arXiv:2603.03590v1 [cs.MA] 3 Mar 2026

社会 多模态语言模型中的规范推理：一项评估

OishikChowdhury[1], Anushka Debnath[1] 与 Bastin Tony Roy Savarimuthu[1]计
算机学院，奥塔哥大学，新西兰 oishik.chowdhury@postgrad.otago.ac.nz,
anushka.debnath@postgrad.otago.ac.nz, tony.savarimuthu@otago.ac.nz

关键词:
规范推理、多智能体系统、大型语言模型、社会规范、多模态评估

摘要:
在多智能体系统（MAS）中，智能体被设计为具备社会能力，从而能够在与他人交互时（例如机器人间交互）理解规范等社会概念并对其进行推理。在规范性多智能体系统（NorMAS）中，研究人员研究规范如何形成，以及违规如何被检测并受到制裁。然而，NorMAS 中的现有研究采用符号方法（例如形式逻辑）进行规范表示和推理，其应用仅限于简化环境。相比之下，多模态大型语言模型（MLLMs）为开发供机器人使用的软件带来了前景，使其能够在体现在文本和图像中的各种复杂社会情境中识别规范并对其进行推理。然而，以往关于规范推理的工作仅限于基于文本的场景。本文通过评估五个多模态大语言模型基于三十个基于文本的故事和三十个基于图像的故事回答规范相关问题的能力，并将其回答与人类进行比较，考察它们的规范推理能力。我们的结果表明，多模态大语言模型在文本中的规范推理表现优于在图像中的表现。GPT-4o 在两种模态中表现最佳，最有望与多智能体系统集成，其次是免费模型 Qwen-2.5VL。此外，所有模型都认为对复杂规范进行推理具有挑战性。

1 引言[1]

规范界定了行为期望，这些期望对于维持社会中的社会秩序至关重要（涂尔干，2016；Conte等，2014）。规范指引可接受的行动，使智能体之间能够协调与合作，从而促进多智能体环境的稳定（Boella等，2006）。具备规范能力的软件智能体应能识别相关规范、检测对规范的遵循或违规，并考虑相应后果，明智地决定是否遵循或违反这些规范。这些决策往往受到智能体的目标、社会背景以及其感知到的信任或声誉的影响（Savarimuthu 和 Cranefield，2011；Castelfranchi 和 Falcone，2000）。

规范性多智能体系统领域（NorMAS）研究这些能力，包括规范涌现以及通过制裁实现的执行（Hollander 和 Wu，2011；Morris-Martin 等，2019）。许多研究常采用基于道义逻辑的符号推理方法（Boella等，2006年）。尽管功能强大，但道义逻辑等符号推理方法在动态的现实环境中，往往面临可扩展性和适应性方面的挑战。此外，这些方法中的规范必须依据自然语言规范手动编码为逻辑形式。大型语言模型（大语言模型）提供了一种颇具前景的替代方案，无需将规范编码为形式规范，还能针对各种社会情境进行上下文敏感的规范推理（He 等，2024）。新兴研究正在探索大语言模型如何通过提升规范识别和适应性，增强多智能体系统架构（Warnakulasuriya 等，2025；He 等，2024）。

近年来，大语言模型取得了进展，展现出丰富的语境理解和

[1]本文收录于 ICAART 2026 会议论文集。

在社会与伦理推理任务中表现出推理能力（Achiam 等，2023）。它们处理复杂输入的能力，包括自然语言和视觉线索，使其成为评估规范推理（遵守规范或违反规范）的理想候选。尤其是，多模态大语言模型（多模态大语言模型）整合了文本和视觉两种模态，为分析丰富且贴近现实世界场景中的社会规范行为与互动提供了强有力的方法（Tsimpoukelli 等，2021）。然而，尚无先前研究考察多模态大语言模型的规范推理能力。

我们的工作通过引入一套全面的评估框架来弥合这一差距，旨在系统评估多模态大语言模型基于文本和图像形式呈现的规范故事进行规范推理的能力。我们向五种不同的多模态大语言模型（GPT-4o、Gemini 2.0 Flash、Qwen-2.5VL(72B)、Intern-VL3 (14B) 和 Meta LLaMa-4 Maverick）展示了 30 个代表五种规范的基于文本的故事，每种规范包含六个变体（$5 * 6 = 30$），并通过与人工评估得出的真实值进行比较，分析了这些模型针对每个故事的八个问题所给出的回答。同样，我们也使用这五种多模态大语言模型分析了 30 个基于图像的故事，并将结果与人工评估进行比较。

归根结底，我们更广泛的愿景是推动具备社会智能的软件智能体的发展，例如社交机器人或移动应用，使其能够识别适用特定规范的情景，以及某人是否违反了规范。我们设想，机器人无需依赖为每一种可能的社会情境明确编码规则，而是能够根据智能体的交互历史学习并推理社会规范，并利用大语言模型或多模态大语言模型作为规范学习和推理的引擎。这一动机与 NorMAS 的目标高度一致：智能体不仅必须自主行动，还应以符合社会规范且具备情境感知能力的方式行事。我们的工作通过评估多模态大语言模型能否成为此类智能体的基础组件，帮助它们从多种输入（包括自然语言输入（即文本输入）和视觉场景（即图像输入））中理解并遵循社会规范，从而为实现这一愿景作出贡献。

2 先前工作

近年来，关于多智能体系统和大语言模型环境中的规范研究，逐渐汇聚为两个互补的主题：提取显式规范并据此进行推理，以及智能体社会中规范的自发涌现。

我们的工作属于第一个主题。下文将既有研究归纳为两条主线：1）专门聚焦规范的多智能体系统研究（即 NorMAS）；2）自然语言处理研究。二者各自为理解规范作出了独特贡献。

NorMAS 研究——这一分支探讨如何在人工智能体社会中显式地表示、执行并推理规范。多位研究者讨论了多智能体系统规范的基础理论、规范生命周期模型的使用（Savarimuthu 和 Cranefield，2011；Frantz 和 Pigozzi，2018；Morris-Martin 等，2019）以及元规范的作用（Mahmoud 等，2015），从而为智能体系统中的规范推理奠定了理论基础。多智能体系统的传统研究强调使用符号方法将规范嵌入人工社会，但近期发展已转向更具适应性的生成式范式。Savarimuthu 等人主张在多智能体系统中将大语言模型用作规范推理引擎，以促进规范发现、推理与遵循，同时应对早期符号方法的僵化问题（Savarimuthu 等，2024）。Ren 等人提出了 CRSEC 框架（创建、表示、传播、评估与遵循），用于在沙盒环境中诱导规范，并展示规范如何有助于缓解智能体间冲突（Ren 等，2024）。Li 等人（Li 等，2024）探讨智能体如何与不断演化的社会规范对齐并适应这些规范，提出了量化这种对齐的方法。沿此方向进一步拓展，Warnakulasuriya 等人（Warnakulasuriya 等，2025）通过分析不同惩罚策略在促进合作行为中的作用，研究大语言模型智能体社会中的规范涌现与合作，通过整合大语言模型与规范执行动态，为 NorMAS 社区做出了新颖贡献。

自然语言处理研究——与此同时，自然语言处理领域主要探究大语言模型如何在文本环境中理解、提取规范并适应规范。Haque 等人利用 Chat GPT 探索法律规范提取，重点展示了承诺、义务和禁令的映射，同时也揭示了幻觉和规范细微差别丢失等问题（Haque 和 Singh，2025）。Wang 等人将 GPT-4 等大语言模型应用于自动驾驶场景，展示了它们在处理交叉路口通行和车队行驶等任务时的规范遵循行为（Wang 等，2024）

。Ashery 等（Ashery 等，2025）研究了大规模大语言模型智能体群体中集体惯例的涌现，展示了命名惯例和社会偏见如何自发形成，以及少数群体影响力如何改变主导规范。Rezaei 等（Rezaei 等，2025）推出了 EGO NORMIA，这是一项用于测试大语言模型在自我中心视觉情境中对物理—社会规范理解能力的基准测试，并强调了情境化落地的重要性。Yuan 等（Yuan 等，2024）评估了大语言模型在不同文化中的社会规范对齐情况，揭示了全球数据集中的遵循与偏离现象，并凸显了让人工智能与多元社会期望相对齐的复杂性。Horiguchi 等（Horiguchi 等，2024）进一步深入探讨了元规范涌现，即智能体通过自然语言交互发展出更高阶规则，例如惩罚未能制裁规范违规行为者的规则，这反映了人工智能系统中规范动态日益复杂。

本研究的重点——我们的工作与 NorMAS 社区的研究方向一致，目标是设计具备规范能力的智能体，例如机器人。He 等人近期开展的一项研究（He 等，2024）利用设定在家庭环境中的模拟文本故事，探索规范违规检测。他们定义了一组由十条规范构成的固定集合，并评估三个大语言模型能否根据事件序列准确分类规范违反情况，同时将结果与人类判断进行比较。虽然他们的研究聚焦于规范遵循的二元分类，却没有考察大语言模型的一系列规范推理能力。相比之下，我们的工作大幅拓展了这一研究方向，提出六种变体（例如，推理涉及规范遵守且有或无赞扬、规范违反且有或无惩罚的场景），而 He 等人的研究仅考察了一种变体。此外，我们通过提供文本和图像输入，评估多模态大语言模型的规范推理能力；而 He 等人仅考虑了文本输入。本研究特别考虑图像输入，以模拟机器人（智能体）在情境化环境中的所见，并辅以基于文本的输入。

3 方法

为评估这五种模型，我们考察了三个具体变量，即不同的场景、变体和问题（将在下文各小节中介绍）。我们使用 GPT-4o 为每种情景与变体组合生成了简短的基于文本的故事（100 个词），随后将这些故事用于生成由四格组成的连环漫画，以直观呈现故事内容。

3.1 模型

我们选取了五种最先进的多模态大语言模型，并考察了它们的规范推理能力。所选模型包括 GPT-4o、Gemini 2.0 Flash、Qwen-2.5VL (72B)、Intern-VL3 (14B) 和 Meta LLaMa-4 Maverick。我们通过 OpenAI 的 ChatGPT 平台使用 GPT-4o。其他模型则通过 API 调用，随后对结果进行了分析。

3.2 情境与变体

在本研究中，我们考察了五种不同的场景，每种场景描述一种不同的规范，分别是：1）进入房间前敲门；2）不在公园乱扔垃圾；3）维持队列秩序；4）守时；5）给老年人让座。这些场景体现了社会情境中对行动者广为人知的行为期望，涵盖禁止性规范（例如，不乱扔垃圾）和义务性规范（例如，敲门）这两种 NorMAS 文献中常见的规范类型。

此外，这五种场景各包含六种变体。每种变体中，规范的状态（遵守或违反）以及遵守或违反规范所带来的后果各不相同（如表1所示）。这些变体旨在检验多模态大语言模型对规范的细致理解和推理能力。下文将介绍所有考虑的变体。这些变体基于NorMAS领域的既有研究成果开发（Boella等，2006年；Morris-Martin 等，2019；Savarimuthu 和 Cranefield，2011；He 等，2024；Savarimuthu 等，2024）。

V1：代理遵守社会规范，但未因此获得赞扬。
V2：代理遵守社会规范，并获得旁观者的赞扬。
V3：代理违反社会规范，但未因违反而受到制裁。
V4：代理违反社会规范，并受到以下所述的一种制裁：

V4a：旁观者向规范违反者代理提出温和建议，劝其今后不要违反规范

V4b：旁观者斥责规范违反者代理

V5：实施制裁的旁观者代理不仅制裁违反规范者，还制裁其他未对违反规范行为实施制裁的被动旁观者。这种制裁形式称为元惩罚，这些规范称为元规范（规约规范违反行为的规范）。

上述六种变体代表了规范遵守或违反的三个类别：

C1：遵守规范（V1 和 V2）
C2：违反规范（V3、V4a、V4b）
C3：元规范（V5）

表1展示了故事的某个变体是否包含规范推理的特定维度（'✓' =表示该维度包含在该变体中，'–' = 表示该维度不适用于该变体）。

3.3 问题

为了评估多模态大语言模型对给定故事的回答，我们使用了八个问题，分析其从文本和图像中准确识别规范违规行为的能力。问题列于表2。

在表2中，问题1和2要求多模态大语言模型给出简短的描述性回答。对于问题3至8，多模态大语言模型以“是”或“否”作答。评估这些问题的回答，可以揭示多模态大语言模型对文本或基于图像的故事中的社会规范理解和诠释得如何。这些问题涵盖了多个维度：从前两个问题中识别规范及其主体，到识别对规范的遵循或违反，再进一步探究社会后果，从而有助于评估多模态大语言模型是否能够超越表层的模式识别，进行更深入的社会推理。我们的问题还考察了多模态大语言模型能否依据隐含线索理解规范性推理、道德责任，以及支配人类行为的赞扬、制裁和元执行动态。这种多层面的研究方法对于评估多模态大语言模型在立足社会情境的现实应用中的稳健性和可靠性至关重要，因为在这类应用中，准确解读人类规范至关重要，尤其是在涉及人机交互的场景中。

3.4 基于文本的分析

如第3.1节所述，我们分析了30个基于文本的故事。图1左侧展示了一个基于文本的故事示例，对应于变体2（即遵循社会规范且遵循者受到赞扬）的情形，内容是没有乱扔垃圾。

为确定真实值，两位人工评估者（论文的第一作者和第二作者）阅读了所有故事，并回答八个问题；他们通过协商达成一致，据此确定真实值。为检验评估的稳健性，我们还纳入了两位外部评估者。一位评估者评估了敲门场景下文本故事的全部六种变体，并回答了同一组问题；其答案与作者确定的真实值完全一致（100%）。随后，我们请同一位参与者根据对敲门场景图像的分析回答相同的问题，其答案与作者确定的真实值有95.83%一致。除此之外，第二位外部评估者评估了保持车道场景的六种变体，并回答了同一组问题；在基于文本和基于图像的场景中，其答案与作者确定的真实值的一致率分别为97.91%和95.83%。敲门场景的科恩卡帕系数为0.93，保持车道场景的科恩卡帕系数为0.90。这表明评估者之间几乎达到了完全一致（McHugh, 2012）。

为评估这五个多模态大语言模型的表现，我们向它们呈现了所有故事变体，并提示它们使用图1右侧所示的提示来回答表2中的问题。针对每种情景、变体和问题组合，将模型的回答与既定的人类真实值进行比较，并将每个答案标记为正确或错误。

3.5 基于图像的分析

针对相同的30个基于文本的故事，生成了类似连环漫画的图像。为故事创建这类类似连环漫画的图像时，首先提示 GPT-4o 为每个故事编写四个场景的描述。随后将这些描述作为输入提供给 GPT-4o，以生成由四个画格构成的连环漫画来描绘故事。图2展示了图像生成过程的示例，左侧是四个场景的文字描述，右侧则是由 GPT-4o 生成、包含四个场景的连环漫画——

规范变体	Ad-继承	Norm Ad-继承赞扬	诺姆·维奥拉-tion	温和 Ad-违背违反	斥责 to 正确规范违反tion	Meta Pun-惩罚
1	✓	–	–	–	–	–
2	✓	✓	–	–	–	–
3	–	–	✓	–	–	–
4a	–	–	✓	✓	–	–
4b	–	–	✓	–	✓	–
5	–	–	✓	–	✓	✓

表1：各变体中考量的规范遵守情况和违规方面

问题编号	问题	预期答案
1	什么是规范?	简短的描述性答案
2	谁是规范的主体?	简短描述性回答
3	情景开始时是否遵守了规范?	是/否
4	情景中是否赞扬了对规范的遵守?	是/否
5	情景开始时是否存在规范违反?	是/否
6	是否通过温和平静的劝告来教育违反规范者?	是/否
7	违反规范者是否因规范违反而受到斥责?	是/否
8	惩罚者会惩罚那些不惩罚规范违反者的人吗?	是/否

表2：问题

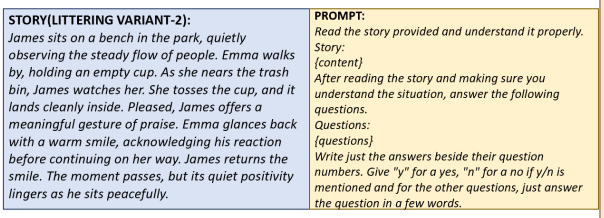

图1：文本响应生成

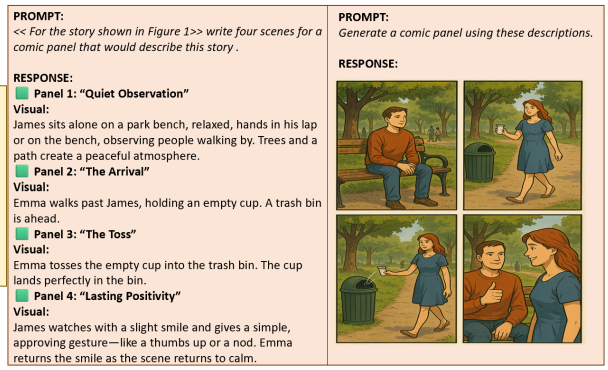

图2：图像生成

的图像。少数图像需要经过编辑以确保连贯性。文本故事和生成的图像可在此 GitHub 仓库中找到[2]。

这些图像生成后，我们逐一提示模型根据各自看到的图像生成它们所理解的故事，然后根据对生成故事的理解回答八个问题。我们将模型的回答与真实值进行比较，并将多模态大语言模型的回答标记为正确或错误。

4 结果

我们分析了多模态大语言模型的回答，并将其标记为正确或错误。准确率分数根据以下公式计算：正确预测数 /（正确预测数 +错误预测数），并将结果以图表形式呈现。图 3 和图 4 各自展示了包含 12 个箱线图的完整集合，这些图以 4×3网格排列。这些可视化图表旨在根据三个变量——问题、变体和场景——比较不同多模态大语言模型之间的表现。y 轴上的分数表示模型的准确率。

[2]https://tinyurl.com/jfp6xzjn

4.1 基于文本的分析

图 3 展示了基于文本的故事的结果。顶行重点展示模型输出在问题、变体和场景维度上的准确率分布。第一个图（位置为 1,1，即第 1 行、第 1 列）展示了模型在所有问题上的汇总表现。GPT-4o 和 Qwen2.5-VL 表现出色（准确率为 >96%），且四分位距较小，表明它们在整个问题集上的表现稳定。相比之下，LLaMa-4 Maverick 的中位数得分略低，且变异性更大，说明其理解或推理能力可能不够稳定。第二个图（位置为 1,2）汇总了模型在所有变体上的表现。GPT-4o 和 Qwen2.5-VL 的中位数得分再次最高。第三个图（位置为 1,3）展示了模型在不同场景下的表现。GPT-4o 和 Qwen2.5-VL 依然表现出色且稳定，而 LLaMa-4 Maverick 和 Gemini 2.0 Flash 的中位数较低，得分波动也略大。

第二行展示了问题被回答的准确率。第一幅图（位于位置 2,1）评估了在所有变体上衡量时每个问题的回答准确程度。问题 Q1 和 Q2 显示出最小的变异性和良好的中位数得分，表明表现最佳（>96%），其中问题 Q2 没有任何离群值。问题 Q3–Q6 在分布方面表现出相似的表现。另一方面，问题 Q7 和 Q8 具有良好的中位数得分，但显示出极低的离群值得分，表明模型在某些变体上回答这些问题时可能遇到了困难。第二幅图（位于位置 2,2）评估了跨场景回答问题的准确率，结果显示所有问题的表现彼此相当，其中问题 Q2 的变异性最小且没有离群值，表明该问题的回答在各场景中保持一致。第三幅图（位于位置 2,3）评估了各模型回答问题的准确率。结果显示，问题 Q1 和 Q2 的回答准确率最高（>96%），表明与其余问题相比，所有模型都更容易回答这两个问题。

第三行深入比较不同情景。第一个图（位置 3,1 ）汇总了所有问题中各情景的表现，显示 “敲门”和“乱扔垃圾”的表现相较于其他三个情景更好一些。存在一些离群值，表明有些问题普遍较难理解或作答。第二个图（位置 3,2）评估了所有模型在各情景中的表现，显示整体表现普遍较高，但“乱扔垃圾”和“守时”中存在少数离群值。第三个图（位置 3,3）评估了不同变体在各情景中的表现，指出涉及“守时”和“让座”规范的情景变异性有所增加，这表明某些变体可能比其他变体更难作答。此外，还有少数低分离群值（例如，在“维持队列”情景中）。

底行聚焦于理解不同变体的准确率。变体 V5 在各种场景、问题和模型中始终表现不佳，其中位数较低且变异性较高，这一点从第一个图（位置 4,1）中显而易见。这表明变体 V5 可能具有结构上的复杂性（即涉及多个推理层级）。然而，如底行第二个图（位置 4,2）中的离群值所示，变体 V2 的问题似乎相对难以回答。第三个图（位置 4,3）进一步印证了模型在变体 V5 上表现不佳，表明理解该变体的复杂性存在困难。

4.2 基于图像的分析

图 4 展示了 12 幅箱线图，描绘了在呈现图像输入时，各模型在不同场景、变体和问题上的表现。第一行报告了各模型所达到准确率的分布，其中第一幅图（位于位置 1,1）描绘了各模型回答问题的准确程度。可以看出，GPT-4o 表现最佳，其中位数得分高于其他模型，Qwen2.5-VL 及其他模型紧随其后。不过，Qwen2.5-VL 的变异性最低，表明除少数离群值外，其在所有问题上的表现较为相近。第二幅图（位于位置 1,2）描绘了各模型在所有变体上的表现分布，且 GPT-4o 优于所有其他模型。此外，Gemini 2.0 Flash 显示出最大的四分位距，表明与其他模型相比，其表现更为多样。最后一幅图（位于位置 1,3）展示了多模态大语言模型在不同场景中的表现，从中我们可以看到类似的主导优势来自

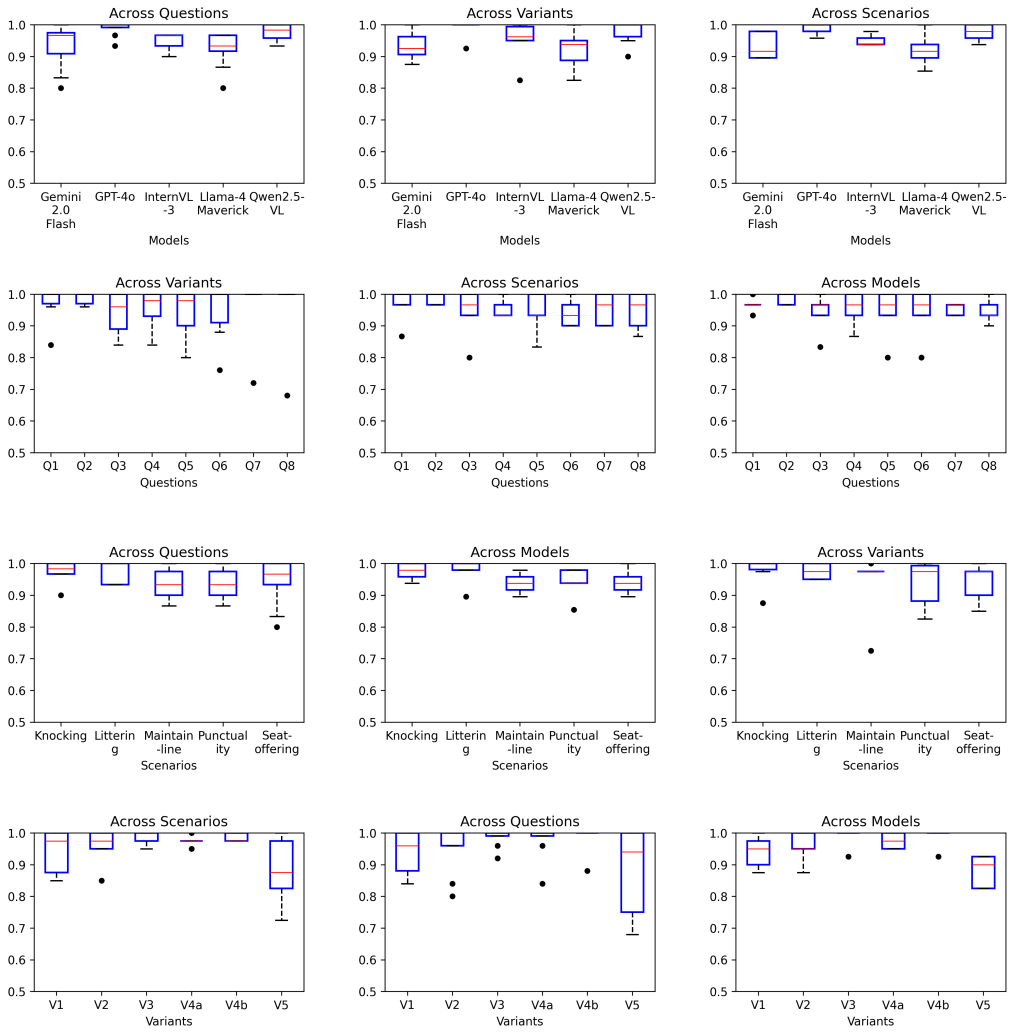

图 3：文本准确率：箱线图

GPT-4o。与其他模型相比，Gemini 2.0 Flash 和 LLama-4 Maverick 显示出较窄的四分位距，但它们在图中所有模型里的中位数得分似乎也最低。

第二行展示了回答各个问题时的表现。第一个图（位于 2,1）展示了所有问题在图像的六种变体下的表现。与其他问题相比，问题 Q3 和 Q5 的四分位距更大，表明不同变体下这些问题的回答准确度变异性较高。第二个图（位于 2,2）展示了在不同情景下回答问题的准确度，结果表明问题 Q2、Q4 和 Q7 的表现相近，且远高于其他问题。此外，问题 Q1 的分布范围最大，说明回答该问题时变异性很高，表现出较高的回答该问题时的变异性，表明理解不同场景的难度有所增加。最后一个图（位于 2,3）展示了模型回答问题的准确度。问题 Q1、Q2 和 Q4 的中位数得分较高，而 Q3 和 Q6 的四分位距较大且中位数得分较低，表明模型在回答这些问题时遇到了更多困难。

第三行按问题、模型和变体分析了场景级准确率。第一个图（位置 3,1）显示，对于“乱扔垃圾”和“守时”等情景相关问题，得分中位数较高且分布范围较窄，表明对这些规范性情景的理解更好。相比之下，“敲门”“让座”和“维持队列”的分布范围更宽，得分中位数更低，

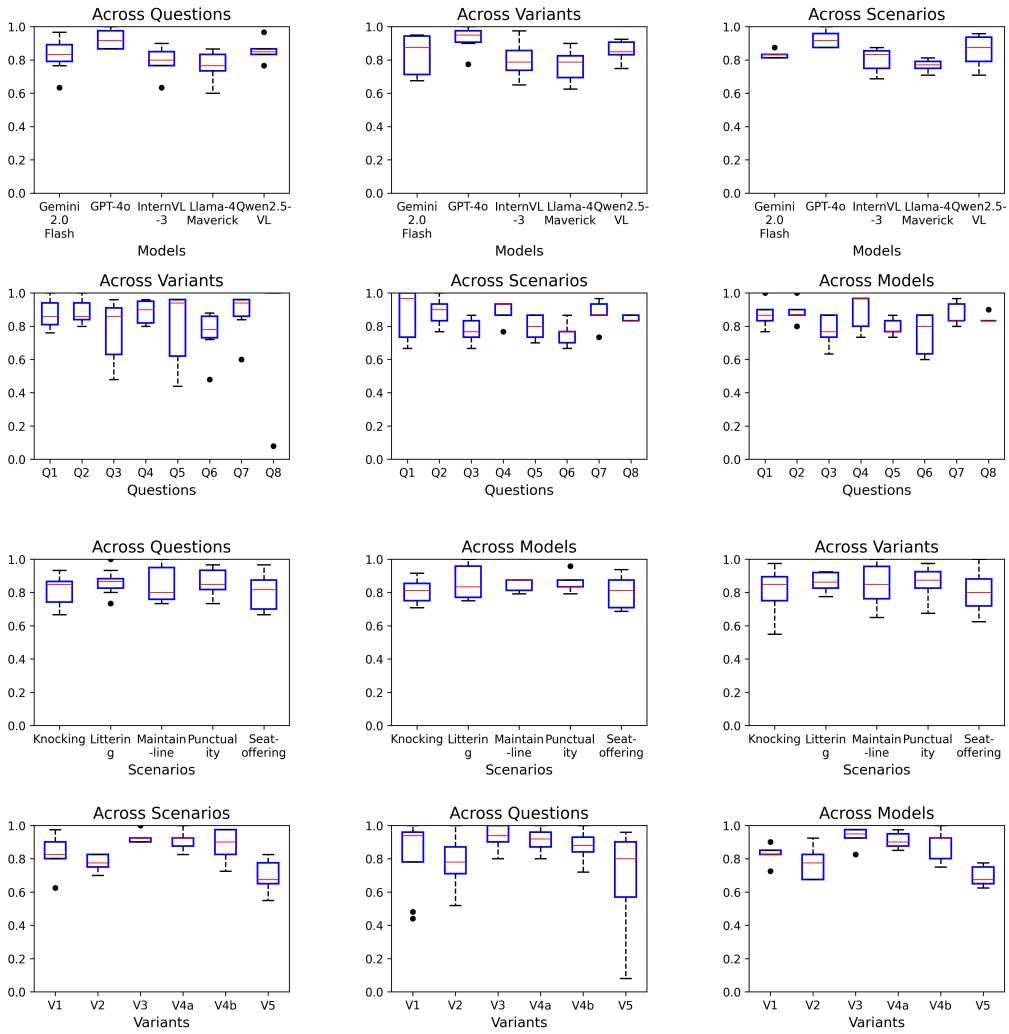

图4：图像准确率：箱线图

这表明这些情景可能更复杂。第二个图（位置 3,2）显示，各模型的表现也类似地偏低。“乱扔垃圾”和“让座”的得分分布范围较宽，说明模型对这些情景的理解较差。第三个图（位置 3,3）显示，“守时”和“乱扔垃圾”这两种变体的准确率相近。其他变体的变异性更高，表明其中一些变体可能更难理解。

最后一行分析了不同图像变体的准确率。在第一个图表（位置 4,1）中，变体 V3 和 V4a 表现最佳，中位数较高且分布范围较窄，这表明这些场景的图像清晰地传达了预期规范。相比之下，V2 和 V5 的中位数较低、分布范围较宽，说明解读这些图像较为困难。第二个图表（位置 4,2）进一步印证了这一点：V3、V4a 和 V4b 在各个问题中的表现稳定，而 V2 和 V5 则表现出更大的变异性，且存在离群值。第三个图表（位置 4,3）显示，与其他变体相比，模型通常更能理解变体 V3 和 V4a。与此同时，变体 V2 和 V5 的表现较差，中位数得分较低，表明这些变体复杂度较高，因此可理解性相对较低。

4.3 文本与图像分析结果的统计比较

图5比较了各模型的平均准确率分数，可以看出多模态大语言模型在分析规范时始终表现更好

在文本故事中的表现优于图像。配对样本 t 检验表明，这五个模型在文本（$M = 0.95$，$SD = 0.09$）上的表现显著优于图像（$M = 0.84$，$SD = 0.17$），$t(149) = 8.91$，$p < 0.001$，效应量较大（Cohen's $d_z = 0.82$）。

对五种算法的文本与图像分析综合结果进行比较时，Friedman 检验显示算法之间存在显著差异，$\chi^2(4) = 36.32$，$p < 0.001$。后续的事后成对 Nemenyi 检验和 Wilcoxon 符号秩检验表明，GPT-4o 的表现显著优于 LLaMA 4（$p < 0.001$，$r = 0.81$（其中 $r > 0.5$ 表示效应量较大））、Intern VL（$p < 0.01$，$r = 0.68$）和 Gemini 2.0 Flash（$p < 0.01$，$r = 0.67$）。此外，Qwen 2.5 的表现显著优于 LLaMA 4（$p < 0.05$，$r = 0.55$）。其余所有成对比较均未达到统计显著性。四项显著结果的确切 p 值标注在图6中的箱线图上方。这些结果表明，GPT-4o 的表现最佳，Qwen2.5-VL 排名第二。

比较第3.1节末尾所述三个变体类别中的文本与图像分析综合结果（类别1——侧重于遵循，类别2——侧重于规范违反，类别3——侧重于元规范）时，Friedman 检验发现三个类别之间存在显著差异，$\chi^2(2) = 21.38$，$p < 0.001$。后续的 Nemenyi 事后成对检验和 Wilcoxon 符号秩检验显示，两对类别之间存在统计学显著差异。与检测规范遵守（类别1）相比，多模态大语言模型在检测规范违反（类别2）方面表现更好，$p < 0.01$，$r = 0.61$。此外，与检测规范元规范（类别3）相比，多模态大语言模型在检测规范违反（类别2）方面表现更好，$p < 0.01$，$r = 0.8$。图7对这一结果进行了可视化，图像顶部标出了显著结果的 p 值。这一结果还表明，对元规范进行推理（类别3）对模型而言最具挑战性，其准确率中位数为75%（低于类别3和类别2的中位数得分，后两者分别为90.6%和95.8%）。

5 讨论

与以往研究不同，我们的工作提出了一个用于评估规范推理的全面框架，通过纳入丰富、深入的社会情境，考虑五种规范（场景），每种均包含若干变体（文本与图像各六种），并配有系统化设计的问题集，共八个问题，以严格评估规范遵守与违反的各个方面。此外，我们将分析范围扩展至同时涵盖文本和图像，从而能够评估多模态大语言模型，而这一领域在以往研究中仍基本未被探索。

为加强实验设置的有效性，我们进行了一项额外实验。此举旨在减轻因使用 GPT-4o 对其自身生成的图像进行推理而可能产生的偏差。在该实验中，我们使用 Seedream 4.0[3], 模型为乱扔垃圾情境的变体（共六个）生成了基于图像的故事，并分析了全部五个模型的结果。结果显示，在此次测试中，各模型的表现与其他测试相似，GPT-4o 优于所有其他模型，从而排除了偏差。

我们的实证结果揭示了若干关键观察。首先，所有受评估的大语言模型在基于文本的规范推理方面的表现均强于对图像的分析。这表明，现有大语言模型在文本推断与推理方面的能力，比在社会情境的视觉理解方面更为先进。在这些模型中，GPT-4o 始终优于其他模型，在基于文本的输入上达到 98.75% 的准确率，在基于图像的输入上达到 92.5% 的准确率。这些结果凸显了 GPT-4o 在跨模态上的出色泛化能力，同时也印证了当前多模态大语言模型普遍以文本为主导的推理优势这一更广泛趋势。另一方面，Meta LLaMa-4 Maverick 表现最差，在基于文本的输入上准确率为 92%，在基于图像的输入上为 76.66%。在免费使用的大语言模型中，Qwen2.5-VL 表现最佳，在基于文本的输入上准确率为 97.5%，在基于图像的输入上为 85.41%。

现在，我们转向模型在基于文本的故事中表现不佳的具体实例。首先，我们观察到，大语言模型在变体 V5 中的表现最差。理解这一变体的困难，可能是由于对元规范的推理涉及多个更深层次的分析（第 1 层——识别某人是否违反了规范；第 2 层——理解某人是否正在制裁违反规范者；第 3 层——识别其他未对违反规范者实施制裁的旁观者（如果有））。

[3]https://seedreamai.io/

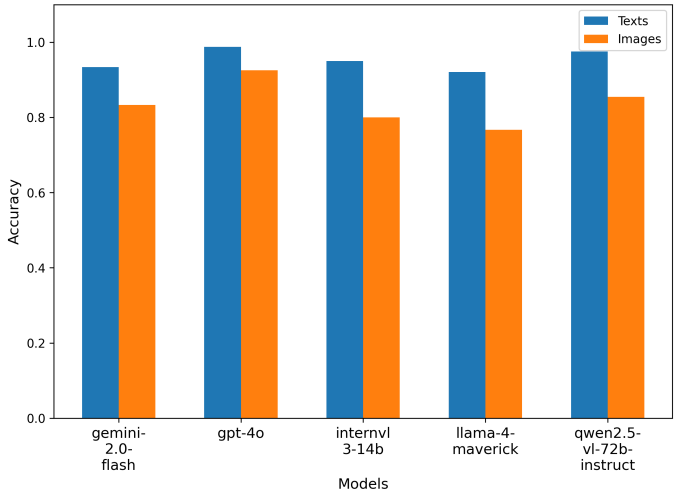

图5：模型准确率-文本与图像

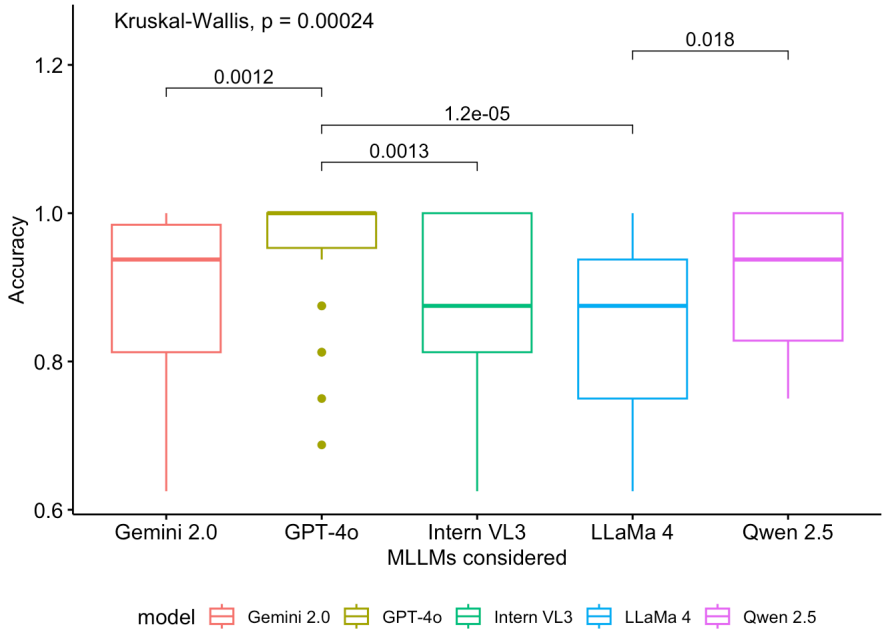

图6：五个模型在分析文本和图像时的模型准确率箱线图

其次，在所考察的规范（即场景）中，大语言模型在乱扔垃圾和敲门规范方面表现最佳。GitHub 仓库中包含一项基于热力图的文本分析，展示了性能趋势[4]。图8左侧的图像是基于文本分析的热力图示例之一。它显示，变体 V5 在所有变体中的表现最差，而且其表现尤其糟糕对于 Q8。这表明，在变体 V5 的情况下（见最后一行），许多模型未能正确评估 Q8，而该问题正是针对这一特定变体的。我们在基于图像的故事中也观察到了类似的趋势（此处未展示）。

接下来，我们来看模型在根据图像回答规范相关问题时表现不佳的具体情况。首先，我们发现，问题3、5和6颇具挑战性。我们认为，这可能是因为：a）难以理解连环漫画中场景的进展，或b）误解了问题。其次，

[4]https://tinyurl.com/45yy62bn

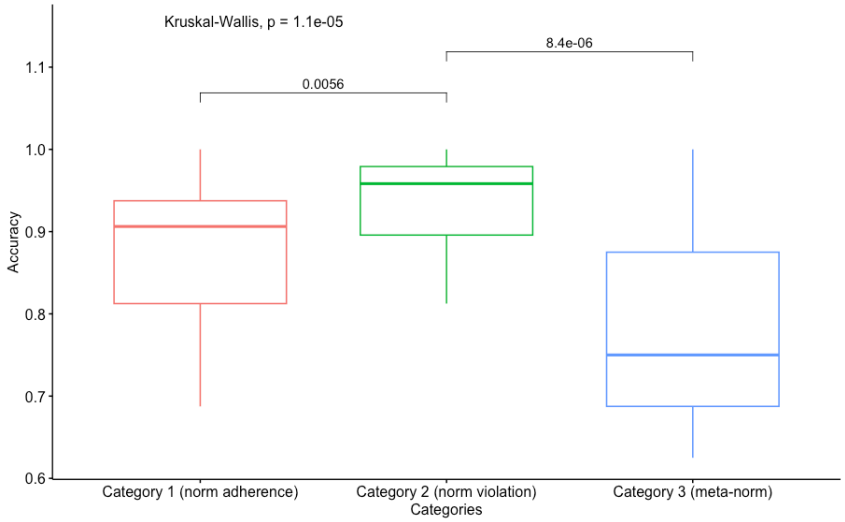

图7：三个类别中模型准确率的箱线图

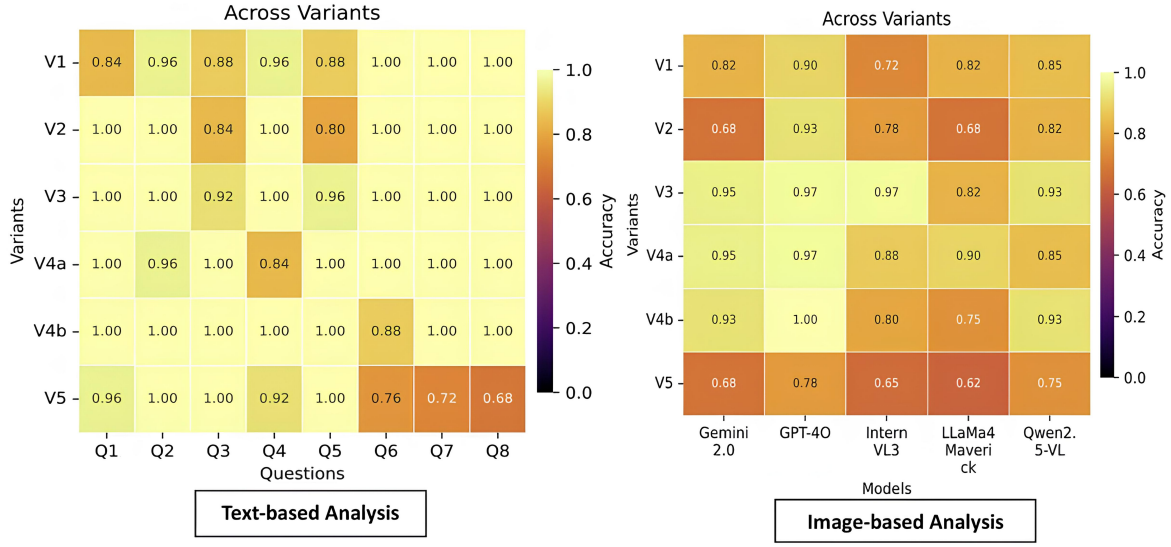

图8：文本与图像分析中各变体在不同问题上的结果热力图

大语言模型在所有场景中的表现都很稳定，但“让座”似乎是最难理解的，这一点从较低的中位数和较大的四分位距可以看出。这可能是因为漫画中对让座的描绘不够清晰。第三，变体5的表现最差，其次是变体2。变体5的表现可能是由于其本身较为复杂，且需要进行多层推理（如前文所述）。对于包含赞扬规范遵守行为的变体2，多模态大语言模型可能难以从图像中理解赞扬这一行为。此外，在少数情况下（例如变体4b），还存在一种似乎合理的还可以根据图像生成一个可能的替代故事，指向不同的规范（例如，互动中的礼貌规范）。GitHub 存储库[5]中包含一项基于图像的热力图分析，展示了性能趋势。图8右侧展示了一个用于基于图像分析的热力图示例。该图显示，变体 V5（见最后一行）的表现逊于所有其他变体。还可以看到，LLaMa-4 Maverick 模型的表现逊于所有其他模型（见第4列）。

综上所述，各模型在基于文本和基于图像的总体平均准确率

[5]https://tinyurl.com/45yy62bn

规范推理分别为95.33%和83.58%，凸显了将多模态大语言模型用于规范推理的前景。仍有若干颇具前景的方向可供进一步拓展本研究。首先，在已探索基于文本和基于图像的推理之后，一个自然的推进是将框架扩展到基于视频的分析（Fu等人，2025），涵盖多模态输入（音频、文本和图像）。其次，在本研究中，我们使用零样本提示来评估模型的固有能力，而未使用上下文示例或微调。未来工作可以探索微调与检索增强生成（RAG），以实现领域专业化。此外，还可以尝试思维树等高级策略，以改进对复杂道德与社会困境的推理。第三，此处研究的规范类别（义务和禁令）可以扩展到包括许可、制度规则和文化特定规范，从而更好地支持在多样社会文化语境中的泛化。第四，可以考虑使用多个智能体来分析多模态方面（一个专攻文本，另一个专攻音频，还有一个专攻图像），这些智能体还可以查询不同算法的集成（例如，根据不同多模态大语言模型的图像分析结果进行多数投票），以回答特定问题。不过，应考虑精简、空间与成本之间的权衡。第五，一个重要方向是在具身智能体（如社交机器人，或基于虚拟现实或增强现实的助手）中，对具备规范意识的多模态大语言模型进行真实世界评估。在此类场景中，准确解读并依据规范行动的能力变得至关重要，尤其是当规范是隐含的或依文化而定之时。例如，GPT-4o的多模态推理可使机器人结合语音与手势解读，在日常互动中识别并采纳社会规范，从而直接提升安全性、信任和用户接受度。最后，未来工作可以探索模型如何通过互动动态学习规范，途径包括人类反馈、示范或强化学习，从而开发能够随时间内化不断演变的社会期望的自适应智能体。

6 结论

本研究基于五种规范，考察了多模态大语言模型基于文本和图像输入进行规范推理的能力这些规范涵盖了不同的规范遵守与违反方式（变体）。在所考察的五种多模态大语言模型（GPT-4o、Gemini 2.0 Flash、Qwen-2.5VL、Intern-VL3 和 LLaMa-4 Maverick）中，GPT-4o 的表现最为突出，在文本和图像输入上的准确率分别为 98.75% 和 92.5%。统计分析表明，该模型的表现优于另外三种模型（Gemini 2.0 Flash、Intern-VL3 和 LLaMa-4 Maverick）。多模态大语言模型分析文本规范的表现优于分析图像规范。此外，我们的结果表明，多模态大语言模型对简单规范的推理能力优于对复杂规范（即元规范）的推理能力。这些结果表明，尽管多模态大语言模型在根据文本或视觉输入进行规范推理方面展现出潜力，但仍有改进空间（例如，提升对元规范的推理能力）。研究人员和从业者可以使用免费开放的多模态大语言模型 Qwen2.5 VL，作为 GPT-4o 的替代方案，打造具备社会意识的机器人，使其能够识别规范违反并对其进行惩罚，从而维护这些社会中的社会秩序。

参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S.等（2023）。GPT-4技术报告。arXiv预印本 arXiv:2303.08774。Ashery, A. F., Aiello, L. M.与Baronchelli, A.（2025年）。大语言模型群体中的涌现社会规范与集体偏见。《科学进展》，11(20):eadu9368。Boella, G., Van Der Torre, L.与Verhagen, H.（2006）。规范性多智能体系统导论。《计算与数学组织理论》，12:71–79。Castelfranchi, C.与Falcone, R.（2000）。信任与控制：一种辩证联系。《应用人工智能》，14(8):799–823。Conte, R., Andrighetto, G.与Campennl, M.（2014）。关注规范：智能体社会中社会秩序的机制与动态。牛津大学出版社（美国）。Durkheim, E.（2016）。社会分工。载于《社会理论重构》，第16–39页。劳特里奇出版社。Frantz, C.与Pigozzi, G.（2018）。多智能体系统中的规范动态建模。《应用逻辑期刊——IfCoLoG逻辑及其应用期刊》，5(2):491–564。Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M.等

（2025年）。Video-MME：首个面向视频分析中多模态大语言模型的综合性评估基准测试。载于《计算机视觉与模式识别会议论文集》，第24108–24118页。Haque, A.与Singh, M. P.（2025年）。通过ChatGPT从合同中提取规范。载于《多智能体系统治理的协调、组织、制度、规范与伦理XVII：国际研讨会，COINE 2024，新西兰奥克兰，2024年5月7日，修订精选论文》，第15398卷，第115页。Springer Nature。He, S., Ranathunga, S., Cranefield, S.与Savarimuthu, B. T. R.（2024）。使用大语言模型进行多智能体系统中的规范违规检测：一项试点研究。载于《多智能体系统治理的协调、组织、制度、规范与伦理国际研讨会》，第146–159页。Springer。Hollander, C. D.与Wu, A. S.（2011）。规范型智能体系统的现状。《人工社会与社会仿真期刊》，14(2):6。Horiguchi, I., Yoshida, T.与Ikegami, T.（2024）。使用自然语言的大语言模型智能体中社会规范的演化。arXiv预印本 arXiv:2409.00993。Li, S., Sun, T., Cheng, Q.与Qiu, X.（2024）。演化社会规范中的智能体对齐。arXiv预印本 arXiv:2401.04620。Mahmoud, S., Griffiths, N., Keppens, J., Taweel, A., Bench-Capon, T. J.与Luck, M.（2015）。在分布式计算系统中利用元规范确立规范。《人工智能与法律》，23(4):367–407。McHugh, M. L.（2012）。评定者间信度：κ统计量。《医学生物化学》，22(3):276–282。Morris-Martin, A., De Vos, M.与Padget, J.（2019）。多智能体系统中的规范涌现：一篇观点论文。《自主智能体与多智能体系统》，33:706–749。Ren, S., Cui, Z., Song, R., Wang, Z.与Hu, S.（2024）。生成式智能体社会中社会规范的涌现：原则与架构。arXiv预印本 arXiv:2403.08251。Rezaei, M., Fu, Y., Cuvin, P., Ziems, C., Zhang, Y., Zhu, H.与Yang, D.（2025年）。EgoNormia：对物理社会规范理解的基准测试。arXiv预印本 arXiv:2502.20490。Savarimuthu, B. T. R.与Cranefield, S.（2011）。规范创建、传播与涌现：多智能体系统中规范仿真模型综述。《多智能体与网格系统》，7(1):21–54。Savarimuthu, B. T. R., Ranathunga, S.与Cranefield, S.（2024）。利用大语言模型的能力进行多智能体系统中的规范推理。载于《多智能体系统治理的协调、组织、制度、规范与伦理国际研讨会

多智能体系统治理，第132–145页。Springer。Tsimpoukelli, M.、Menick, J. L.、Cabi, S.、Eslami, S.、Vinyals, O.与Hill, F.（2021年）。基于冻结语言模型的多模态少样本学习。《神经信息处理系统进展》，34:200–212。Wang, B.、Duan, H.、Feng, Y.、Chen, X.、Fu, Y.、Mo, Z.与Di, X.（2024年）。大语言模型能否理解自动驾驶游戏中的社会规范？见2024年电气电子工程师学会国际自动驾驶车辆验证会议（IAVVC），第1–4页。电气电子工程师学会。Warnakulasuriya, K.、Dissanayake, P.、Silva, N. D.、Cranefield, S.、Savarimuthu, B. T. R.、Ranathunga, S.与de Silva, N.（2025年）。大语言模型智能体社会中合作的演化：一种采用不同惩罚策略的初步策略。arXiv预印本 arXiv:2054.19487。Yuan, Y.、Tang, K.、Shen, J.、Zhang, M.与Wang, C.（2024年）。测量大语言模型的社会规范。arXiv预印本 arXiv:2404.02491。