023年。使人工智能与共同的人类价值观保持一致。*NeurIPS*。

Zaijing Li、Gongwei Chen、Rui Shao、Yuquan Xie、DongmeiJiang 和 Liqiang Nie。2024年。提升——

通过情感思维链增强大语言模型的情感生成能力。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Saif Mohammad、Felipe Bravo-Marquez、Mohammad Salameh 和 Svetlana Kiritchenko。2018年。SemEval-2018 任务 1：推文中的情感。载于第 *12* 届语义评测国际研讨会论文集。

Soujanya Poria、Devamanyu Hazarika、Navonil Ma- jumder、Gautam Naik、Erik Cambria 和 Rada Mi- halcea。2019年。MELD：用于对话情绪识别的多模态多方数据集。载于*ACL* 论文集。

Hannah Rashkin、Eric Michael Smith、Margaret Li 和 Y-Lan Boureau。2019 年。迈向富有同理心的开放领域对话模型：一种新的基准测试和数据集。载于 *ACL* 论文集。

Sahand Sabour、Siyang Liu、Zheyuan Zhang、June M. Liu、Jinfeng Zhou、Alvionna S. Sunaryo、Juanzi Li、Tatia M. C. Lee、Rada Mihalcea 和 Minlie Huang。2024年。EmoBench：评估大语言模型的情商。

Carlo Strapparava 和 Rada Mihalcea。2007年。SemEval-2007 任务 14：情感文本。载于第四届国际语义评测研讨会论文集。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 Denny Zhou。2023年。思维链提示能激发大语言模型的推理能力。*NeurIPS*。

Pei Zhou、Jay Pujara、Xiang Ren、XinyunChen、Heng-Tze Cheng、Quoc V. Le、Ed H. Chi、Denny Zhou、Swaroop Mishra 和 Huaixiu Steven Zheng。2024年。Self-Discover：大语言模型自行组合推理结构。