arXiv:2611.31009v1 [physics.soc-ph] 23 Nov 2026

城市声誉的起源：使用智能体人工智能和自然语言处理测量在线评论中的信任度和偏见

Rafiazka Hilman,[1,2], J´ulia Koltai[1,3]

1. MTA–TK Lend¨ület “Momentum”数字社会科学社会分层研究组，ELTE社会科学中心，布达佩斯，匈牙利 2. 信息学研究所，阿姆斯特丹大学，阿姆斯特丹，荷兰 3. 实证研究所，ELTE社会科学学院，布达佩斯，匈牙利

摘要

在线评论将数百万个体的判断汇总为影响人们就餐、住宿和旅行地点的信号，人们也越来越多地将其视为地点声誉的证据。然而，这些声誉究竟真正反映了地点本身，还是更多地反映了用于测量它们的工具，尚不明确。在此，我们比较了两种用于测量来自十个美国城市的100,000条Yelp 评论中的信任度和偏见的工具：一种包含评估器-批评器循环的基于大语言模型的智能体系统，以及一种建立在预训练分类器、词典和基于规则的特征之上的基于语料库的工具。结果显示，这两种工具在可信度方面达成一致，但在合理性方面却完全不一致。它们对词语和短语的不同使用方式也会作出不同反应。基于语料库的工具更擅长标记投诉和概括性判断，而智能体则更常准确引用涉及社会群体的内容。我们还发现，仅提供五个示例时，智能体将普通的最高级表达判定为偏见的频率，远高于不提供示例或仅提供一个示例时的情况。将结果汇总到城市层面后，两种工具都发现城市之间的变异很小，因为评论层面声誉的方差中只有不到百分之零点五存在于城市之间。这些城市之间的微小差异，主要源于不同程度的信任（对于智能体式工具而言）和偏见（对于基于语料库的工具而言），而且两种工具对城市的排名也不同。因此，根据平台评论得出的声誉取决于评论如何被处理和分析。随着平台将这一任务交给语言模型，内置于这些模型中的定义和示例也就成为地点获得评价的方式之一。

关键词： 在线评论、城市声誉、大语言模型、智能体人工智能、平台估值

1 引言

如今，在选择餐厅、酒店或要去游览的城市之前，大多数人都会先阅读陌生人写下的相关内容。评论会影响消费者选择哪些产品和地点（Babi´c Rosario 等，2016年、2020年）。对于体验品——其质量只有在消费后才能评判——消费者尤其认为发布在在线社区中的评论值得信赖（Bae 和 Lee，2011年）。当众多消费者都依赖评论时，这种依赖会对商家产生可衡量的经济后果。在 Yelp 上，餐厅平均评分每提高一颗星，其收入就会增加 5-9%，迄今为止，该平台在一定程度上已经取代了连锁归属等较为传统的声誉形式（Luca，2016年）。因此，评论平台的作用不只是记录意见；它们还将数以千计的个人判断汇总为信号，引导人们选择在哪里用餐、住宿和旅行。

这些信号也会附着在旅游目的地上。旅游研究将在线旅游评论视为一种电子口碑，它揭示了游客旅行后对目的地形成的形象（Guo 等，2021；Marine-Roig 和 Huertas，2020）。评论表达了写作者眼中某个地方的安全程度、吸引力或真实感，并将这种形象传递给潜在游客（Marine-Roig 和 Huertas，2020）。这一过程影响着哪些形象得以浮现、哪些形象逐渐消退，而其走向在一定程度上取决于平台，因为同一目的地在不同评论网站上的呈现方式可能有所不同（Guo 等，2021）。评论还可能将社会偏见带入某个地方的形象之中。例如，在布鲁克林，Yelp 评论者用 贫民区 和 破败不堪等词语描述一个以黑人居民为主的街区里的餐馆，同时称赞一个以白人居民为主的街区舒适而真实。Zukin 等（2017）将这种现象称为“话语性种族隔离”。大量评论汇总起来后，这类语言可能塑造整座城市或整个地区的声誉。

评论内容能否为某个地方建立可信的声誉，取决于两个特性：评论在多大程度上值得信任，以及在多大程度上存在偏见。信任至关重要，因为只有读者相信评论，基于评论建立的声誉才会发挥作用。消费者和旅行者会采纳他们认为可信的评论信息，并对其余信息打折看待（Erkan and Evans, 2016; Ismagilova et al., 2020）。信任并非理所当然，而信任的缺失也不是一种会在平均值中相互抵消的随机噪声。Yelp 上约 16% 的餐厅评论会因可疑而被过滤；当商家的声誉不佳或竞争激烈时，他们更有可能进行评论欺诈（Luca and Zervas, 2016）。当利益驱使时，企业也会发布宣传自己的促销性评论，以及贬低竞争对手的负面评论（Mayzlin et al., 2014），因此评论网站必须核实评论，以维持用户的信任（Zelenka et al., 2021）。即使是真实评分，也只是选择性样本，因为人们通常只会在经历特别好或特别糟的体验后撰写评论，这使得平均星级评分成为对质量的有偏估计（Schoenmueller et al., 2020）。因此，判断一条评论在多大程度上值得信任，必须依据其文本，而不能只看评分。

偏见之所以重要，是因为评论谈论的不只是商家。它们还会描述周围的人物和地点，而这种语言中的偏见会从个体作者传播开来，影响整个区域的形象（Zukin 等，2017年）。偏见的表现形式多种多样，从明显的侮辱到对人群或地点的微妙概括性判断。信任与无偏见这两种属性彼此独立：一份关于餐食的详细、可信描述，仍可能对餐厅所在的街区抱有刻板印象；而一条毫无偏见的评论也可能是捏造的。因此，地方声誉测量必须同时评估这两者，并根据每条评论的可信度和无偏见程度加以权衡。

Comp计算研究主要使用基于语料库的自然语 n-

言处理（自然语言处理），例如情感分类器（Hutto and Gilbert，2014；Loureiro 等人，2022）、偏见语言词典（Recasens 等，2013）、用于检测不一致性的推理模型（Williams 等人，2018；Laurer 等人，2024），以及能够揭示语言中类人偏见的基于嵌入的关联测试（Caliskan 等人，2017）。这些测量工具具有透明性，因为每个分数都可以追溯到固定规则或模型输出；它们也具有可扩展性，因为只需很低的成本，就能将其应用于数百万条评论。然而，这些工具检测的是文本中是否出现某些线索，而非这些线索在具体语境中的含义。同一个词可能表达截然不同的意思。*Never*在“这个地区的人从来不付钱”中标示的是一种以偏概全的概括，但在“食物一直没送到”中只是普通的抱怨，而“一点也不粗鲁”则包含了词典会判定为负面的那个词。反过来，偏见往往并不通过任何被标记的词语表达，而是通过“你知道镇上这一带是什么样子”之类的说法来传达。由于不同群体的写作者用词有所差异，这些错误是系统性的，而非随机的（第1.3节）；当分数汇总到地点层面时，它们也不会相互抵消。因此，建立在这类计数之上的声誉，反映的可能更多是人们的写作方式，而非他们所报告的内容。此外，许多用于衡量其中偏见的自然语言处理指标，都建立在定义模糊的构念之上（Blodgett 等人，2020年）。

大语言模型为信任度和偏见的测量提供了一种不同类型的工具。在许多任务中，它们能够以不逊于甚至超过众包工作者的质量对文本进行标注（Gilardi 等人，2023年），但其可靠性在不同任务和构念之间存在很大差异（Ziems 等人，2024年）。将其配置为智能体后，它们还能够解释自己的判断，引用支持这些判断的证据，并根据第二个模型的反馈进行修订（Madaan 等，2023；Wang 等，2024）。此类智能体评估还具有另外两项优势。它们通过人类编码员会使用的相同定义和编码方案，应用以文字定义的构念，而不是通过一组固定的代理线索来进行判断。语言模型可以根据提供给人工编码员的同一套编码手册说明对文本进行标注（Gilardi 等人，2023年）。此外，每个分数都与评论中的引用片段相对应，因此可以将每项判断与其所依据的文本进行核对。将模型输出建立在引用证据之上，可以直接完成此类核查（DeYoung 等人，2020年；Gao 等人，2023年）。然而，目前尚不清楚，它们测量的是否与基于语料库的方法所测量的信任度和偏见这一理论构念相同，还是不同的构念。答案的重要性超越了方法论层面，并具有认识论意义：如果两种方法得出的结果不一致，那么一座城市的声誉将取决于用于测量它的工具，而不仅仅取决于评论者写下的内容。

为检验这一点，我们比较来自十个美国城市的 100,000 条 Yelp 评论中，信任度和偏见的智能体式测量与基于语料库的测量，并追踪这些测量方法如何塑造城市层面声誉。我们分别对隐性偏见和显性偏见进行操作化处理，同时通过可信度和合理性对信任度进行操作化处理。一个基于 Gemma 4 31B 构建的智能体通过评估器-批评器循环，对这些隐性偏见、显性偏见、可信度和合理性进行评估，并分别采用零个、一个或五个示例进行提示。对于另一种工具，基于语料库的测量方法则使用预训练分类器、词典和基于规则的特征，对相同维度进行测量。两种工具都针对每条评论，按照从 0 到 1 的相同五点评分量表报告分数。我们将信任度和偏见结合为评论层面的声誉信号，其城市均值定义了城市声誉。我们旨在回答两个研究问题。第一，智能体评估的偏见和信任度在多大程度上与基于语料库的测量结果一致（RQ1）？第二，这两种方法如何塑造我们对城市声誉这一在线评论聚合结果的理解（RQ2）？

1.1 评论作为估值基础设施

在线评论属于声誉系统，使陌生人在缺乏私人熟悉关系的市场中能够彼此信任（Resnick 等，2000）。评分总结了过去的体验，并将其呈现给未来的买家；评分对选择和销售额的影响已有充分记录（Chevalier 和 Mayzlin，2006；Luca，2016；Babi´c Rosario 等，2016）。平台研究进一步指出，评论并不是对质量的中性记录。评论网站决定哪些声音能够被计入、如何汇总这些声音，以及如何展示结果。在这一过程中，它们重组了酒店、餐馆和目的地获得评价的方式（Orlikowski 和 Scott，2014）。这一过程属于更广泛的社会生活的平台化：平台成为生产和争议公共价值的基础设施（van Dijck 等，2018）。对内容进行排名、筛选和评分的算法也是这一基础设施的一部分，因为它们塑造了在所有已被报告的信息中，哪些内容能够被视为相关知识（Gillespie，2014）。

由此看来，任何基于评论计算出的指标，本身都是一种估值行为。当平台或城市营销人员根据评论的可信度或偏见对其评分，并汇总这些分数时，他们便为平台的运作机制增添了另一层。因此，工具的选择并非技术细节。它决定哪些评论值得信任，哪些地点显得声誉良好。

1.2 从评论内容到旅游地点的形象与特征

评论也描述旅游目的地。旅游研究利用在线旅游评论重构游客对目的地形成的形象，包括他们对安全性的感知（Marine-Roig 和 Huertas，2020年），并表明这种形象取决于评论所在的平台（Guo 等，2021年）。地方品牌研究将此类旅游目的地形象视为众多线上对话共同形成的动态结果，而非仅仅由官方营销塑造（Kavaratzis 和 Hatch，2013年）。

然而，评论不仅塑造目的地形象，也能让我们了解整个区域真实的社会变化。城市研究利用同类数据测量城市内部的社会变化，例如，Yelp 上的商家信息可以预测社区变化（Glaeser 等，2018年），而人们光顾的餐馆则揭示了城市消费的隔离程度（Davis 等，2019年）。基于位置的平台同样表明，人们大多光顾与自身社会经济阶层相同的群体所去的地点，因此消费数字痕迹承载着有关城市社会分化的信息（Moro 等，2021年；Hilman 等，2022年）。评论所使用的语言也有助于理解城市或地区的社会特征。Zukin 等（2017年）发现，在以黑人居民为主的布鲁克林街区，Yelp 评论者通过危险和肮脏来描绘餐馆；而在以白人居民为主的街区，他们则通过真实性来描绘餐馆，这种表述可能会引导投资。因此，评论可能将个体作者的偏见传递到整个区域的形象之中。这种情况是否会发生在城市层面，以及自动化工具能否可靠地检测此类语言，仍有待探讨。

1.3 使用自然语言处理衡量评论中的信任度和偏见

关于信任，对评论可信度的研究经常会识别出读者用来判断该相信谁的文本线索，这些线索可能是文本的详尽程度或平衡性，也可能是

是否存在夸大（Filieri，2016年）。读者做出此类判断时，往往依赖启发式方法，而非仔细审查在线信息（Metzger等人，2010年）；平台也面临着宣传性评论和虚假评论带来的蓄意操纵（Luca和Zervas，2016年；He等人，2022年；Zelenka等人，2021年）。计算方法会依据语言特征检测欺骗性评论（Ott等人，2011年），并根据具体性、长度和第一人称叙述等代理指标估算可信度。这些代理指标只能部分反映人类判断。其中一些与读者使用的线索相符，例如，较长且更详细的评论会被认为更有帮助（Hong等人，2017年）。然而，人们并不擅长判断欺骗行为。在一项受控研究中，人类评判者识别捏造的酒店评论时，准确率接近随机水平，而且往往会将评论当作真实评论；而根据语言特征训练的分类器则能可靠地区分两者（Ott等人，2011年）。同一项研究发现，虚假评论比真实评论更常使用第一人称，因此，一种直觉上能表明评论者有亲身经历的线索，也可能表明评论存在欺骗性。因此，计算代理指标可能会编码人们使用的启发式方法，包括其中的错误；仅仅与人类直觉一致，并不能证明可信度测量指标有效。

针对偏见，基于语料库的自然语言处理提供了偏见语言词典（Recasens 等，2013年）、毒性分类器，以及能够揭示语言中编码的刻板印象的基于嵌入的关联测试（Caliskan 等人，2017年；Garg 等人，2018年）。这些工具具有透明性和可扩展性，但自身也带有偏见。毒性分类器会不成比例地标记方言和身份术语（Davidson 等人，2019年；Sap 等人，2019年），而偏见可能通过数据、标注、表示、模型和研究设计进入自然语言处理系统（Hovy 和 Prabhumoye，2021年）。一项批判性综述发现，大多数关于偏见的自然语言处理研究都未对其底层构念作出定义（Blodgett 等人，2020年）。测量理论对这一问题进行了精确界定：只有在操作化确实测量了其声称要测量的构念时，该操作化才是有效的（Jacobs 和 Wallach，2021年）；文本即数据方法也必须经过验证，而不能想当然地认为其有效（Grimmer 等人，2022年）。

1.4 使用语言模型和智能体测量评论中的信任与偏见

大语言模型能够提升自动化文本分析的质量。在许多任务中，它们对文本的标注质量达到或超过众包工作者（Gilardi 等人，2023年）；与基于词典的文本分析相比，它们能够更加准确地识别情感倾向和冒犯性（Rathje 等人，2024年）；还能够再现人类不同亚群体的反应模式（Argyle 等人，2023年），因此日益被提议作为社会科学的工具（Bail，2024年）。然而，它们的表现会因任务而有很大差异；对于需要社会或语境解释的构念，其表现最差（Ziems 等人，2024年）。此外，它们的输出也容易受到提示方式的影响，包括少样本示例的选择、顺序和格式；这些因素可能会显著改变结果（Zhao等，2021年；Lu等，2022年；Sclar等，2024年）。

不过，智能体设计可以为大语言模型的选择提供解释，并在必要时加以修订。评估模型可以引用支持其判断的证据，第二个模型则可以对其进行审核和完善（Madaan 等，2023；Wang 等，2024）。根据以往研究，在某些情境下，作为评判者的模型与人类偏好具有较高的一致性（Zheng 等，2023）；但若没有外部反馈，自我纠正往往无法改善推理，只有在获得可靠的外部反馈时才较为有效（Huang 等，2024；Kamoi 等，2024）。此类智能体能否衡量信任

和偏见，并且比基于语料库的测量工具更有效，还是仅仅有所不同，尚未在同一批评论上进行检验。

基于对现有研究的综述，我们发现了以下两个有待探讨的空白。首先，尚未有研究在相同文本和共享量表上比较智能体式与基于语料库的信任度和偏见测量，因此尚不清楚二者是否捕捉到相同的构念（RQ1）。其次，从在线评论推断地点声誉的研究将城市声誉视为这些评论的聚合结果，但很少探讨测量方法如何塑造这一结果，因此也不清楚不同方法是否会导致对同一城市声誉的不同理解（RQ2）。完整流程见补充材料 A。

2 数据与方法

2.1 数据

我们分析了 Yelp 开放数据集（Yelp 公司，2026年）中的餐厅和商家评论，研究对象为美国评论数量最多的十个城市：印第安纳波利斯、纳什维尔、新奥尔良、费城、里诺、圣路易斯、圣彼得堡、圣巴巴拉、Tampa 和 Tucson。我们从每个城市的文件中随机抽取 10,000 条评论，先剔除文本为空或评论标识符重复的评论，共得到 100,000 条评论。在每种测量中都使用相同的评论。每条评论均根据其所属文件分配至相应城市。测量工具仅接收评论文本，而星级评分、投票数和商家属性则在之后作为声誉模型中的协变量使用。

2.2 智能体测量

这些智能体基于拥有 310 亿参数的 Gemma 4 构建（Gemma 团队，2026年），并通过 Ollama 调用。两个智能体分别负责评估偏见和信任，每个智能体都会单独评估每条评论。每个智能体都由一个评估器和一个批评器组成，二者使用相同的模型、定义和编码方案。编码方案包括提供给智能体的概念定义和等级描述。它将每个分数划分为序数量表中的五个等级（0、0.25、0.50、0.75、1），并用文字说明各等级的含义，因此模型会像人工编码员一样，通过匹配最接近的描述来为评论评分。

构念和编码方案。 偏见智能体对隐性偏见 （微妙的偏向性表述、未被直接陈述的刻板印象或假设）、显性偏见 （公开且不公平地偏袒或贬低个体、群体或实体的词语）以及总体偏见得分 b_i 进行评分。其定义明确指出，与所描述经历相关的强烈观点（“汤是凉的”）不属于偏见。偏见是指对人物或地点作出不公平、无依据或概括性的陈述，例如针对工作人员、顾客、社区或城市的刻板印象。信任智能体对可信度 （基于具体且可能可验证的细节，例如菜品、价格、等待时间或工作人员行为的主张）、合理性（描述的内部一致性和真实性）以及总体信任分数t_i进行评分。其定义指出，信任关注的是描述的可信度，而不是评论者是否喜欢该商家，因此负面评论也可能完全值得信任。对于偏见，等级范围从无偏见（0，即中立或基于亲身经历的描述），到单个

带有倾向性的词语（0.25）、明显的片面框架或无依据的概括（0.50），以及强烈的带有倾向性的语言或对人物或地点的刻板印象（0.75），再到主导整篇评论的普遍性偏见（1）。对于信任，等级范围从不可信（0，即自相矛盾、难以置信或纯粹宣传性的描述），到大多是模糊主张（0.25）、夹杂具体细节与泛泛主张（0.50），以及基于亲身经历的具体、连贯描述（0.75），再到基于可观察事实的详细且一致的描述（1）。同一量表也适用于两个维度得分以及每个智能体的总体得分。本研究用于设计信任智能体和偏见智能体的提示词见补充材料 C。

评估器。 评估器接收定义、编码方案、分步说明和评论文本，并返回结构化评估。它首先从文本中提取准确引文作为证据（偏见方面提取带有偏见的词语或短语，信任方面提取支持性证据或危险信号），然后撰写至多三句话的理由说明，并在其中引用这些引文，最后才根据编码方案选择最符合该评论的等级，为各项指标评分。这种先证据、后评分的顺序，能将每条评论的评分与文本联系起来，使读者可以对照文本审视每项评分。说明要求智能体仅依据评论文本，不对商家、其评分或受欢迎程度作任何假设。

依据核验工具和批评器。 自动化工具会在对 Unicode、引号、空白字符和字母大小写进行标准化处理后，核实所有引文是否与原评论中的文字完全一致。批评器接收评论、当前评估和依据核验报告。它会检查每条引文是否符合定义、是否遗漏了明显的证据，以及根据引述的证据，各项评分是否符合编码方案中相应等级的描述。对于这两个智能体，定义都明确指出，评论的语气不是评判标准：有强烈观点但有描述的经历作为依据，并不构成偏见；负面评论也完全可能是可信的。信任批评器还会检查评论的情感倾向是否影响了信任分数，因为读者往往认为负面评论比正面评论更可信（Lo and Yao，2019；Chen and Lurie，2013），而缺乏具体内容的热情表达则会引发对宣传性写作的怀疑（Filieri，2016）。若信任分数受语气左右，声誉度量就会变成情感倾向度量。批评器可以接受评估，也可以对其进行修订。只有在指出能够从评论文本中查证的具体问题时，才可修订，例如引文并未出现在评论中、引文不符合定义、评估器遗漏了明显证据，或评分等级与引述的证据不符。否则，默认接受评估；并且批评器收到复核要求时，不得仅因此更改评分，因为语言模型在被要求检查自己的答案时，往往会修改原本正确的答案（Huang 等，2024）。批评器最多运行三轮；若其接受评估，或没有任何评分变动达到或超过0.05，则停止。每轮评分都会被保存，因此评估器的初始评分仍可供比较。

提示设计。 我们采用三种设计运行每个智能体，分别为零样本、单样本和五样本。示例是由人类撰写并评分的合成评论，涵盖全部五个量表等级。单样本示例是五个示例中的第一个，因此这些设计是嵌套的。我们将示例同时展示给评估器和批评器，并要求它们根据每条评论本身的优劣进行判断，而不是照搬示例的分数。

推理设置。 解码过程中，输出受到 JSON模式的约束。由于语言模型以文本形式生成分数，它仍可能返回量表范围之外的值，例如 1.5 或 75。因此，每条回复在使用前都会依据同一模式进行验证。验证会检查所有必需字段是否齐全（引文、理由、分数，以及批评器的判定），每个字段的类型是否符合预期，每个分数是否介于 0 和 1 之间，以及判定是否为“接受”或“修改”。未通过任何一项检查的回复都会被弃用，而不会进行更正。比较方法时，范围内但介于两个量表等级之间的分数会四舍五入到最近的等级。模型还必须重复评论的标识符，任何不匹配都会被记录下来。

我们将温度设为 0.2，使采样结果接近确定性，并固定每次调用的随机种子，以便复现每个输出。种子的数值本身没有实质意义。我们预先将其设为 42——这是一个常见惯例——并对每条评论都使用该值，而不从多次运行中择优。每次调用都有各自的种子：评估器使用 42，$42 + 1000k$ 批评器第 k 轮使用相应种子；重试时，则在失败尝试所用种子的基础上加 1。轮次之间相差 1,000，可确保同一条评论的任意两次调用不会使用相同的种子，因此批评器轮次不会重复评估器或此前轮次的随机抽样。上下文窗口设为 16,384 个词元，足以容纳最长的提示词、包含五个示例的批评器提示词以及一条长评论。模型的扩展推理模式则关闭，因为评估器已在评分前说明其证据和理由说明。

当回复无效、因达到最大输出长度而被截断、服务器返回错误或未能及时响应时，我们会重试请求；若服务器未能及时响应，请求会等待 5 至 120 秒后再重试。连续失败五次后，该评论将不予评分。提示词包含固定指令、定义、示例；对于批评器，还包含当前评估结果以及评论本身，因此过长的评论可能导致提示词超出上下文窗口。在这种情况下，Ollama 会悄然丢弃部分提示词，而不返回错误。因此，我们会检查每个请求报告的提示词长度，并在提示词占据上下文窗口 98% 或以上时停止运行，以确保不会依据被截断的文本为评论评分。每条评论在每种设计下只评分一次。智能体测量所用的正式表达式见补充材料B。

2.3 基于语料库的测量

该基于语料库的工具使用固定的预训练模型、词汇资源和基于规则的特征来测量相同的维度。在每一步中，它都独立于智能体式测量。它只接收评论文本，不使用智能体的提示词、示例、引文、理由说明或得分。其任何模型或阈值都没有根据智能体的输出进行训练、调优或调整，并且其映射规则在两种测量结果进行比较之前就已固定。这两个工具唯一共享的是评论样本，以及它们据此报告得分的五点评分量表。评论使用 PySBD（Sadvilkar and Neumann，2020）拆分为句子，再使用 spaCy（Honnibal et al.，2020）处理词元、词元原形和命名实体。每个预训练模型都应用于单个句子，或在推理模型的情况下应用于句子对，而不是整篇评论。这些模型每次输入最多接受 512 个词元（句子嵌入模型为 384 个），因为它们对词位置的表示仅定义到该长度，输入更长时会导致错误，而不是得到较弱的结果。因此，较长的句子会被截断至此上限，并且

每篇评论只处理前 60 个句子。由于这些限制适用于单个句子，因此只有在文本异常长时才会产生影响，例如没有标点、被分词器作为一个整体分段的评论，或包含超过 60 个句子的评论。在此类评论中，被省略文本中的线索不会被计入，因此这些评论的基于语料库的得分可能低估其中包含的线索。

信任。 每个线索都将赋予智能体的定义中的一部分具体化。由于可信度被定义为建立在具体且可能可验证的细节之上，其核心线索统计了此类细节，即数值细节（价格、时间、数量、日期）和命名实体。读者会认为包含更多细节的评论更可信（Filieri，2016年），而真实评论相比虚假评论会更具体地描述具体情境（Ott等人，2011年）。之所以纳入评论长度，是因为较长的评论有更多空间提供此类细节，而且会被认为更有帮助（Hong等人，2017年）。第一人称的使用表明是在描述个人经历，符合定义要求；不过，这一线索存在歧义，因为虚假评论比真实评论更常使用第一人称（Ott等人，2011年）。夸大程度，即感叹号和全大写词语的出现率，会降低可信度，这与读者不信任夸大评论的证据一致（Filieri，2016年）。合理性指评论的内部一致性。无法从文本中验证所描述的访问是否确实发生，但可以验证评论是否自相矛盾。我们依据DeBERTa-v3自然语言推理模型（Laurer 等人，2024年），将其测量为1减去其句子对之间最大的矛盾概率，因为单个矛盾就足以削弱一段描述的可信度。该指标针对所有句子对计算；对于较长的评论，则针对可复现的45个句子对随机子集计算，因为句子对的数量会随句子数量呈二次增长（45个句子对对应一篇包含十个句子的评论）。单句评论不定义合理性。

偏见。关于信任，我们选择每个线索来捕捉赋予智能体的定义中的一个要素，而非一般的负面情绪，因为与所描述经历相关的强烈观点并不构成偏见。显性偏见，即公开且不公平地偏袒或贬低人物或群体的词语，通过 unbiased-toxic-roberta 分类器在句子层面给出的身份攻击和侮辱概率来衡量（Hanu 和 Unitary 团队，2020年）。我们使用该分类器的这两个标签，而非总体毒性，因为它们针对的是贬损人物及其群体归属的语言，而餐食引发的脏话或愤怒并不属于偏见。该分类器经过训练，以减少对提及身份群体内容的误报（Borkan 等人，2019年）；这是此类工具众所周知的弱点（Davidson 等人，2019年；Sap 等人，2019年）。Recasens 等人（2013年）的偏见词典源自维基百科中删除了部分措辞的编辑记录，可捕捉毒性分类器会遗漏的片面框架。隐性偏见，即未直接表达的偏向性框定或刻板印象，通过两个线索来识别。第一个是泛化标记（例如总是、从不、典型的、这些人）的出现频率，因为抽象、笼统地描述人物，而不是描述具体行为，是刻板印象的一种语言特征（Semin 和 Fiedler，1988年；Maass，1999年）。第二个是提及人物、群体或地点的句子在句子嵌入空间中沿愉悦—不愉悦效价轴的最大绝对投影；该效价轴由 WEAT 属性词构建，用于捕捉附着于社会目标的评价性关联（Caliskan 等人，2017年）。第三个是单类别嵌入关联测试（SC-WEAT），它使用 all-mpnet-base-v2 句子嵌入，衡量一个句子与概括性陈述（“这附近的人总是很粗鲁”）相比，与具体的、基于经历的陈述（“我们的服务员忘了给我们上饮料，过了二十分钟才想起来”）的相似程度差异，

表1：将基于语料库的线索映射至五点评分量表的 Apriori规则。按从上到下的顺序应用规则。第一个适用的规则决定等级；若没有规则适用，则等级为 0。计数单位为每条评论。毒性值为句子级概率的最大值。SC-WEAT 值为句子级效应量的最大值。“提及人物”指至少有一个句子提及人物、群体或地点。严重夸张（每 100 个词中出现 ≥ 5 个感叹号或全大写词语）会使两个信任维度均降低一个等级。

等级	显性偏见	隐性偏见
1.00	身份攻击 ≥ 0.8，或身份攻击和侮辱均为 ≥ 0.5	≥ 3 概括标记，SC-WEAT ≥ 0.8 并提及人物
0.75	身份攻击 ≥ 0.5，或侮辱 ≥ 0.8	SC-WEAT ≥ 0.8 并提及人物
0.50	侮辱 ≥ 0.5，或 ≥ 3 偏见词典词语	SC-WEAT ≥ 0.5，或 ≥ 2 标记和提及人物
0.25	侮辱或身份攻击 ≥ 0.2，或 ≥ 1 偏见词典词语	≥ 1 标记，或 SC-WEAT ≥ 0.2
等级	**可信度**	**合理性（最大矛盾概率）**
1.00	≥ 4 具体细节、第一人称和 ≥ 80 词语	< 0.20
0.75	≥ 2 具体细节、第一人称和 ≥ 40 词语	< 0.50
0.50	≥ 1具体细节，或第一人称和 ≥ 40 词语	< 0.80
0.25	第一人称，或 ≥ 20 词语	< 0.95

（Reimers 和 Gurevych，2019年）。与词表不同，它还能识别用词表中没有的词语表达的概括性判断。我们没有使用有监督偏见分类器，因为不存在符合这些定义的评论偏见标注数据。

映射两种测量方式。 为了比较基于语料库的评估得分与智能体测量结果，我们按照在与智能体进行任何比较之前预先确定的规则，将自然语言处理线索映射到同一个五点评分量表上（表1）。总体偏差取隐式水平和显式水平中的最大值，因此，一条评论的偏差程度取决于其最强的偏差形式。总体信任度是可信度与合理性的平均值。我们报告了所有阈值，以便复现和调整该映射。关于基于语料库的测量的技术说明见补充材料 D。

2.4 声誉

根据本研究的概念模型，一条评论对其所在城市声誉的贡献，与其可信度和无偏见程度成正比。我们将评论层面的声誉信号定义为 $R_i = t_i(1 - b_i)$，并将城市声誉定义为该城市 c 的 N_c 条评论的平均值：

$$R_c = \frac{1}{N_c} \sum_{i \in c} t_i \, (1 - b_i). \tag{1}$$

我们使用信任度和偏见的总体得分，分别计算每种方法的 R_c，即三种智能体设计和基于语料库的测量。

2.5 研究问题的统计分析

由于所有方法对相同的评论进行评分，我们可以使用配对统计进行比较，并在全文报告效应量。这一点很重要，因为每个城市约有 10,000 条评论，因此按照传统的显著性标准，几乎任何差异都会达到显著水平。

方法间一致性（RQ1）。我们使用斯皮尔曼秩相关 ρ 和二次加权 Cohen's κ（Cohen，1968）测量评论层面的等级一致性：将分数四舍五入至最接近的量表等级，并依据 Landis 和 Koch（1977） κ 进行解读。为检验各构念的两个维度在不同方法下是否以相同方式相互关联，我们使用 Vargha–Delaney A 统计量（Vargha 和 Delaney，2000），比较每条评论的两项分数：隐性偏见与显性偏见，以及可信度与合理性。 A是第一项分数高于第二项分数的评论所占比例，其中两项分数相等的评论按半条计。由于分数仅分为五个等级，分数相等的情况很常见；将其按半条计入 $A = 0.5$ 意味着两项分数都没有更高的倾向。我们使用威尔科克森符号秩检验检验差异。我们使用弗里德曼检验检验所有方法之间平均分数的差异，并以 Kendall's W 作为效应量，随后进行两两威尔科克森符号秩检验，并采用 Holm 校正（Holm，1979）。为展示两种方法出现较大分歧的频率，我们报告智能体与基于语料库的测量结果相差至少两个五点评分量表等级的评论所占比例。

判断背后的语言。为了更好地理解模型评估结果之间潜在相似性与差异性的来源，我们探究模型在评估时是否依赖不同的语言特征。我们比较这些文本中最常见的术语，即模型据此进行评估的文本。我们使用带有信息性狄利克雷先验的加权对数优势比（Monroe 等人，2008）将这些频率进行比较。对于智能体式模型，目标文本由其作为证据引用的段落构成，而比较文本则是所有获其评分的评论的全文，包括摘取引文的评论。因此，如果某个术语被智能体引用的频率高于根据评论中的频率所预期的水平，那么该术语就具有区别性。由于引文也属于比较文本的一部分，这种对比略显保守。对于不引用任何内容的基于语料库的模型，目标文本由其评为高分（偏见 ≥ 0.5; 信任 ≥ 0.75）或低分（信任 ≤ 0.25）的评论构成，比较文本则由其余评论构成。为从文本中提取语言特征，我们统计句子中的词语和双词短语，排除缩写形式、功能词以及长度少于三个字母的词语，保留 $always$ 和 $never$ 等泛化标记，并要求至少出现五次。

城市差异与声誉（RQ2）。使用克鲁斯卡尔–沃利斯检验检验各项得分在城市之间的差异，以$\eta^2_H = (H - k + 1)/(n - k)$作为效应量（Tomczak和Tomczak，2014），并通过城市排名进行汇总；我们使用斯皮尔曼 ρ来衡量不同方法所得排名之间的一致性。

为确定城市声誉差异究竟源于信任还是偏见，我们将每座城市相对于所有评论平均声誉的偏差，精确分解为

信任部分、偏差部分和余项

$$R_c - \bar{R} = \underbrace{(\bar{t}_c - \bar{t})(1 - \bar{b})}_{\text{trust}} \underbrace{- \bar{t}\,(\bar{b}_c - \bar{b})}_{\text{bias}} + \text{remainder}, \tag{2}$$

其中，R、$\bar{t}$ 和 b 分别是所有评论中声誉、信任和偏见的均值，$\bar{t}_c$ 和 b_c分别是城市 c 的评论中信任和偏见的均值，余项则汇总了交互作用项和城市内协方差项。该分解遵循将两个率之差拆解为各组成部分的思路（Kitagawa，1955；Das Gupta，1993）。每一部分都显示了：当另一个因素处于总体均值时，一个因素的城市偏差会使其声誉发生多大变化。将另一个因素固定在总体均值是几种可能的参照选择之一，但在这里影响不大，因为交互作用项是两个较小城市偏差的乘积。我们通过各部分所占的城市偏差绝对值比例来概述这一分解（补充材料E）。

为将城市效应与评论特征区分开来，我们针对每种方法拟合了一个线性混合模型，R_i 其中每个城市设有一个随机截距，并纳入五个标准化评论层面预测变量，即星级评分、有用、好笑和酷的投票数加一后的对数，以及评论文字长度加一后的对数。投票数和评论长度呈明显右偏分布。大多数评论没有投票或仅获得少量投票，长度也适中；但少数评论会获得大量投票，或篇幅特别长。如果不取对数，这些少数极端评论将主导估计系数及其标准化贡献。对数变换还反映了每增加一个单位所带来的重要性递减，因为从没有投票到一票，或从20个词到120个词的差异，比从100票到101票，或从1,020个词到1,120个词的差异，更能说明一条评论的特征（Gelman 等，2020）。由于计数中包含零，而零的对数没有定义，因此取对数前先加一。模型采用限制最大似然并结合多个优化器进行估计，保留具有最高有限似然值的解。对于多层模型而言，十个城市属于少量高层级单位，经验法则通常要求至少20至30个高层级单位（Bryan 和 Jenkins，2016）。这些规则针对的是高层级预测变量效应以及组间方差的推理；当群体数量较少时，后者的标准误过小（Maas 和 Hox，2005）。在线性模型中使用限制最大似然时，即使聚类数量较少，系数和方差成分本身仍可在无偏的情况下估计（Elff 等，2021；McNeish 和 Stapleton，2016）。因此，我们仅使用该模型估计城市间方差所占的比例以及评论层面预测变量的效应，不检验城市间方差。聚类稳健标准误无法避免这一问题，因为在聚类数量较少时其数值过小（Cameron 等，2008），且无法估计城市间方差。无需对城市效应作任何假设的Kruskal–Wallis效应量，可独立检验城市差异的大小。我们报告Nakagawa和Schielzeth（2013）提出的组内相关系数（城市间方差所占的比例）以及边际和条件 R^2 。预测变量的贡献以SHAP值表示，即等于 $\beta_j\,(x_{ij} - \bar{x}_j)$ （Lundberg 和 Lee，2017），并按城市取平均，以描述各城市的特征概况。为探究城市效应是否与城市特征相关，我们将估计的城市效应分别与人口、面积和餐厅数量的对数，以及餐厅密度进行相关分析，采用斯皮尔曼 ρ 并进行Holm 校正。由于本研究仅包含十个城市，我们将这些关联视为探索性结果。

3 结果

我们比较了来自十个美国城市的相同100,000条Yelp 评论的四种测量结果：一种基于语料库的自然语言处理方法，以及以零个、一个或五个示例进行提示的Gemma 431B智能体。四种方法均使用从0到1的统一五点评分量表报告偏见和信任。由于每种方法评估的都是相同的评论，我们采用配对统计并报告效应量，而不是报告 p值；在每个城市有10,000条评论的情况下，几乎所有差异都会在传统显著性水平上达到显著。

3.1 智能体式与基于语料库的测量方法对相同评论的评估各不相同

两种方法首先在评论中包含多少信任度和偏见这一问题上存在分歧（图1）。基于自然语言处理的模型发现，三分之二的评论存在隐性偏见，只有33%的评论得分为0，另有22%的评论达到0.75分，即强偏见等级。没有示例的智能体在93%的评论中未发现偏见；提供一个示例时，这一比例为94%（图1a）。在同一条评论中，基于自然语言处理的模型在大多数情况下更可能判定隐性偏见高于显性偏见（Vargha-Delaney $A = 0.81$）；而智能体对两者的评分几乎同样低（$A = 0.50$-0.54）。因此，尽管基于自然语言处理的模型与智能体方法对隐性偏见的估计存在差异，但相比之下，所有方法都认为显性偏见很少见，86%至95%的评论得分为0。此外，示例会改变智能体的校准结果。使用五个示例后，没有隐性偏见的评论比例从93%降至80%，另有16%的评论被归为轻微等级。我们将在下文进一步讨论这一影响，因为它不仅改变智能体发现偏见的数量，也改变它将哪些情况视为偏见。

信任呈现出不同的模式。两种方法都将可信度分布在整个量表范围内，其中大多数评论的得分介于 0.50 和 1 之间（图1b）。然而，合理性却使两种方法明显区分开来。智能体将 92–95% 的评论评为 1 分，而基于自然语言处理的模型则将 35% 的评论评为 0 分。在同一条评论中，对这两个维度的比较与偏见的比较相似。基于自然语言处理的模型将可信度评得高于合理性的频率与低于合理性的频率大致相同（Vargha–Delaney $A = 0.53$），而智能体在几乎每条评论中都将合理性评得高于可信度（$A = 0.08$–0.11）。因此，对于智能体而言，合理性基本处于量表上限，信任度的变异几乎完全来自可信度；而对于基于自然语言处理的模型而言，两个维度都会发生变异。这表明，两种方法以“合理性”之名测量的是不同的构念，而不是对同一构念的两种估计。差异在于定义。基于自然语言处理的模型只测量逻辑一致性，即用 1 减去句子对之间最大的矛盾概率（第 2.3 节）；而智能体的定义还涵盖描述是否具有逻辑连贯性，以及其中的主张是否现实可信（补充材料 C）。基于自然语言处理的模型在 35% 的评论中发现，至少有一对句子的矛盾概率达到 0.95 或更高，对应于得分 0（表1）；而智能体仅在 5–8% 的评论中发现合理性问题。推理模型检测到的矛盾究竟反映了真实的不一致，还是误报，在没有人工编码的情况下无法确定。

平均分数的差异证实了这一模式。各方法在隐性偏见（弗里德曼检验，肯德尔 $W = 0.52$）、总体偏差（$W = 0.52$）和合理性（$W = 0.51$）方面差异显著，但在显性偏见（$W = 0.03$）、可信度（$W = 0.03$）和总体信任度（$W = 0.07$）方面差异甚微。最大的差异出现在基于语料库的工具与

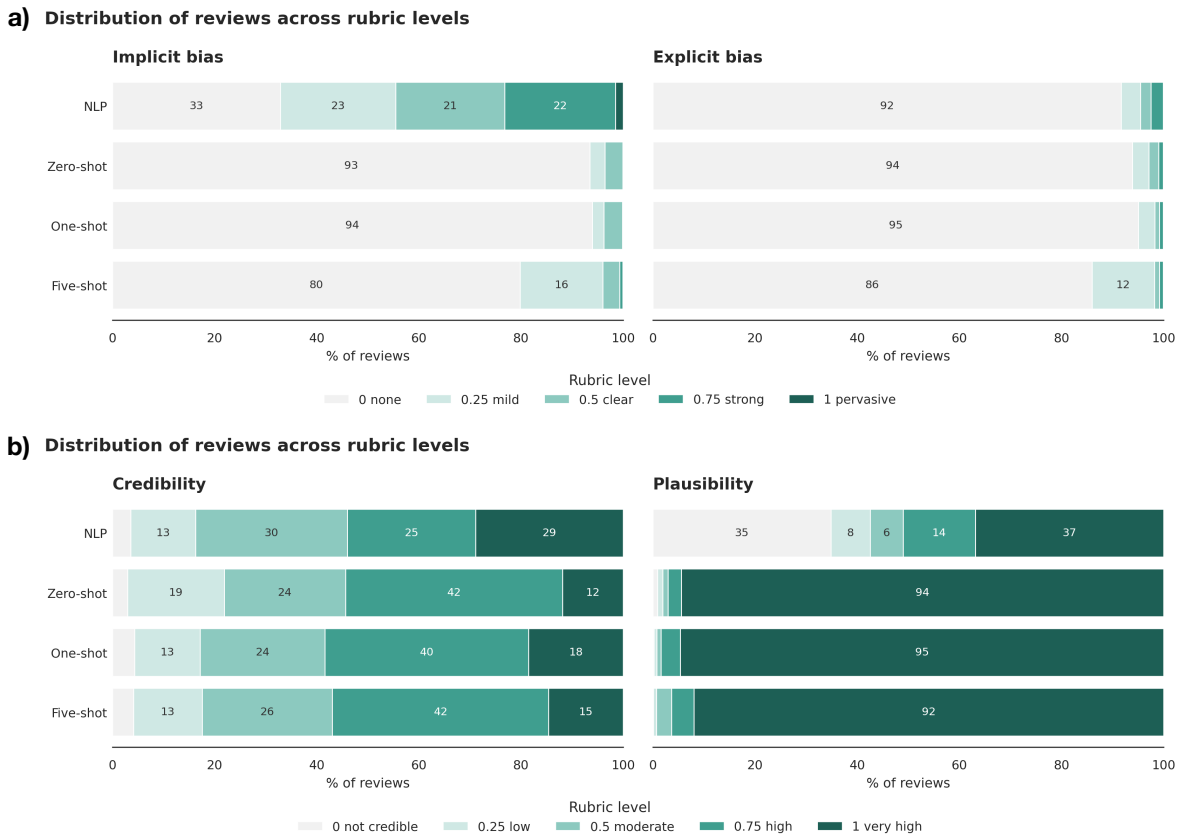


图1：智能体和自然语言处理以截然不同的方式使用同一量表。基于自然语言处理的模型与采用零样本、单样本和五个示例的 Gemma 4 31B 在各量表等级（0 到 1，步长为 0.25）上的评论占比。(a) 偏见。自然语言处理模型将三分之二的评论判定为存在隐性偏见（33% 得分为 0），而智能体认为 93% 至 94% 的评论（零样本和单样本）以及使用五个示例时 80% 的评论不存在隐性偏见；所有方法判定存在显性偏见的情况都很少（86% 至 95% 得分为 0）。(b) 信任。两种方法对可信度的评分都分布在整个量表上，但对合理性的评价差异显著：智能体认为 92% 至 95% 的评论非常合理，而基于自然语言处理的模型将 35% 的评论评为最低等级。柱内数字为百分比；低于 6% 的区段不标注。

智能体之间，两者的隐性偏见平均值分别为 0.341 和 0.025–0.04，合理性平均值分别为 0.526 和最高 0.981。由于样本量较大，即使在 W 可忽略不计的情况下，几乎所有成对差异经 Holm 校正后仍具有统计显著性（补充材料 G）。

3.2 两种方法对可信度的判断一致，但对偏见或合理性的判断并不一致

在单条评论层面，一致程度取决于所测量的维度（图 2）。对于偏见，基于自然语言处理的模型与智能体对评论的排序仅呈较弱的一致性（隐性偏见：斯皮尔曼 $\rho = 0.20$–0.21；显性偏见：$\rho = 0.19$–0.26；总体偏差：$\rho = 0.23$–0.25）。它们给出相同评分的情况也很少（二次加权 $\kappa = 0.06$–0.10，适用于隐性偏见；显性偏见为 0.24–0.27；总体偏差为 0.08–0.12）。对于显性偏见，两种方法的评分一致性略高，因为它们给大多数评论的评分都是 0，但一致性仍然较弱。对于可信度，两者的一致性较高（$\rho = 0.65$-0.67，$\kappa = 0.62$-0.65）。然而，对于合理性，两者并无一致性（$\rho \leq 0.08$）。总体信任度是这两个维度的平均值，因此合理性方面的差距使总体信任度的一致性降至 $\rho = 0.04$-0.05，掩盖了双方在可信度上的高度一致。

基于自然语言处理的模型与智能体模型之间的分歧存在于测量方法，而非提示词之间。三种智能体设计在总体信任度方面高度一致（$\rho = 0.90$-0.93），在总体偏见方面则达到中度至高度一致（$\rho = 0.56$-0.83），其中五样本的一致性最低。较大的分歧十分常见。在38%至41%的评论中，基于自然语言处理的模型对总体偏见的估计至少比智能体高出两个量表等级，而反向情况不足1%。对于信任，方向则相反：智能体对16%至29%的评论的评分至少比基于自然语言处理的模型高出两个等级，而基于自然语言处理的模型在12%的评论中出现这种情况。

这些结果从两个方面回答了第一个研究问题。当信任建立在文本的可观察特征（例如具体细节）之上时，智能体式测量与基于语料库的测量结果一致。而当判断需要进行解释时，例如判断某种概括是否没有根据，或某段经历是否可信时，两者则会出现分歧。

3.3 不同方法对不同语言作出响应

为了理解不同模型为何会以不同方式估计合理性和偏见，我们分析了构成模型决策特征的评论语言特征。我们发现，这种差异具有词汇特征（图3）。为了确定最能体现每种方法判断特点的术语，我们将某种方法所关注的语言与所有评论中的语言进行了比较（Monroe 等人，2008）。对于智能体而言，这指的是作为证据引用的文本；而对于基于自然语言处理的方法（不引用任何内容），则指的是其评分较高的评论。在比较中，我们使用了加权对数优势比 z-分数值，具体如第2.5节所述。

根据结果，在偏见方面，基于自然语言处理的模型对泛化性语言和抱怨性语言较为敏感。其判定为高偏见的评论中，标记出的词语包括从不 ($z \approx 20$)、总是、粗鲁、最差 和荒谬 （图3a）。其中许多只是表达不满的普通说法，而非针对人物的带偏见陈述。相比之下，零样本和单样本智能体会引用对社会群体和地点的提及，例如嬉皮士、贫民区、无家可归者、女孩 以及国籍术语。它仅在 7-8% 的评论中标记这类措辞，这与其较低的流行率估计相符。因此，智能体将偏见定位于关于人们身份的陈述中，而自然语言处理则将其定位于评论者抱怨的强烈程度。有趣的是，五个示例改变了这一情况。五样本智能体在 21% 的评论中引用了带有偏见的措辞，其最具特色的词语变成了最佳 ($z \approx 38$)，其次是有史以来、最佳地点和城市名称。编码方案没有明确界定这类最高级表达是否算作偏见。它将偏见定义为对个体、群体或

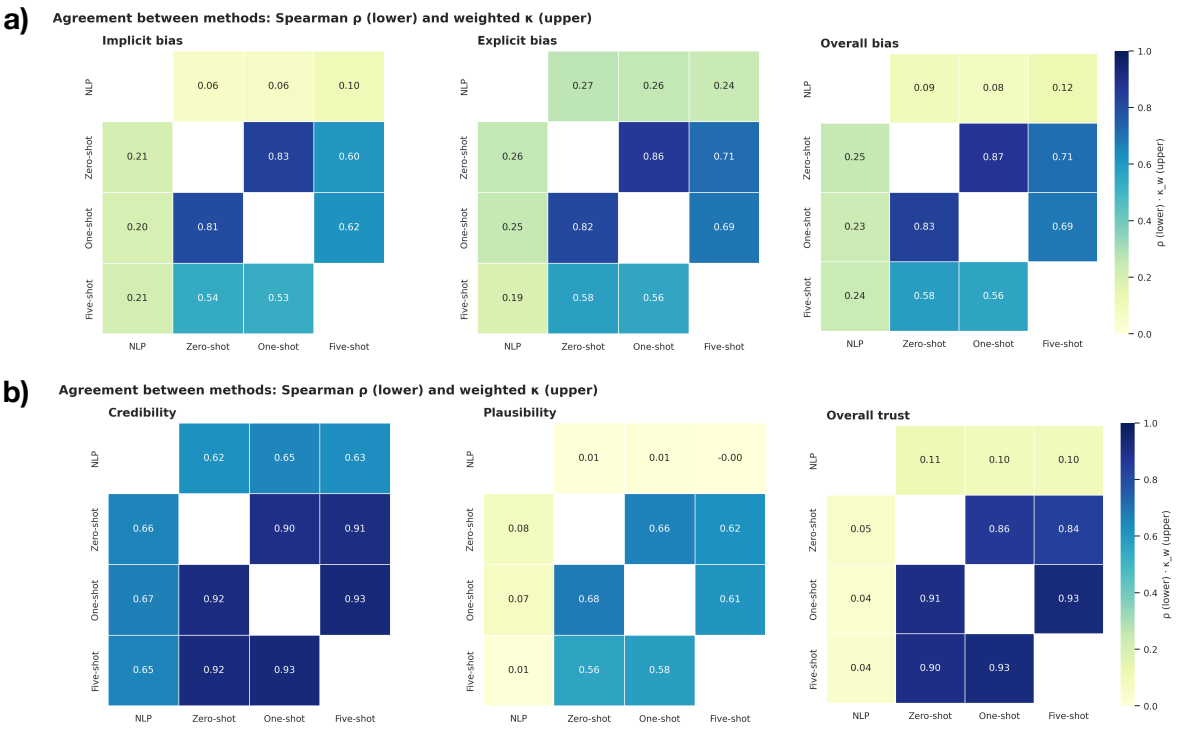


图2：智能体式与基于语料库的测量方法在可信度上达成一致，但在偏见或合理性上并不一致。 所有方法两两之间的评论等级一致性：斯皮尔曼的 ρ（下三角）和量表等级上的二次加权 Cohen κ（上三角）。（a）偏见。基于语料库的工具与智能体在隐性偏见（$\rho = 0.20$–0.21，$\kappa = 0.06$–0.10）、显性偏见（$\rho = 0.19$–0.26，$\kappa = 0.24$–0.27）和总体偏差（$\rho = 0.23$–0.25，$\kappa = 0.08$–0.12）上的一致性较弱。（b）信任。两种方法在可信度上相当一致（$\rho = 0.65$-0.67，$\kappa = 0.62$-0.65），但在合理性上并不一致（$\rho \leq 0.08$），因此总体信任度的一致性仍可忽略不计（$\rho = 0.04$-0.05）。三种智能体设计彼此高度一致（信任 $\rho = 0.90$-0.93；偏见 $\rho = 0.53$-0.83），因此差异源于测量方法之间，而非提示词之间。

实体进行不公平、缺乏依据或泛化的评价；零样本和单样本智能体主要将这一界定用于关于人物和地点的陈述。然而，五个示例中的一个将一条针对餐厅、缺乏依据的最高级评价（“这座城市里最好的披萨，别的都远远比不上”）评为轻微偏见（补充材料 C）。五样本智能体似乎将这一单独案例泛化到了普通赞美；任何针对偏见的少样本测量都应检验这种提示设计效应。

对于信任，两种方法依赖的是性质相反的术语（图3b）。智能体模型将信任建立在具体且可核查的细节之上。在全部评论中，有95%的评论里，每种智能体设计都会引用至少一段文字作为支持性证据，平均每条评论约引用3.5段。在其余评论中，由于当评论不包含具体细节时指令允许不引用，因此不会引用任何内容。与整体评论相比，这些引文中最具特色的术语是菜品（鸡肉、奶酪、虾）以及等待时间（分钟）等具体信息。基于自然语言处理的模型则将高信任度与泛化赞扬联系起来，例如很棒、友好、推荐和绝对。相比之下，智能体会将这类词汇（很棒、最佳和惊艳）恰恰视为危险信号。一种方法将其解读为满意且可信的顾客所发出的信号，另一种方法则将其解读为宣传性或不真实写作的信号。

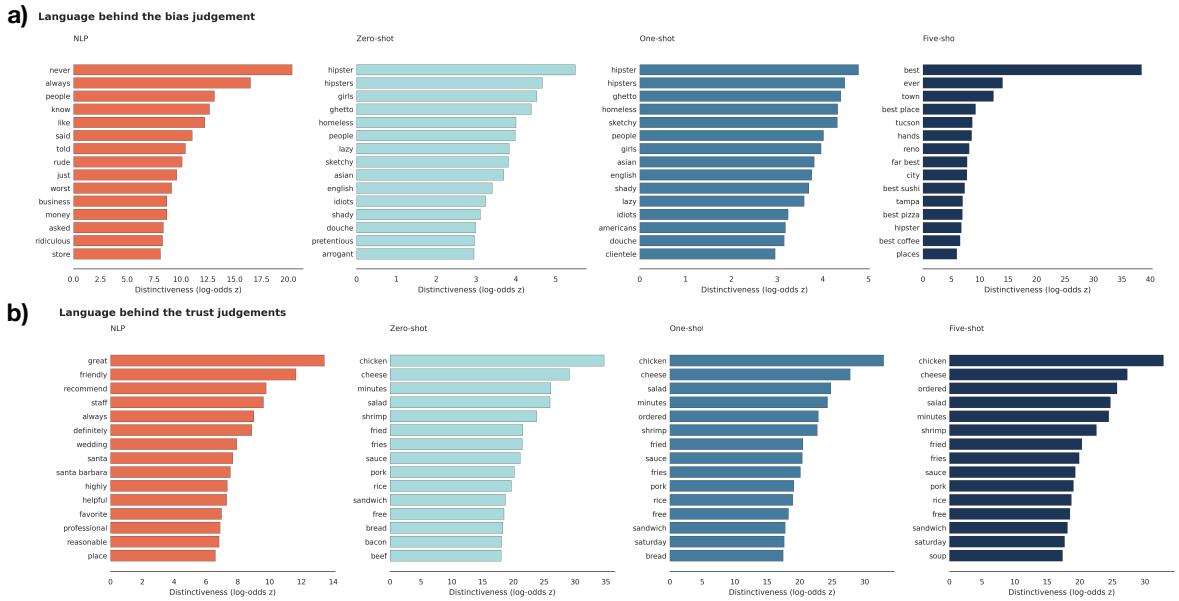


图3：两种方法关注不同的语言。 根据信息性狄利克雷先验中的加权对数优势比 z-得分，对每种方法评价中最具区分性的术语进行排名： z 越高，表示其代表性越强。对于智能体而言，目标文本是它引用的段落；对于不引用任何内容的基于自然语言处理的方法，目标文本则是其评分较高的评论，并与其他评论进行比较。(a) 偏见。基于自然语言处理的方法会关注泛化词和抱怨词（从不、总是、粗鲁、最差），而零样本和单样本智能体则会引用提及社会群体和地点的内容（嬉皮士、贫民区、无家可归者）。在提供五个示例后，智能体会标记最高级表达（最佳、有史以来、最佳地点），并在21%的评论中引用带有偏见的措辞，而低样本模型的比例为7%—8%；这表明示例可以改变其对偏见的判定方式。(b) 信任。智能体模型将信任建立在具体且可核查的细节上（菜品、价格、等待时间），而基于自然语言处理的方法则将高信任度与泛化赞扬（很棒、友好、推荐）联系起来，这些词会被智能体引用为危险信号。缩写形式、功能词以及少于三个字母的术语均被排除；泛化标记（总是、从不）予以保留。

3.4 城市间的偏见几乎没有差异，信任程度也仅略有不同

第二个研究问题探讨这些测量如何塑造城市声誉。总体而言，首要结论是，对于每种方法，城市之间的差异都很小（图4）。就偏见的平均等级而言，城市归属几乎无法解释变异，这一点体现为基于自然语言处理的方法的 Kruskal-Wallis $\eta^2_H \leq 0.003$，以及每种智能体设计的 $\eta^2_H \leq 0.001$ （图4a）。因此，按偏见进行的城市排名在很大程度上反映的是噪声，并且会因方法不同而有所差异。这表明，在任一方法下，城市之间的平均偏见都没有差异。

信任度是城市之间确实存在差异的方面。不同智能体设计下的平均可信度呈现出虽小但始终如一的模式（$\eta^2_H = 0.005$–0.006；图4b）。圣彼得堡在每种设计中的可信度都最低，而费城、新奥尔良和圣路易斯则名列前茅。每座城市约有 10,000 条评论，因此城市均值的标准误约为 0.003；这一模式在三种独立提示设计中反复出现，不太可能是噪声。然而，城市层面只能解释不到百分之一的方差。根据智能体的评估，城市之间的合理性没有差异（$\eta^2_H = 0.000$），因为几乎所有评论都获得了最高分。基于自然语言处理的方法发现，可信度方面的城市差异规模相近且同样很小（$\eta^2_H = 0.004$）；与智能体不同，该方法在合理性方面也发现了差异（$\eta^2_H = 0.003$），但对城市的排序不同。圣巴巴拉

在基于自然语言处理的方法中，总体信任度排名第一，但在智能体评估中仅排第七至第八；新奥尔良在智能体评估中位于前半部分，在基于自然语言处理的方法中则降至第九。由于两种测量工具下的城市差异都很小，即使测量内容略有差异，也足以改变城市排名。因此，研究报告中的城市排名取决于所使用的测量工具。

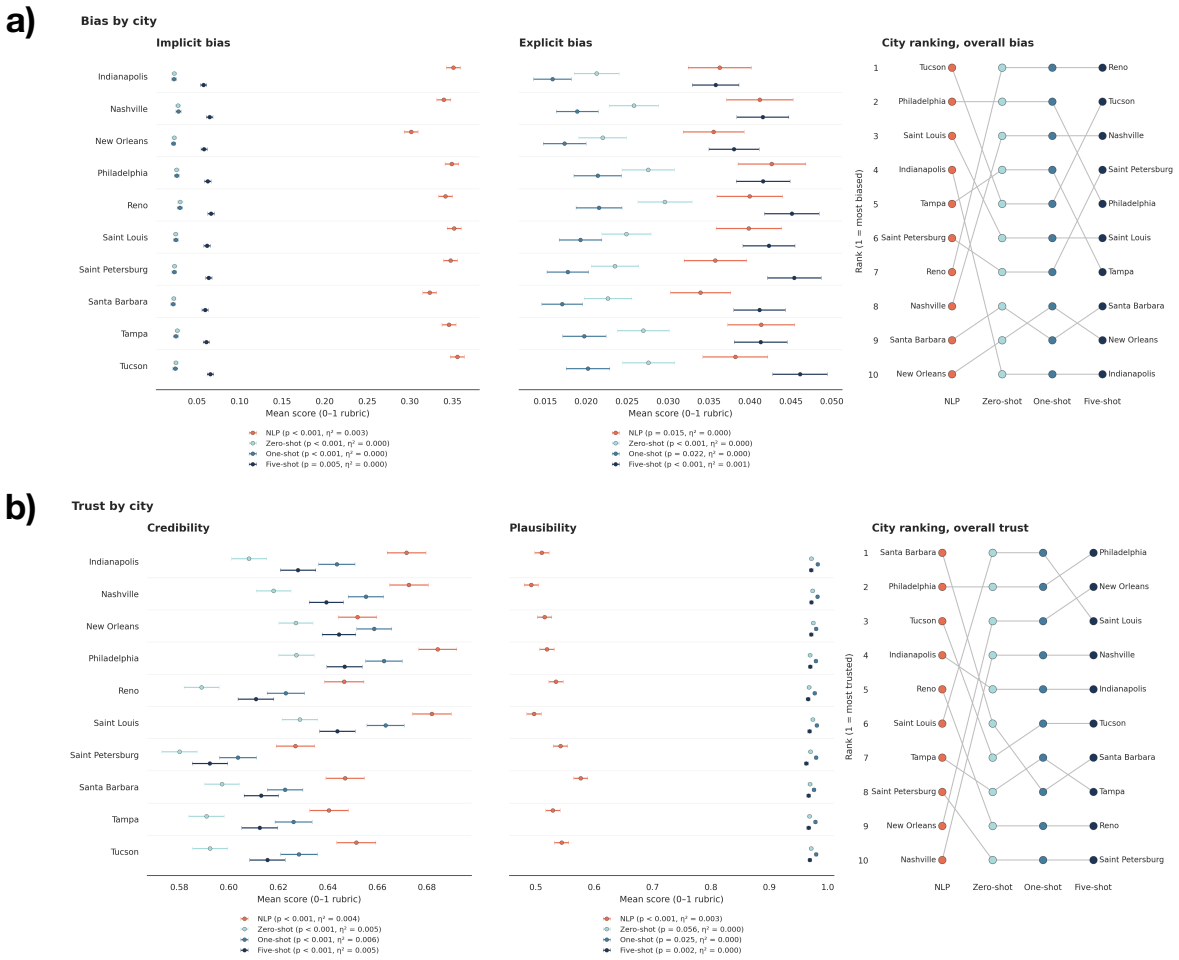


图4：城市在偏见方面几乎没有差异，在信任方面的差异也很小，而且不同方法对城市的排名各不相同。城市均值及95%置信区间（左图和中图），以及按总体得分排列的城市排名（右图；连线表示不同方法中的同一城市）。（a）偏见。城市归属几乎无法解释变异（对于每种方法，Kruskal–Wallis $\eta^2_H \leq 0.003$），因此排名在很大程度上反映的是噪声。（b）信任。不同智能体设计之间的平均可信度呈现出小幅但一致的模式（$\eta^2_H = 0.005$–0.006），其中圣彼得堡最低，费城、新奥尔良和圣路易斯最高。对于该智能体，平均似然性在城市之间没有差异（$\eta^2_H = 0.000$）。基于自然语言处理的方法发现，各城市在可信度（$\eta^2_H = 0.004$）和似然性（$\eta^2_H = 0.003$）方面的差异同样非常小，但对城市的排序不同。对于基于自然语言处理的方法，圣巴巴拉的总体信任度排名第一，但在该智能体中排名第七至第八。排名 1 = 偏见最严重（a）或最受信任（b）的城市。

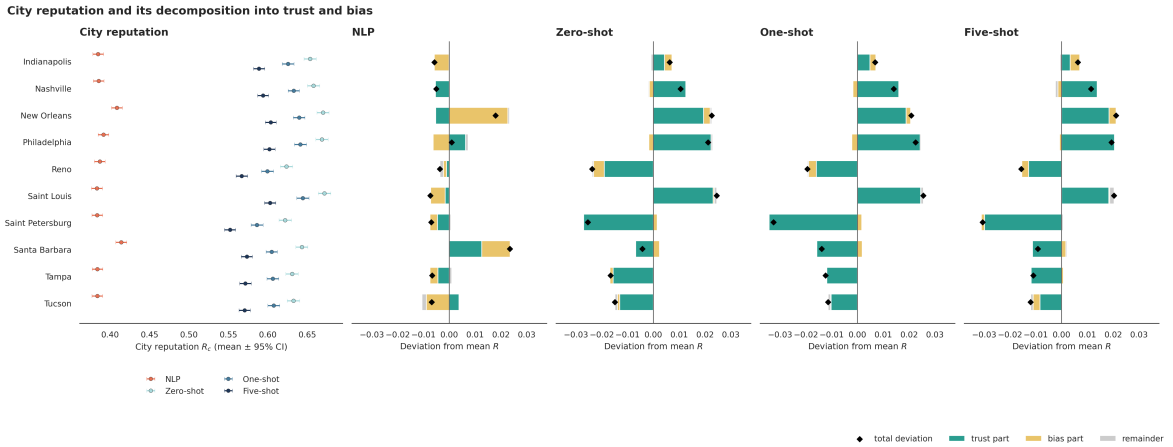


图5：声誉的细微差异源于智能体的信任，以及自然语言处理的偏见。 (a) 各方法的城市声誉 $R_c = \frac{1}{N_c}\sum_{i \in c} t_i (1 - b_i)$（均值 ± 95% 置信区间）。(b)–(e) 各城市相对于平均声誉（菱形）的偏差，精确分解为信任部分、$(\bar{t}_c - \bar{t})(1 - \bar{b})$、偏差部分 $-\bar{t}(b_c - \bar{b})$，以及较小的余项（城市内部的交互作用和协方差）。对于智能体而言，信任约占城市偏差绝对值的 90%；在所有设计中，圣路易斯、费城和新奥尔良均高于均值，而圣彼得堡和里诺均低于均值。对于自然语言处理而言，偏见约占 60%，主要抬高了新奥尔良和圣巴巴拉的声誉，这两个城市是其检测到的隐性偏见最少的城市。没有任何城市与平均声誉相差超过 0.03（量表范围为 0 至 1）；与单条评论中声誉的离散程度相比，这一差距很小。

3.5 城市声誉差异源于不同机制

各方法得出的城市间差异都很小，但基于评论得出的地点排名仍会被发布和使用，即使其基础差异很小。因此，我们关注的不是城市之间是否存在显著差异，而是这两种方法对各城市声誉高低的判断是否一致，以及它们如何解释一个城市为何排名高于另一个城市。将两项指标合并为城市声誉 R_c（公式 1）后可以看出，两种方法得出的等级已经有所不同（图5a）。根据设计不同，智能体将所有城市的得分定在约 0.55 至 0.67 之间，而基于自然语言处理的方法则将得分定在约 0.38 至 0.41 之间，因为它结合了较低的合理性和较高的隐性偏见。方法之间的这一差距，远大于同一方法下各城市之间的任何差异。

为识别城市间差异的驱动因素，我们将每个城市相对于所有评论平均声誉的偏差分解为信任部分、偏差部分和一小部分余项（公式 2；图5b 至 e）。两种方法将声誉归因于不同的机制。对于智能体而言，信任约占城市偏差绝对值的 90%。在三种设计中，圣路易斯、费城和新奥尔良均高于均值，圣彼得堡和里诺则低于均值。对于基于自然语言处理的方法，偏见约占 60%。该方法主要因为检测到新奥尔良和圣巴巴拉的评论中隐性偏见较少，便将这两座城市的声誉推高至均值之上。在 0 到 1 的范围内，没有城市与均值的差距超过 0.03。因此，两种方法得出了不同的声誉地理格局。智能体设计对城市声誉的排名几乎完全一致（斯皮尔曼 $\rho \geq 0.95$），而基于自然语言处理的方法与智能体的排名一致性较弱（均值 $\rho = 0.24$）。某座城市在一种工具下显得声誉良好，是因为另一种工具无法检测到相关特征。

3.6 声誉在不同评论之间的差异大于不同城市之间的差异 s

以标准化评论特征和每个城市的随机截距建立的评论层面声誉多层模型 R_i，进一步印证了此前关于城市本身贡献甚微的发现（图 6）。 R_i 的方差中，介于 0.1% 至 0.4% 的部分来自城市间差异（组内相关系数，组内相关系数 $= 0.001$-0.004，因方法而异）。五项评论特征本身并非偏见或信任分数，却能解释远多于基于自然语言处理方法所得声誉评分的、基于智能体的声誉评分方差（边际 $R^2 = 0.34$–0.40）。这五项特征分别是星级评分、评论收到的有用票数、搞笑票数和酷票数，以及评论的字数；其中投票数和字数以对数形式纳入模型。城市效应，即在考虑这些特征后，各城市评论相对于总体均值的估计偏移，几乎没有增加解释力，因为同时计入这些效应的条件 R^2 几乎等于边际R^2。对于零样本智能体，这些效应在 0 到 1 的范围内，从里诺的 -0.018 到新奥尔良的 $+0.028$（图 6b）；圣路易斯（$+0.009$）和圣巴巴拉（$+0.005$）也高于均值，而 Tampa 和圣彼得堡（-0.009）低于均值。我们考察的四项城市特征，即人口、面积、餐厅数量和餐厅密度，在任何方法中经过 Holm 校正后，均与城市效应无显著关系。与餐厅密度相关的程度最大，各方法的相关值介于 $\rho = 0.31$ 和 0.50 之间，但对于仅有的十座城市，这些相关性无法与偶然结果区分开来（补充材料 F）。

在所选的评论特征中，有一项特征占据主导地位。对三种智能体设计而言，评论长度的 SHAP 贡献均最大，平均约为R 个单位，即 0.12–0.14，并占智能体声誉评分方差的约 35–40%（图 6a、c）。一种可能的解释是，较长的评论包含更多具体细节，而智能体会将这些细节视为可信度证据。对基于自然语言处理的方法而言，星级评分最为重要，这与其依赖泛化赞扬的情感倾向相符。在城市层面，城市特征呈现出相同的情况（图 6b）。由于智能体给较长的评论更高的声誉评分，评论长度高于平均水平的城市，其预测声誉会提高；评论较短的城市则会降低。对于零样本智能体，评论长度使费城的预测声誉提高约 0.02，使圣路易斯提高约 0.017；同时使圣彼得堡降低约 0.017，使圣巴巴拉降低约 0.011。这些贡献与最大的城市效应相当，后者的范围为 -0.018 至 $+0.028$；在费城和圣彼得堡，评论长度的贡献甚至超过了城市效应本身。

这些结果回答了第二个研究问题。每种测量工具都根据评论为每座城市计算声誉，但两种测量工具得出的声誉在等级、排名以及将差异归因于何种机制方面各不相同。在每种测量工具内部，所选城市之间的声誉差异微乎其微。变异主要存在于单条评论之间，并且与评论长度等特征有关；这在意料之中，因为两种测量工具评分的都是评论文本。因此，无论对于平台还是城市营销，基于评论得出的声誉评分都取决于生成该评分的测量工具，不能将其视为城市本身的属性。



图 6：影响智能体声誉评分的是评论长度，而非城市。基于包含标准化评论特征和每个城市随机截距的线性混合模型，分析对评论层面声誉的贡献 R_i 。(a) 全局重要性：各预测变量的 SHAP 值绝对值均值，$\beta\,(x - \bar{x})$。对于三种智能体设计，评论长度的影响均占主导地位（约为 R 的 0.12–0.14 个单位），而星级评分对自然语言处理最为重要。(b) 零样本智能体的城市特征：各城市和预测变量对应的 SHAP 值均值，以及估算的城市效应；根据相似程度对城市进行聚类。(c) 对 R_i 方差的近似分解。评论长度约占智能体声誉方差的 35–40%，而城市仅占 0.1–0.4%（不同方法的组内相关系数 ICC 为 $= 0.001$-0。004）。因此，声誉在不同单条评论之间的差异远大于在所选城市之间的差异。

4 讨论

智能体式和基于语料库的测量工具在可信度方面得出相似结果，但在偏见和合理性方面得出不同结果。此外，在同一种测量工具下，各城市的声誉差异很小，但同一城市的声誉在不同测量工具下却有所不同。目的地形象研究通常将一个地方的形象——即访客如何感知和评价该地，例如对其质量、安全性或真实性的评价——视为该地的一种特征，并认为评论能够揭示这一特征（Marine-Roig 和 Huertas，2020年；Guo 等，2021年）。按照这一观点，更好的测量工具只会更精确地测量同一种形象。我们的研究结果对这一假设提出了质疑。在我们的数据中，测量工具既影响哪些城市显得声誉良好，也影响其原因。

4.1 各测量工具测量的内容

当构念以可观察到的文本为锚点时，两种方法的结果趋于一致。两者都将可信度定义为具体的第一手细节，因此得出了较高程度的一致结果（$\rho \approx 0.66$,$\kappa \approx 0.63$；图2）。当构念需要解释时，两者便产生分歧。对于合理性，两种测量工具用同一个名称测量了不同的构念。基于自然语言处理的模型只评估句子对之间的逻辑一致性，并发现三分之一的评论中存在明显矛盾；而智能体采用更广泛的定义，还将真实性纳入其中，因此几乎没有发现问题。将这一维度与可信度取平均值，合并为一个信任分数，会掩盖可信度方面的高度一致，因为合理性方面并无一致性（图2b）。报告单一可信度或信任指数的评论研究也面临同样的掩盖风险。因此，分别报告各个维度可以避免这一问题。对于偏见，基于语料库的方法会对泛化标记和投诉词汇（图3a）作出反应，而这些线索也会出现在普通的不满表达中。相比之下，智能体会引用提及社会群体和地点的内容，这正是 Zukin 等（2017年）在 Yelp 评论中记录的话语性种族隔离所使用的词汇。综合考察两种方法可见，基于自然语言处理的方法衡量评论者泛化和抱怨的程度，而智能体衡量他们是否对人物身份作出判断。单凭任一结果，都无法说明哪一种才是编码方案所定义的有效偏见测量方式（Jacobs 和 Wallach，2021年）。编码方案将偏见定义为不公平地偏袒或贬损个体、群体或实体的语言，例如针对工作人员、顾客、社区或城市的刻板印象，或根据一次访问得出的笼统判断；同时指出，与所描述经历相关的强烈观点不属于偏见（第2.2节；补充材料 C）。这一界定与相关研究一致：这些研究将语言中的偏见视为对社会群体作出的无根据泛化（Maass，1999年），以及将主观评价呈现为事实的框架（Recasens 等，2013年）。智能体引用的内容提及社会群体和地点，与该定义中列出的刻板印象相符；而粗鲁或最差等投诉则属于该定义所列的例外，即与某种经历相关的观点。然而，从不和总是等泛化标记，也可能表示该定义所认定的偏见性笼统判断。因此，两种方法所回应的语言差异表明，智能体的解读更接近该定义的核心；但要证实这一点，仍需对存在分歧的评论进行人工编码。

信任背后的语言揭示了更尖锐的冲突。智能体以具体细节为信任依据，并将泛化赞扬视为危险信号，这与有关可信度线索的研究一致（Filieri，2016），也与虚假评论所含具体细节少于真实评论的证据相符（Ott等人，2011）。这种基于语料库的自然语言处理方法会考虑细

节、第一人称叙述和篇幅，但不评判其内容，却将同样的赞扬与高信任度联系起来。一个奖励热情表达的工具，可能恰恰偏向平台试图过滤掉的那些评论。

4.2 提示设计是测量的一部分

智能体设计彼此高度一致，因此其结果与自然语言处理的偏离并非任何单个提示所造成的偶然现象。然而，示例会改变智能体对偏见的判定标准。在五个示例的情况下，它会将 最佳和有史以来等最高级表达视为偏见，并估计21%的评论存在偏见，而零样本模型和单样本模型估计的比例为7%–8%（图3a）。这可能是因为五样本模型使用的五个示例中，有一个示例（单样本模型未使用该示例）将一条缺乏依据的最高级表达（“城市里最好的披萨，其他任何披萨都无法望其项背”）评为轻微偏见（补充材料 C）。五样本智能体似乎将这一单一案例泛化到了普通赞誉，这种提示设计效应是任何少样本偏见测量都应进行检验的。该结果与已知的语言模型对少样本示例的敏感性一致（Zhao等，2021年；Lu等，2022年；Sclar等，2024年），并表明示例必须像编码方案本身一样经过严格验证。

4.3 城市为何几乎没有差异

评论层面的声誉变异中，不到百分之零点五来自城市之间的差异（图 6）。这可以由两种机制解释。首先，评论描述的是与单个商家的体验，而这些体验在城市内部的差异远大于城市之间的差异。此外，评论撰写者评价这些体验时，依据的是相对于自身预期的表现，而不是绝对尺度。他们的预期是在当地形成的，受到所支付价格（Abrate 等人，2021年）以及同一商家先前评分的影响；后者会作为其自身评分的锚点（Wang 等人，2022年），因此每条评论都是以当地标准来评价一家商家。以这种方式锚定的评价，很可能在每个城市都围绕相近的平均值变动，从而几乎没有留下城市之间差异的空间。其次，评论撰写是一种具有共同惯例的体裁。在 Tucson 和费城，人们描述食物、服务和价格的方式大体相同。相对而言，变异最大的是评论撰写者写了多少内容；评论长度主导着智能体的声誉评分，因为较长的评论包含更多智能体视为可信度依据的细节。

此处，分析的尺度至关重要。Zukin等（2017年）记录的带有偏见的表述，以及Davis等（2019年）、Moro等（2021年）和Hilman等（2022年）记录的消费隔离，发生在社区之间，并且可能在城市平均值中相互抵消。因此，我们关于城市的零结果并不意味着评论不存在基于地点的偏差。相反，这意味着城市平均值过于粗略，无法将其呈现出来。

4.4 对平台和城市的影响

对于一项从评论中推断地点形象的研究，我们的结果表明，应将工具视为研究对象的一部分。同一批评论会得出不同的城市排名，并将属性声誉归因于不同的机制：使用智能体时归因于信任，使用基于自然语言处理的方法时则归因于偏见（图5）。同时报告两者，并将信任度和偏见分解呈现出来，可以使这些选择一目了然。对于平台和城市营销人员，任何

根据评论计算出的声誉评分，都会承袭其工具内置的价值取向（Orlikowski 和 Scott，2014；Gillespie，2014）。如果此类评分被用于地点排名或内容审核，就应像记录和审计平台估值体系的其他部分一样，对该工具进行记录和审计。

4.5 局限性

这些结论受到若干局限性的约束。本研究涵盖一个平台、十个美国城市和英语评论。智能体基于单一开放权重模型运行，每条评论仅运行一次，且温度较低；其他模型可能会以不同方式划定该构念的边界。此外，语言模型自身携带社会偏见，这些偏见可能影响其对偏见的判断（Gallegos 等，2024年）。提示词中使用的示例是合成的，其中至少有一个改变了五样本智能体对偏见的理解。基于语料库的线索使用了事先固定的阈值，而其他阈值会改变其等级，并可能影响其与智能体的一致性。最后，由于缺乏对相同评论的人工标注，我们可以展示这些测量工具在哪些方面达成一致以及为何产生分歧，但无法判断哪一个更接近人类判断。

每项局限性都指向一个后续步骤。对一组分层样本中的一致评论和分歧评论进行人工编码，将为测量这两种工具的效度提供机会。随机抽样或按商家分层抽样、在邻里尺度上开展分析，以及跨模型、平台和其他城市进行重复验证，都可以检验城市差异近乎不存在这一结果的普遍性。选取真实的、经过人工编码的评论作为示例，并测试每个示例如何改变智能体的判断，将使少样本测量具备可审计性。

5 结论

我们比较了两种用于测量来自十个美国城市的 100,000 条 Yelp 评论中的信任度和偏见的工具：一种是智能体式工具，另一种是基于语料库的工具。两者在可信度这一维度上的结论一致，因为该维度最牢固地锚定于文本之中；但在偏见和合理性方面存在分歧，因为这些方面的判断需要进行解释。它们对不同的语言特征作出反应：基于语料库的自然语言处理关注投诉和概括性判断，而智能体关注关于社会群体的陈述。将结果汇总到城市层面后，两类模型都发现变异很小；剩余的细微差异，一方面源于其中一种工具对信任度的判断，另一方面源于另一种工具对偏见的判断。

在所研究的十个城市中，在线评论因此显示出城市之间的声誉差异很小。研究表明，声誉如何从这些评论中产生，取决于阅读评论的角度。随着平台和城市营销人员越来越多地将评论解读工作委托给语言模型，融入这些模型的定义、示例和检查机制，也就成为地点获得评价的方式的一部分。明确说明这些选择，并在社区层面以人类判断为参照对其进行检验，是迈向声誉测量指标的下一步；这样的指标应当能够揭示地点的真实特征，而不是揭示为其评分的测量工具。

CRediT 作者贡献声明

Rafiazka Hilman：概念化、Methodology、软件、验证、形式分析、调查、资源、数据整理、撰写——初稿、写作——审阅与编辑、可视化、经费获取

J´ulia K oltai：概念化、验证、资源、写作——审阅与编辑

监督、项目管理、经费获取

利益冲突声明

作者声明，他们不存在任何已知的竞争性经济利益或可能被认为影响本文所报告工作的个人关系。

资金来源

Rafiazka Hilman 感谢 Momentum MSCA计划 的支持。该计划由欧盟委员会通过HORIZON-MSCA-2023-COFUND计划以及匈牙利科学院秘书处（MTA）根据资助协议编号101179854共同资助。Julia Koltai 感谢匈牙利科学院 Lend¨ulet 计划的资助：LP2022-10/2022。

参考文献

Ana Babi´c Rosario、Francesca Sotgiu、Kristine De Valck 和 Tammo H. A. Bijmolt。《电子口碑对销售额的影响：关于平台、产品和指标因素的元分析综述》。《营销研究杂志》，53(3)：297–318，2016年。doi：10.1509/jmr.14.0380。

Ana Babi´c Rosario、Kristine de Valck 和 Francesca Sotgiu。电子口碑过程的概念化：关于电子口碑创建、曝光和评价，我们已知和仍需了解的内容。《营销科学学会杂志》，48(3):422–448，2020年。doi: 10.1007/s11747-019-00706-1。

Soonyong Bae 和 Taesik Lee。产品类型与消费者对在线消费者评论的认知。《电子市场》，21(4):255–266，2011年。doi: 10.1007/ s12525-011-0072-0。

Michael Luca。评论、声誉与收入：Yelp.com 案例。工作论文 12-016，哈佛商学院，2016年。

Xinxin Guo、Juho Pesonen 和 Raija Komppula。比较作为目的地形象信息代理的在线旅游评论平台。《信息技术与旅游》，23 (2):159–187，2021年。doi: 10.1007/s40558-021-00201-w。

Estela Marine-Roig 和 Assumpci´o Huertas。安全性如何影响通过在线旅游评论投射的目的地形象。《目的地营销与管理杂志》，18:100469，2020年。doi: 10.1016/j.jdmm.2020.100469。

Sharon Zukin、Scarlett Lindeman 和 Laurie Hurson。杂食者社区？在线餐厅评论、种族与绅士化。《消费者文化杂志》，17（3）：459–479，2017年。doi: 10.1177/1469540515611203。

Ismail Erkan 和 Chris Evans。社交媒体中的电子口碑对消费者购买意愿的影响：一种扩展的信息采纳方法。《人类行为中的计算机》，61：47–55，2016年。doi: 10.1016/j.chb.2016.03.003。

Elvira Ismagilova、Emma Slade、Nripendra P. Rana 和 Yogesh K. Dwivedi。信息源可信度特征对消费者行为的影响：一项元分析。《零售与消费者服务杂志》，53:101736，2020年。

Michael Luca 和 Georgios Zervas。《弄假成真：声誉、竞争与 Yelp 评论欺诈》。《管理科学》，62(12):3412–3427，2016。doi: 10.1287/ mnsc.2015.2304。

Dina Mayzlin、Yaniv Dover 和 Judith Chevalier。《促销性评论：在线评论操纵的实证调查》。《美国经济评论》，104(8):2421– 2455，2014。doi: 10.1257/aer.104.8.2421。

Josef Zelenka、Tracy Azubuike 和 Martina P´askov´a。旅游服务在线评论的信任模型与目的地评价。《行政科学》，11(2):34，2021年。doi: 10.3390/admsci11020034。

Verena Schoenmueller、Oded Netzer 和 Florian Stahl。在线评论的极性：普遍程度、驱动因素及影响。《营销研究杂志》，57(5):853–877，2020年。doi: 10.1177/0022243720941832。C. J. Hutto 和 Eric Gilbert。VADER：一种简约的基于规则的社交媒体文本情感分析模型。载于国际*AAAI*网络与社交媒体会议论文集，第8卷，第216–225页，2014。Daniel Loureiro、Francesco Barbieri、Leonardo Neves、Luis Espinosa Anke 和 Jose Camacho-Collados。TimeLMs：基于Twitter的历时语言模型。载于第*60*届计算语言学协会年会论文集：系统演示，第251–260页，2022年。

Marta Recasens、Cristian Danescu-Niculescu-Mizil 和 Dan Jurafsky。用于分析和检测偏见语言的语言模型。载于计算语言学协会第*51*届年会论文集，第1650–1659页，2013年。

Adina Williams、Nikita Nangia 和 Samuel Bowman。通过推理进行句子理解的广覆盖挑战语料库。载于计算语言学协会北美分会*2018*年会议论文集：人类语言技术，第1112–1122页，2018年。doi：10.18653/v1/N18-1101。

Moritz Laurer、Wouter van Atteveldt、Andreu Casas 和 Kasper Welbers。减少标注，增加分类：运用深度迁移学习和 BERT-NLI 解决监督机器学习的数据稀缺问题。《style id='1'>政治分析》，32(1)：84–100，2024年。doi：10.1017/pan.2023.20。

Aylin Caliskan、Joanna J. Bryson 和 Arvind Narayanan。语言语料库中自动推导出的语义包含类人偏见。《科学》，356(6334):183–186，2017年。doi: 10.1126/science.aal4230。

Su Lin Blodgett、Solon Barocas、Hal Daum´e III 和 Hanna Wallach。语言（技术）即力量：对自然语言处理中“偏见”的批判性综述。载于计算语言学协会第*58*届年会论文集，第5454–5476页，2020年。

Fabrizio Gilardi、MeysamAlizadeh 和 Ma¨el Kubli。ChatGPT 在文本标注任务上的表现优于众包工作者。《美国国家科学院院刊》，120 (30):e2305016120，2023年。doi: 10.1073/pnas.2305016120。

Caleb Ziems、William Held、Omar Shaikh、Jiaao Chen、Zhehao Zhang 和 Diyi Yang。大语言模型能否改变计算社会科学？*ComputationalLinguistics*，50(1):237–291，2024年。doi: 10.1162/coli a 00502。

Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegr- effe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Shashank Gupta、Bod- hisattwa Prasad Majumder、Katherine Hermann、Sean Welleck、Amir Yazdanbakhsh 和 Peter Clark。Self-refine：结合自反馈的迭代优化。发表于*Advances inNeural Information Processing Systems*，第 36 卷，2023年。

Lei Wang、Chen Ma、Xueyang Feng、Zeyu Zhang、Hao Yang、Jingsen Zhang、Zhiyuan Chen、Jiakai Tang、Xu Chen、Yankai Lin、Wayne Xin Zhao、Zhewei Wei 和 Jirong Wen。《基于大语言模型的自主智能体综述》。《计算机科学前沿》，18(6):186345，2024年。doi: 10.1007/s11704-024-40231-1。

Jay DeYoung、Sarthak Jain、Nazneen Fatema Rajani、Eric Lehman、Caiming Xiong、Richard Socher 和 Byron C. Wallace。ERASER：用于评估合理化自然语言处理模型的基准。载于计算语言学协会第*58*届年会论文集，第4443–4458页，2020年。doi: 10.18653/v1/2020.acl-main.408。Tianyu Gao、Howard Yen、Jiatong Yu 和 Danqi Chen。使大语言模型能够生成带引文的文本。载于*2023*年自然语言处理实证方法会议论文集，第6465–6488页，2023年。doi: 10.18653/v1/2023.emnlp-main.398。Paul Resnick、Ko Kuwabara、Richard Zeckhauser 和 Eric Friedman。声誉系统。《*ACM*通讯》，43(12):45–48，2000年。doi: 10.1145/355112.355122。Judith A. Chevalier 和 Dina Mayzlin。口碑对销售额的影响：在线图书评论。《营销研究杂志》，43(3):345–354，2006年。doi: 10.1509/jmkr.43.3.345。Wanda J. Orlikowski 和 Susan V. Scott。评价转移到线上后会发生什么？探索旅游业中的价值评估装置。《组织科学》，25(3):868–891，2014。doi: 10.1287/orsc.2013.0877。Jos´e van Dijck、Thomas Poell 和 Martijn de Waal。《平台社会：连接世界中的公共价值》。牛津大学出版社，纽约，2018年。Tarleton Gillespie。算法的相关性。载于 Tarleton Gillespie、Pablo J. Boczkowski 和 Kirsten A. Foot 编，《媒体技术：关于传播、物质性与社会的论文》，第167–194页。麻省理工学院出版社，马萨诸塞州剑桥，2014年。Mihalis Kavaratzis 和 Mary Jo Hatch。地方品牌的动态：基于身份的地方品牌理论方法。《营销理论》，13(1):69–86，2013年。doi: 10.1177/1470593112467268。Edward L. Glaeser、Hyunjin Kim 和 Michael Luca。绅士化的即时预测：利用Yelp数据量化社区变化。《美国经济学会论文集与会议录》，108:77–82，2018年。doi: 10.1257/pandp.20181034。Donald R. Davis、Jonathan I. Dingel、Joan Monras 和 Eduardo Morales。城市消费的隔离程度如何？《政治经济学杂志》，127(4):1684–1738，2019年。doi: 10.1086/701680。Esteban Moro、Dan Calacci、Xiaowen Dong 和 Alex Pentland。流动模式与美国大城市中实际经历的收入隔离相关。《自然通讯》，12:4633，2021年。doi: 10.1038/s41467-021-24899-8。Rafiazka Millanida Hilman、Gerardo I˜niguez 和 M´arton Karsai。美国大都市区城市混合模式中的社会经济偏差。《*EPJ*数据科学》，11:32，2022年。doi: 10.1140/epjds/s13688-022-00341-x。Raffaele Filieri。是什么让在线消费者评论具有可信度？《旅游研究年鉴》，58:46–64，2016年。doi: 10.1016/j.annals.2015.12.019。

Miriam J. Met$_{zger,}$ Andrew J. Flanagi^{n,} 一个d Ryan B. Medders. Social and heuristi
在线可信度评价的……方法。《传播学杂志》，60(3):413–439，2010年。doi: 10.1111/j.1460-2466.2010.01488.x。Sherry He、Brett Hollenbeck 和 Davide Proserpio。虚假评论市场。*Mar-keting Science*，41(5):896–921，2022年。Myle Ott、Yejin Choi、Claire Cardie 和 Jeffrey T. Hancock。无论如何都能发现欺骗性意见垃圾信息。载于第*49*届计算语言学协会年会：人类语言技术论文集，第309–319页，2011年。Hong Hong、Di Xu、G. Alan Wang 和 Weiguo Fan。理解在线评论有用性的决定因素：一项元分析研究。《决策支持系统》，102:1–11，2017年。doi: 10.1016/j.dss.2017.06.007。Nikhil Garg、Londa Schiebinger、Dan Jurafsky 和 James Zou。词嵌入量化了100年来的性别和族裔刻板印象。《美国国家科学院院刊》，115(16):E3635–E3644，2018年。doi: 10.1073/pnas.1720347115。Thomas Davidson、Debasmita Bhattacharya 和 Ingmar Weber。仇恨言论和辱骂性语言检测数据集中的种族偏见。载于第三届在线辱骂性语言研讨会论文集，第25–35页，2019年。Maarten Sap、Dallas Card、Saadia Gabriel、Yejin Choi 和 Noah A. Smith。仇恨言论检测中的种族偏见风险。载于第*57*届计算语言学协会年会论文集，第1668–1678页，2019年。doi: 10.18653/v1/ P19-1163。Dirk Hovy 和 Shrimai Prabhumoye。自然语言处理中的五种偏见来源。《语言与语言学指南》，15(8):e12432，2021年。doi: 10.1111/lnc3.12432。Abigail Z. Jacobs 和 Hanna Wallach。测量与公平性。载于*2021*年*ACM*公平性、问责制与透明度会议论文集，第375–385页，2021年。doi: 10.1145/3442188.3445901。Justin Grimmer、Margaret E. Roberts 和 Brandon M. Stewart。《文本即数据：机器学习与社会科学的新框架》。普林斯顿大学出版社，普林斯顿，新泽西州，2022年。Steve Rathje、Dan-Mircea Mirea、Ilia Sucholutsky、Raja Marjieh、Claire E. Robertson 和 J ay J. Van Bavel。GPT是一种有效的多语言心理学文本分析工具。《美国国家科学院院刊》，121(34):e2308950121，2024年。doi: 10.1073/pnas.2308950121。Lisa P. Argyle、Ethan C. Busby、Nancy Fulda、Joshua R. Gubler、Christopher Rytting 和 David Wingate。由一到多：使用语言模型模拟人类样本。《政治分析》，31(3):337–351，2023年。doi: 10.1017/pan.2023.2。Christopher A. Bail。生成式人工智能能否改善社会科学？《美国国家科学院院刊》，121(21):e2314021121，2024年。doi: 10.1073/pnas.2314021121。

赵子豪、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。使用前校准：提升语言模型的小样本性能。载于第*38*届国际机器学习会议论文集，第139卷《机器学习研究》，第12697–12706页，2021年。

Yao Lu、Max Bartolo、Alastair Moore、Sebastian Riedel 和 Pontus Stenetorp。Fantastically ordered prompts and where to find them：克服少样本提示顺序敏感性。载于*Proceedings of the 60th Annual Meeting of the Association for Compu-tational Linguistics*，第 8086–8098 页，2022年。

Melanie Sclar、Yejin Choi、Yulia Tsvetkov 和 Alane Suhr。量化语言模型对提示设计中伪特征的敏感性，或：我是如何学会开始担心提示格式的。载于国际学习表征会议，2024年。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、EricP. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。使用 MT-Bench 和 Chatbot Arena 评审大语言模型评审。载于*Advances in Neural Information Processing Systems*，第 36 卷，2023年。

Jie Huang、Xinyun Chen、Swaroop Mishra、Huaixiu Steven Zheng、Adams Wei Yu、Xinying Song 和 Denny Zhou。大语言模型尚无法进行自我纠正推理。载于国际学习表征会议，2024年。

Ryo Kamoi、Yusen Zhang、Nan Zhang、Jiawei Han 和 Rui Zhang。LLM 何时才能真正纠正自身错误？关于 LLM 自我纠正的批判性综述。计算语言学协会会刊，12：1417–1440，2024年。doi：10.1162/tacl a 00713。

Yelp 公司 Yelp 开放数据集。 https://www.yelp.com/dataset，2026年。访问日期：[DAT E].

Gemma 团队。Gemma 4 技术报告，2026年。

Ada S. Lo 和 Sharon Siyu Yao。是什么使酒店在线评论具有可信度？对评论者专业性、评论评分一致性和评论效价作用的研究。《国际当代酒店管理杂志》，31(1)：41–60，2019年。doi：10.1108/IJCHM-10-2017-0671。

Zoey Chen 和 Nicholas H.Lurie。在线口碑影响中的时间邻近性与负面偏见。《营销研究杂志》，50(4):463–476，2013年。

Nipun Sadvilkar 和 Mark Neumann。PySBD：实用的句子边界消歧。在《第二届自然语言处理开源软件研讨会 (NLP-OSS)》中，第 110–114 页，2020年。

Matthew Honnibal、Ines Montani、Sofie Van Landeghem 和 Adriane Boyd。spaCy：Python 中的工业级自然语言处理。Zenodo，2020年。

Laura Hanu 和 Unitary 团队。Detoxify。 https://github.com/unitaryai/detoxify，2020。

Daniel Borkan、Lucas Dixon、Jeffrey Sorensen、Nithum Thain 和 Lucy Vasserman。用于利用真实数据测量文本分类中意外偏差的细致指标。载于*2019*年世界互联网大会论文集，第491–500页，2019年。G¨un R. Semin 和 Klaus Fiedler。语言类别在描述人物时的认知功能：社会认知与语言。《人格与社会心理学杂志》，54(4):558–568，1988年。Anne Maass。语言群体间偏差：通过语言延续刻板印象。载于 Mark P. Zanna 编，《style id='6'>《实验社会心理学进展》》，第31卷，第79–121页。Academic Press，1999年。Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生BERT网络的句子嵌入。载于*2019*年自然语言处理实证方法会议暨第九届国际自然语言处理联合会议论文集，第3982–3992页，2019年。Jacob Cohen。加权κ：在允许分级不一致或部分得分的情况下衡量名义尺度一致性。《心理学公报》，70(4):213–220，1968年。doi：10.1037/h0026256。J. Richard Landis 和 Gary G. Koch。分类数据的观察者一致性测量。《生物统计学》，33(1):159–174，1977年。doi：10.2307/2529310。Andr´as Vargha 和 Harold D. Delaney。对 McGraw and Wong 的 CL 通俗语言效应量统计量的批评与改进。《教育与行为统计学杂志》，25(2):101–132，2000年。doi：10.3102/10769986025002101。Sture Holm。一种简单的序贯拒绝多重检验程序。《斯堪的纳维亚统计学杂志》，6(2):65–70，1979年。Burt L. Monroe、Michael P. Colaresi 和 Kevin M. Quinn。斗争性词语：用于识别政治冲突内容的词汇特征选择与评价。《政治分析》，16(4):372–403，2008年。doi：10.1093/pan/mpn018。Maciej Tomczak 和 Ewa Tomczak。重新审视报告效应量估计的必要性：若干推荐效应量指标概述。《体育科学趋势》，21(1):19–25，2014年。Evelyn M. Kitagawa。两个率之差的组成部分。《美国统计学会会刊》，50(272):1168–1194，1955年。Prithwis Das Gupta。《率的标准化与分解：用户手册》。《当前人口报告》，系列P23-186。美国人口普查局，华盛顿特区，1993年。Andrew Gelman、Jennifer Hill 和 Aki Vehtari。《回归与其他故事》。剑桥大学出版社，剑桥，2020年。Mark L. Bryan 和 Stephen P. Jenkins。国家效应的多层次建模：一个警示案例。《欧洲社会学评论》，32(1):3–22，2016年。

Cora J. M. Maas 和 Joop J. Hox。多层次建模的充分样本量。*Methodology*，1(3):86–92，2005年。doi: 10.1027/1614-2241.1.3.86。Martin Elff、Jan Paul Heisig、Merlin Schaeffer 和 Susumu Shikano。少量聚类下的多层次分析：改进基于似然的方法，以提供无偏估计和准确推断。*British Journal of Political Science*，51(1):412–426，2021年。Daniel M. McNeish 和 Laura M. Stapleton。极少聚类情况下的聚类数据建模。*Multivariate Behavioral Research*，51(4):495–518，2016年。A. Colin Cameron、Jonah B. Gelbach 和 Douglas L. Miller。基于自助法的改进方法，用于对聚类误差进行推断。*The Review of Economics and Statistics*，90 (3):414–427，2008年。Shinichi Nakagawa 和 Holger Schielzeth。从广义线性混合效应模型中获得 R^2 的一种通用且简单的方法。*Methods in Ecology and Evolution*，4(2): 133–142，2013年。doi: 10.1111/j.2041-210x.2012.00261.x。Scott M. Lundberg 和 Su-In Lee。解释模型预测的统一方法。载于*Advances in Neural Information Processing Systems*，第30卷，2017年。Graziano Abrate、Sarah Quinton 和 Rebecca Pera。支付价格与酒店评论评分之间的关系：期望失验还是安慰剂效应? *TourismManagement*，85:104314，2021年。doi: 10.1016/j.tourman.2021.104314。Qian Wang、Michael Chau、Chih-Hung Peng 和 Eric W. T. Ngai。运用锚定效应和文化维度理论研究顾客的在线评分行为。*Information SystemsFrontiers*，24(5):1451–1463，2022年。doi: 10.1007/ s10796-021-10148-2。Isabel O. Gallegos、Ryan A. Rossi、Joe Barrow、Md Mehrab Tanjim、Sungchul Kim、Franck Dernoncourt、Tong Yu、Ruiyi Zhang 和 Nesreen K. Ahmed。大语言模型中的偏见与公平性：综述。*Computational Linguistics*，50(3):1097–1179，2024年。doi: 10.1162/coli a 00524。