arXiv:2305.01615v1 [cs.HC] 2 May 2023

Judgment Sieve：通过针对模糊性与分歧的干预措施降低群体判断中的不确定性

QUAN ZE CHEN，美国华盛顿大学AMY X. ZHANG，美国华盛顿大学

当一群人被要求作出判断时，往往会出现不确定性问题。现有的降低不确定性的方法通常侧重于通过迭代方式提高总体任务说明的具体性。然而，不确定性可能源于多个来源，例如，由于上下文有限，被判断项目本身存在模糊性；或者由于参与者拥有不同视角以及任务说明不够明确而产生分歧。如果没有针对正确的不确定性来源，统一方法可能会收效甚微。本文介绍了一种新的工作流程 Judgment Sieve，旨在以针对性的方式降低涉及群体判断的任务中的不确定性。通过利用在初始判断 elicitation 轮次中区分不同不确定性来源的测量结果，我们可以选择针对性干预措施，为每个被判断项目添加上下文或审议，从而最有效地降低不确定性。我们在两项任务上测试了该方法：评定词语对的相似度以及在线评论的毒性。结果表明，定向干预措施降低了最不确定案例中的不确定性。在排名前 10% 的案例中，对于这两项任务，我们分别观察到模糊性降低了 21.4% 和 25.7%，分歧降低了 22.2% 和 11.2%。此外，我们通过模拟发现，与统一方法相比，我们的针对性方法降低了两种不确定性来源的平均不确定性得分；而在统一方法中，一个来源的平均不确定性降低往往伴随着另一个来源的不确定性增加。

CCS 概念： • **以人为中心的计算 → 协作交互**。

补充关键词与短语：众包；标注；模糊性；校准

ACM 引用格式：
Quan Ze Chen 和 Amy X. Zhang。2023。Judgment Sieve：通过针对模糊性与分歧的干预措施，减少群体判断中的不确定性。载于*CSCW '23*：第 *26* 届 *ACM* 计算机支持的协同工作与社会计算会议，线上举行。ACM，纽约，美国，26 页。https://doi. org/10.1145/nnnnnnn.nnnnnnn

1 引言

不确定性是许多涉及对项目作出判断的任务中无法避免的挑战。尤其是，涉及群体的判断往往必须应对不确定性，因为群体环境中的不确定性既可能源于群体中个体所经历的不确定性，也可能产生于群体层面。群体中的每个人都可能因所判断项目的模糊性 而感到一定程度的不确定，难以独自作出判断。与此同时，即使每个人对自己的个人判断都很确定，群体层面的不确定性仍可能源于群体成员之间的分歧 ； 这种分歧可能来自群体成员视角的差异，而任务说明中并未通过明确规定来解决这些差异。

允许免费制作本作品的全部或部分内容的数字副本或纸质副本，供个人或课堂使用，前提是这些副本不得用于或分发以营利或商业利益为目的，并且副本须包含本声明及首页的完整引文。对于本作品中归 ACM 以外其他方所有的部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新出版、发布至服务器或重新分发至列表，须事先获得具体许可和／或支付费用。请通过 permissions@acm.org 申请许可。*CSCW '23*，美国明尼苏达州明尼阿波利斯© 2023 计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

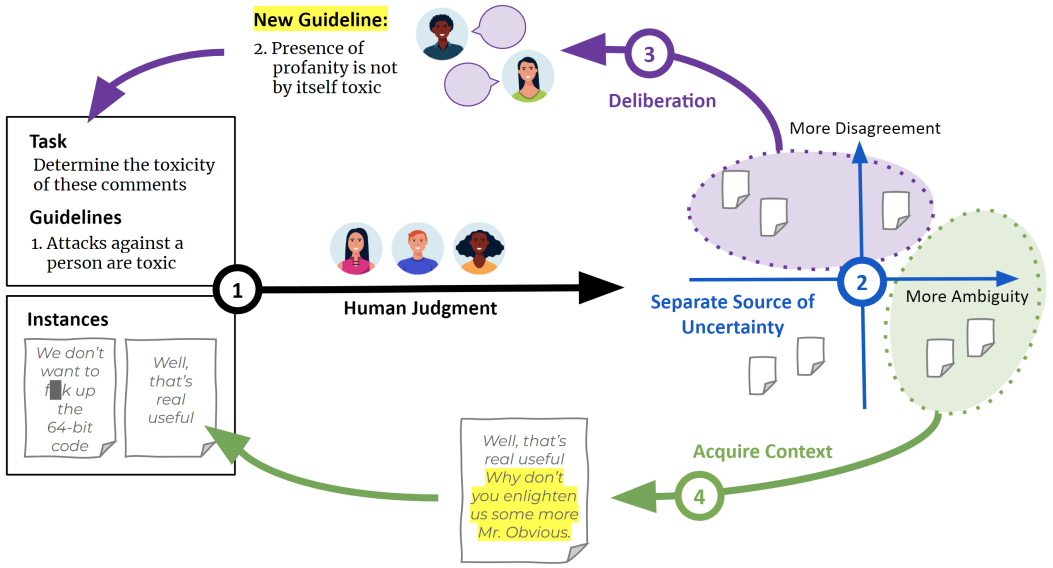


图 1. 工作流程概览：（1）使用能够量化不同不确定性来源的标注工具收集人类判断；（2）针对每个实例，计算对应于不确定性来源（即模糊性 和 分歧）的分数；（3）对分歧较大的实例进行审议干预，以解决分歧并制定新指南；（4）为模糊性较大的实例收集上下文信息，并将其纳入实例中。

例如，一群社区版主在判断是否应删除有害帖子时，可能会因用语的模糊性以及发帖者的意图而感到不确定 [66]。与此同时，背景和文化方面的差异 [45] 意味着社区成员往往会对什么算有害 [3] 以及应采取何种后续行动 [2]产生分歧。同样，在教育场景中，助教和教师在为开放式作业评分时也经常面临不确定性。评分的目标是评估学生的理解程度，但设计不佳的作业题目可能导致答案无法明确表明学生是否理解。另一方面，如果共用的评分量规没有说明遇到特定情况时该如何处理，不同评分者最终可能会依赖个人判断，从而造成分歧和不一致 [75]。

在人类标注过程中未能考虑不确定性，可能导致测量结果不可靠且不一致 [89]，即使是在涉及专家判断的领域中也是如此 [16]。此外，如果未妥善考虑群体成员不同的背景和视角所造成的偏差，也可能导致群体判断产生偏差 [71, 72]。基于这些观察，许多领域已开发出测量数据中不确定性的方法和流程，以剔除不可靠的判断 [6, 34, 67]，或创建能够利用细分数据的系统 [27, 51, 87]。虽然考虑现有的不确定性对于构建更稳健的流程十分重要，但在实际做出决策时，许多情况下都需要通过干预来减少不确定性。

一种减少不确定性的干预措施，是通过提供更多上下文来消除待判断项目中的模糊性，从而帮助人们作出决定。例如，研究表明，提供评论的父帖子等信息，有时会导致对其毒性的判断完全相反 [66]。然而，上下文十分复杂，要收集所有相关上下文

涵盖的所有情况，可能很困难 [78]。因此，通常难以提前为所有实例收集上下文。另一种减少不确定性的方法，是应对群体中个体判断之间的分歧。为解决分歧，人们提出了多种方法，例如通过使用锚定示例 [13]，让人们对任务的理解有据可依；以及利用测得的不确定性来找出含糊不清的指南 [56]，从而进一步明确任务要求。另一些方法则针对视角不同这一根本问题；研究表明，审议有助于提高共识 [12, 25, 74]。然而，这些干预措施的成本也很高，通常需要同步协作。

这些干预措施不仅成本高昂，如果不确定性的根源另有所在，却采用旨在解决某一种不确定性的干预措施，就可能白费力气。例如，先前研究发现，当上下文模糊或缺失时，通过审议解决分歧可能会失败 [74]。另一方面，如果群体成员对自己的判断很确定，但仍需要解决分歧，提供更多上下文可能并无帮助。与其在每个出现不确定性的案例中都采用这些成本高昂的干预措施，不如测量并区分每个案例中导致群体不确定性的来源，从而针对每个案例选择更有针对性且更有效的干预措施。

本文提出了一种名为 Judgment Sieve 的新工作流程，用于高效降低群体判断中的不确定性。Judgment Sieve 采用一种决策过程，根据每个项目初始标注期间观察到的不确定性类型选择针对性干预措施（图 1）。当检测到个体不确定性时，我们会着重获取更多上下文，以减少项目中的模糊性；当检测到标注者之间存在分歧时，我们会着重让标注者参与审议，以协调彼此不同的视角，并更好地明确任务说明。

本文的贡献如下：

- 我们提出了 Judgment Sieve，这是一种工作流程，通过利用与每个实例的模糊性和分歧相关的测量结果，降低群体判断场景中的不确定性。我们还为标量评分任务提供了这一工作流程的原型实现。
- 我们针对两个标量评分任务领域开展了标注：词对相似度评分（*wordsim*）和毒性评分（*toxicity*），并验证了添加上下文和开展审议分别是**减少模糊性和分歧的有效干预措施**。
  - 在最模糊的 10% 案例中，我们观察到，引入上下文使模糊性降低了 21.4%（*wordsim*）和 25.7%（*toxicity*）。– 同样，对于分歧最高的 10% 案例，我们发现，引入根据审议制定的指南使分歧降低了 22.2%（*wordsim*）和 11.2%（*toxicity*）。
- 然而，我们还观察到，对所有项目广泛应用干预措施总体上可能会在某些情况下增加不确定性；添加上下文使分歧增加了 2.06%（*wordsim*）和 3.54%（*toxicity*）。
- 我们开展了一项模拟实验，以评估 Judgment Sieve 的针对性干预功能。该功能会根据初始标注中测得的不确定性类型来选择干预措施。我们发现，与统一方法相比，针对不确定性最高的实例所采取的干预措施能够降低两种不确定性来源的总体均值；而在统一方法中，一种不确定性来源的降低会伴随另一种来源的上升。不过，我们也注意到，如果将动态方法未分配任何干预的实例纳入统计，这种降低在统计上并不显著。

2 相关工作

机器学习和社会计算领域日益认识到，对群体判断中的不确定性进行考量并加以处理，是一项亟待解决的重要问题。

本节将回顾这方面的先前研究，重点包括：（1）阐明错误与不确定性之间的区别；（2）了解现有系统和工作流程如何运用（汇总）不确定性度量；（3）探讨区分不确定性来源的一些理论框架；（4）讨论先前关于上下文和审议的研究，以及这些研究如何为我们设计用于应对不确定性来源的干预措施提供参考。

2.1 错误与不确定性

在众包情境中的群体判断里观察到的不确定性，历来被归因于错误。由于众包任务具有时间限制，且众包工作者的专业知识和注意力有限，错误确实时常发生 [44, 76]。因此，许多旨在减少不确定性的先前研究，都着重于减少错误 的发生及其对判断结果的影响。一种观点认为，错误源于说明不充分，导致工作者感到困惑，进而在任务中出错 [63, 92]。另一些研究则指出，通过更好的培训 [29]或筛选 [54] 工作者，可以减少那些仅凭文本和示例难以说明的复杂任务中的错误。还有研究探讨了如何让工作者保持专注 [4]，以便在较长任务期间 维持质量 [37]。从劳动激励的角度来看，一些研究将错误 问题视为恶意行为（即垃圾信息）所致 [64, 85],，因此提出了基于激励的解决方案 [22]。最后，一种关注数据（而非任务）的观点提出，可以利用模型事后纠正错误，通过冗余标注推断工作者质量 [19]，或根据已知答案的标准答案题目评估其表现 [91]。

然而，当这种以错误为中心的不确定性观被应用于对人类判断日益增长的新需求时，其局限性便显现出来：人们需要对某些案例作出判断，而对于何为真实情况可能并不存在共识。事实上，将错误减少机制应用于复杂的、具有社会情境的，甚至是主观的判断，已经导致下游模型出现问题 [32]。人口统计偏差 [21] 或认知偏差 [38] 等问题无法像错误一样通过“修补”来消除。实际上，在许多此类情况下，即使引入专业知识并消除时间限制，也无法完全缓解所有不确定性 [1, 24, 65]。

上述情况促成了本研究。在本研究中，我们认识到，尽管错误 仍然是观察到的判断不确定性的一部分，但一个日益重要的挑战还在于：如何考量并非由标注者错误 造成，而是源于任务其他因素的不确定性来源，例如模糊案例 [5]，或由于标注者具有不同的视角和背景而产生的合理分歧 [71]。因此，本研究并不寻求取代传统的错误缓解，而是旨在改进我们应对非错误 不确定性来源的方式。

2.2 对人类判断中的不确定性进行考量

如今，对不确定性最直接的看法是将其视为一种筛选标准——如果标注者意见不一致，那么可靠判断就未能达成。按照这种看法，解决方案着重于衡量和评估分歧，例如使用评审者间信度指标 [36]。如果一致性过低，数据可能会被丢弃，或者增加标注者，直到一致性达到令人满意的水平 [70, 96]。因此，达到一定的标注者一致性阈值被视为数据集质量的证明 [76]。然而，正如已有观察所示，待判断的实例本身可能处于一个连续谱上，其可判断的确定程度各不相同 [81, 89],；因此，舍弃示例或过度增加标注者可能会导致偏差，而由于数据集构建过程的透明度有限，这些偏差可能会被忽视 [31]。

测量或不确定性的直接引出，也被提出作为构建任务本身的一种方式。通过让标注者也参与其中的流程，共同完善含糊不清的指南或定义不充分的任务 [9, 11, 46, 55]。

关于不确定性的最后一种观点，着眼于人类判断结果的最终应用，并提出应将不确定性原样传递给下游模型。例如，已有系统被用于从不确定标签中学习 [93, 94] ；在机器学习领域，越来越多的研究致力于利用未聚合的标签作为模型的训练数据，以实现更高的性能 [27, 42]。不确定标签的应用并不限于培训，它们也被用于评估 [34]，以便对性能作出更贴近现实的评估。

然而，可能不存在一种适用于所有情况的不确定性处理方法，即使使用不确定性数据训练的系统，也仍可能缺乏社会认知能力 [7]。即使在同一个数据集中，不同的不确定性来源也不应一概而论。因此，我们的研究关注到近期的一种认识：要以最优方式应对不确定性，就应对其分类并分别量化。

2.3 不同的不确定性来源

传统上，不确定性的量化采用统计学视角，通过呈现并推断根据所收集的判断得出的概率分布 [17] 。基于这一思路，近期研究致力于通过捕捉整个回答集合上的分布 来量化不确定性。然而，生成答案分布的成本很高 [14] ，而且仅凭分布未必能深入揭示哪些因素可能导致了观察到的结果，因此还需要进一步分析和数据收集 [89]。

因此，近年来，已有研究尝试分别归类 不确定性的不同来源。受早期统计学中有关测量的研究启发 [39, 48], ，其中一个框架提出了以下区分：随机 不确定性（或偶然不确定性）——不确定性源于所测量属性或现象中自然存在的、不可预测的方差；以及认知不确定性——不确定性源于模型、工具和理解能力的局限 [41, 86]。然而，这种不确定性表述在实际应用中的价值有时也受到批评，因为随着我们对问题的理解不断深入，曾经看似不可约的不确定性，可能实际上是由后来才发现的因素造成的 [28]。

在本研究中，我们并未直接照搬这一既有分类，而是通过模糊性 和分歧的视角，对不确定性提出了一种经过调整的分类。我们选择这种视角，是因为它与我们为实验任务收集判断的机制更直接契合 [13] 不过，更一般地说，我们也指出，对模糊性 和分歧 的特定表述并非旨在全面分类不确定性；具体而言，依据判断的性质和任务的最终目标，采用其他不确定性的量化与分类框架可能更为合适 [77],

2.4 提供上下文以消除歧义

人类判断中不确定性的一个来源，被认为是所判断事物本身的模糊 性。从这个意义上说，模糊性可能源于缺乏足够的上下文来了解待判断的案例，因此，减少模糊性的一种干预方式是补充更多信息。在毒性评级任务中，已有研究表明，添加父级帖子的上下文会影响结果 [66, 88] ；而在调查网络虐待案件时，上下文也可能必不可少 [59]。自然语言处理任务中也有通过添加上下文来提升表现的做法 [40]。上下文的用途并不局限于众包标注，在法律领域等传统场景中，人们也常借助上下文来减少模糊性；其中的判断过程往往包含由专家执行的程序，专门用于建立上下文、消除模糊性 [53]。在教育领域，有效的学生表现评估也会结合上下文，例如考察自由形式回答和中间步骤 [57] ，而不只是最终答案。

尽管上下文非常重要，但要提前知道应该收集哪些上下文可能并不容易。例如，在维基百科的内容审核中，审核员可能会在作出审核决定之前，调查各种因素，例如过往行为、元数据（如 IP 地址）以及相关活动 [78]。其他社区可能会选择调查不同类型的上下文，例如，内战肖像识别社区会查阅历史文献，并依据时间线进行推理，将这些内容作为判断身份关联真实性的上下文 [61]。即使是传统标注也可能需要上下文，不过在这种情况下，可能需要额外的图像来消除遮挡区域的歧义 [52]。因此，只有在知道收集哪些信息最有用时，我们才希望投入额外工作来收集不确定性。

2.5 通过评分细则和审议解决分歧

人类判断中另一种常见的不确定性来源，是个案裁定者之间潜在的分歧 。导致这些分歧的因素有很多，从对任务标准的解读不一致，到裁定者背景各异、视角不同，都可能造成分歧。

对于涉及众包标注的任务，评分细则 [30] 被认为是传达要求、消除对任务各方面困惑的有效工具。然而，即使是专家，也很难制定出涵盖所有边缘案例的评分细则，因此先前研究邀请众包参与者协助制定评分细则 [9, 56, 68] ： 通过找出分歧较大的方面，并征集指南建议。评分细则和指南也可以是隐含的，例如以示例 [50] 或锚点的形式呈现，以便与先前案例进行比较 [13]。

除了制定评分细则面临的挑战之外，评分细则和指南在实际应用中也有局限性。即使采用专家制定的指南，在作出判断后，也可能需要调整和完善，以解决原有指南中的问题 [79]。此外，不仅非专业众包群体如此，专家群体在判断实例时，也可能对标准应为何种内容存在分歧 [73]。在教育 [75], 、医学诊断 [8], 或法律裁决 [82], 等某些高风险领域，人的专业知识往往会使现有指南无法被严格照搬；相反，指南常常会被有条件地偏离，甚至受到质疑并被推翻。在内容审核等社会嵌入型领域，随着数据或裁定者的分布特性发生变化，指南也可能逐渐与实际情况脱节，例如在线平台 [33, 80] 或更广泛社会中的社会规范发生变化时。在这些情况下，未来的裁定者可能会质疑过去的判断及其所依据的标准。

最后，任务不明确并非分歧的唯一来源。即使任务目标和指南都很明确，视角不同的标注者仍可能基于不同的推理视角，对如何评判某个项目产生分歧 [47]。先前关于通过众包实现审议自动化并扩大其规模的研究表明，简单的基于反思的方法能够有效解决分歧 [23, 49]。近期研究利用同步审议 [12, 74]开展情境化审议，使参与审议者能够迅速针对具体分歧点提出有针对性的论据。也有研究考察了各种审议设计选择（例如参与者、审议过程、沟通媒介等）之间的权衡，并发现有效的审议需要营造最契合任务的环境 [18]。当然，更广泛地说，成功运用审议来解决分歧也可能取决于其他因素。例如，确保审议参与者接受有效论证方面的培训 [12] ，并以合作、包容的方式进行沟通 [10], ，可能十分重要，尤其是考虑到审议中存在冲突。此外，审议作为一种智力型任务所具有的群体动态，也意味着确保被安排到审议团队中的成员彼此适配可能很重要 [90]。

在我们的研究中，我们借鉴了有关解决分歧的现有文献，将审议构建为我们的干预措施之一。然而，设计恰当的审议颇具挑战，而且取决于参与者和任务，因此，我们采用较简单的审议形式，很可能不如针对相关任务定制审议有效。

3 设计

本节将介绍 Judgment Sieve 工作流程的设计。我们的工作流程包含以下步骤（如图 1 所示）：

(1) 通过一种能够测量每个实例的个体模糊性和群体分歧的流程，收集群体成员对各实例的判断。(2) 根据上一步的测量结果，计算每个实例的两个得分：模糊性（M_a）和分歧（M_d）。(3) 根据每个实例的模糊性和分歧测量结果，为其分配可能的干预措施：

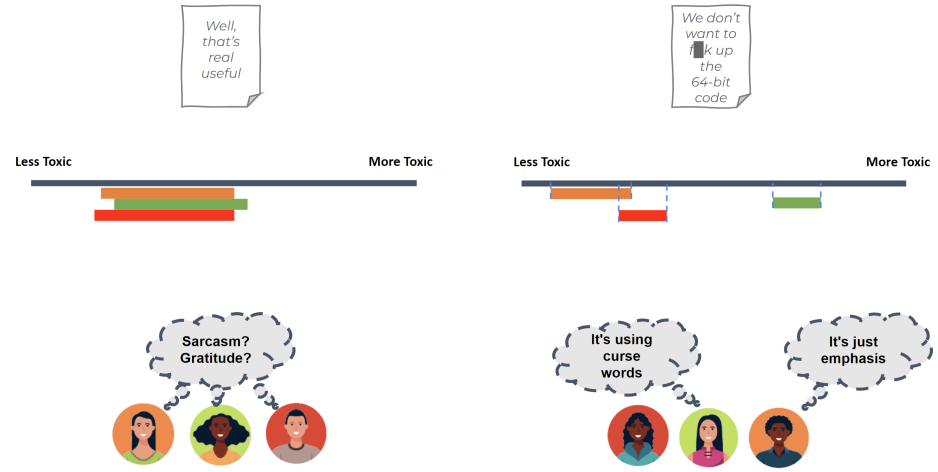
- 如果模糊性评分高于设定阈值，则为该实例分配上下文干预。在此干预下，将收集该实例的额外上下文信息并将其纳入该实例中。
- 如果分歧分数高于设定阈值，则为该实例分配审议干预。在此干预下，将招募一个新的群体重新标注该实例，随后围绕他们对该实例判断的分歧展开审议。针对每个实例的审议结束时，群体将共同提出一项新的通用指南，以他们认为最能解决该分歧的方式进行表述。(4) 纳入干预措施产生的新信息。收集到的额外上下文信息将作为相应实例的一部分。制定出的额外指南将纳入判断任务的定义中。(5)

根据需要重复此过程，直到不确定性达到可接受的水平
本节其余部分将详细介绍工作流程设计的各个方面。

3.1 衡量不确定性来源

为了选择合适的干预，我们首先需要一种方法来了解哪些来源可能导致群体对每个实例当前存在的不确定性。在我们的工作流程中，我们关注两种不同的不确定性来源：每个判断本身固有的模糊程度（M_a），以及不同判断之间的分歧程度（M_d）。从宏观层面看，区分这两种不确定性来源的一种方式，是考察不确定性由谁造成：模糊性反映了每位标注者对其判断的确信程度，这一信息通过我们的标注界面直接收集（图 3）；分歧则是一种涌现属性，源于汇总各位标注者的判断。

在我们的实验中，我们考察了工作流程的一种常见应用场景：在连续量表上作出评分判断。在应用定向干预之前，我们首先需要一种标注方法，以便区分不同的不确定性来源。大多数专注于评分判断的常见标注工具都通过汇总来测量不确定性，并将分歧作为不确定性的替代指标 [26, 89]。然而，要实施有效的干预，我们需要一种能够区分不确定性来源的方法。一些先前研究纳入了让标注者通过直接报告自身不确定性的估计值来表明其信心的方式 [14]，然而，人们通常并不擅长作出这类评估 [84]。对于我们的具体标量评分应用，我们



(a) 当标注者发现实例存在歧义时，会产生更宽的范围——这反映了他们认为可以接受的更多评分等级——由此产生。

(b) 当标注者对评分意见不一致时，我们会看到，他们的范围被放置在量表上的不同位置，从而导致重叠更少。

图 2. 示意图展示了两种不确定性来源——模糊性和分歧——如何体现为基于范围的评分标注工具（如 Goldilocks [13]）生成的范围测量值。

采用了先前研究提出的一种标注方法 Goldilocks [13]，，该方法提出了一种方式，可在每位标注者的标注过程中，分别收集对其各自评估的不确定性来源的测量结果。Goldilocks 通过将评分判断改为范围标注任务来实现这一点：评分者不再给出单个评分，而是给出一个范围（$[l_i^{(x)}, u_i^{(x)}]$），以表示他们认为适合赋予该项目的数值。

借助这种方法收集的范围标注，我们可以定义两个度量指标，用于量化每个实例的不同不确定性来源（x）。首先来看模糊性——即单个标注者不确定所判断实例的评分时出现的情况。采用基于范围的标注流程时，我们可以看到，这种模糊性会反映在生成范围的大小上；范围"越宽"，表示对评分的模糊性越大（图 2a）。因此，我们可以将每个实例的模糊性评分定义为参与判断过程的标注者群体所提交的所有范围的平均大小。

$$\text{Ambiguity}(x, i) = u_i^{(x)} - l_i^{(x)}$$

$$M_a(x) = \frac{1}{|N|} \sum_{i \in N} \text{Ambiguity}(x, i)$$

关于参与者之间的一致性（或分歧），我们可以看到，使用基于范围的标注时，标注者越一致，他们给出的范围就越有可能重叠。因此，一种自然的度量方式是考察一位标注者的范围与另一位标注者的范围重叠的程度——在本例中即比率（图 2b）。然而，与和固定量表相关的范围大小不同，直接计算重叠会得到一种也受范围大小（或在本例中受模糊性）影响的度量。我们可以看到，随着

任一范围的绝对大小增加（反映出模糊性更高），该范围与另一范围重叠的可能性也会增加，从而导致更高的重叠率。为此并得出分歧度量，我们不直接使用重叠率，而是比较测得的重叠率与根据所比较范围的大小计算出的预期重叠率之间的差异。我们注意到，对于任意范围 $[l, u]$，与另一随机均匀放置的范围 $[l\prime u']$ 相比，其预期重叠率等于另一范围的大小（$u' - l\prime$）。根据上述观察，我们将一致性（或分歧）定义为：

$$\text{Overlap}(l, u, l', u') = \max(\min(u, u') - \max(l, l'), 0)/(u - l)$$

$$\text{Agreement}(x, i) = \sum_{j \neq i \in N} \text{Overlap}(l_i, u_i, l_j, u_j) - (u_j - l_j)$$

$$M_{\text{d}}(x) = -\frac{1}{|N|} \sum_{i \in N} \text{Agreement}(x, i)$$

直观来看，某位标注者对某个实例的一致性得分越高，意味着该标注者与其同伴越一致。正分表示该实例上的一致性高于随机水平——标注者倾向于一致；负分则表示一致性低于随机水平——标注者倾向于分歧。我们指出，这种针对每位标注者定义一致性的方式通常不具备交换性（即从标注者 A 的角度衡量的一对标注者 A、B 之间的一致性，不一定等同于从 B 的角度衡量的结果）。这反映了一致性中自然存在的不对称性，可通过范围重叠体现出来：假设一对标注者中，一方的范围“较窄”（是另一方范围的子集），那么由于双方都接受“较窄”范围内的评分，较窄范围的一方可能会认同范围“较宽”（是其超集）的同伴；然而从同伴的角度来看，自己认为可接受的一些评分却未被范围“较窄”的一方接受。最后，为了使该指标更直观，我们可以对“一致性”指标取反，定义分歧。通过计算所有标注者的平均分歧，我们可以得到每个实例的分歧分数。

3.2 收集额外上下文

当个体判断之间的模糊性较高时，上下文 已被证明是降低这种不确定性的有效方式，既适用于传统的人类判断场景 [73]，也适用于以众包任务形式开展的群体判断 [59]。然而，具体到不同的判断任务，上下文 本身可能涵盖种类繁多的信息，而获取这些信息的成本各不相同，且不一定能有效减少模糊性。即使获取上下文的成本很低，提供过多上下文也可能耗尽人工裁定者有限的注意力容量，拖慢判断过程 [69]；而误导性上下文则可能导致错误的决策 [83]。因此，要实现可扩展的人类判断，试图全面收集上下文很可能是个无法实现的目标。由此，我们需要建立一种收集额外上下文的流程，并以测量结果为依据，判断哪些案例确实存在模糊性、可能从上下文中获益。

在 Judgment Sieve 中，我们将上下文的收集设计为一个开放式流程，并根据工作流程所应用的具体领域进行定制。虽然我们无法为所有应用提供全面指导，但会通过几个具体示例，简要说明如何在两种场景中收集上下文：一种是社区参与度较高的**内容审核**任务，另一种是更侧重标注的**数据集标注**任务。

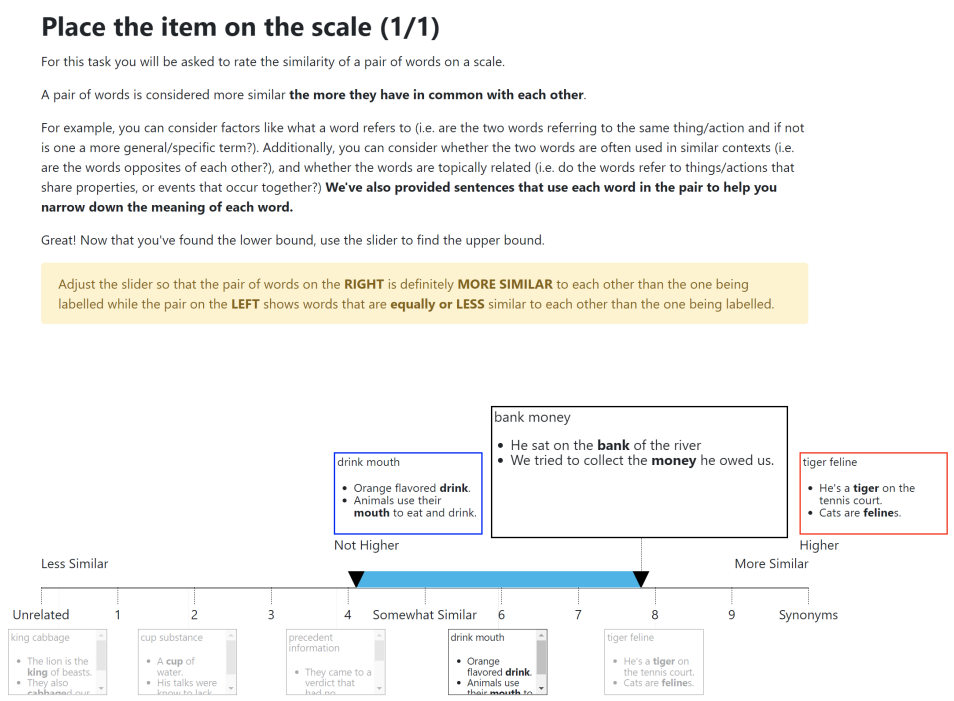


图 3。标注过程中所使用界面的屏幕截图。该标注工具允许我们收集标注者对每个项目所观察到的模糊性进行个体判断的测量结果，并通过比较不同标注者的范围来测量分歧。

在许多在线社区中，**内容审核**通常由一个审核员委员会负责，成员共同决定采取何种审核措施（例如降低内容的展示优先级、删除内容、封禁用户或不采取任何行动）[25, 58]。虽然有些案例有明确证据支持某项措施，但历史上也曾出现备受关注的案例：由于上下文有限，新闻内容被归类为色情内容 [33, 43, 62]。对于此类判断任务，Judgment Sieve 可以识别出因模糊性较高而可能需要补充上下文的案例集合。随后，委员会可以逐一审视这些案例，指出哪些上下文可能有助于消除相应案例的模糊性，并委派另一个调查小组收集相关证据。该小组会查阅日志和元数据，并整理成一份“案例档案”。这份调查“案例档案”随后会作为上下文附加到案例中 [25]。对于**数据集标注**，Judgment Sieve 则会通过模糊性指标，直接向众包任务发布者指出模糊案例。随后，众包任务发布者可以发起额外任务，让工作人员为模糊实例补充假设性上下文，以消除模糊性，并将澄清后的案例重新纳入数据集。另一种做法是，如果数据可以重新收集——例如，不清晰的图像 [20] 是由他们自己的数据收集过程造成的——任务发布者可以针对该实例重新启动数据收集流程，并将新数据作为上下文添加。

在我们的实验中，我们并未专注于优化任何特定的上下文收集策略。相反，我们通过选取一个已收集上下文线索的数据集，并将上下文暂时隐去，来模拟一种尚未开展补充上下文工作的情形。

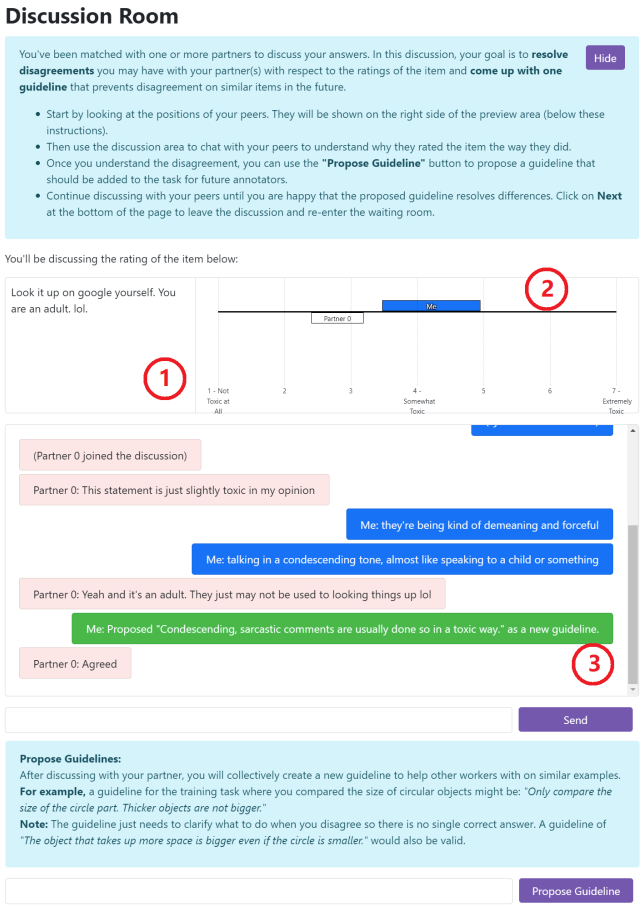


图 4. 我们在实验中使用的审议界面截图。该界面包含 3 个主要部分：(1) 对所评估实例的预览；(2) 在同一尺度上显示每位参与者范围答案的可视化；(3) 同步讨论区。

3.3 通过审议解决分歧

我们的审议过程设计受到先前研究的启发，旨在通过同步审议来解决分歧 [12, 74]。在我们的工作流程中，分歧指标 M_d 用于自动查找最有可能从审议中受益的候选实例——即分歧是不确定性的主要来源的情形。随后，小组先独立对每个示例中的项目作出判断，再进行同步集体讨论。在审议过程中，每位小组成员的判断都会以可视化方式呈现，并提示小组共同借此将自己的判断与同伴的判断进行比较。审议参与者会受到提示，思考并向同伴阐述自己作出判断时所依据的标准。然而，与传统审议系统以对实例作出判断作为审议结果不同，我们的审议过程旨在形成一份可推广的指南，以解决类似的分歧。在参与基于讨论的审议过程后，参与者会受到提示，思考他们在审议中观察到的视角以及审议结果，以便共同提出一份适用于未来示例的指南，

从而解决该实例中视角的差异。

所有审议结束后，可以收集提出的指南，并在必要时去除重复项。这样便能形成一套最终指南，并将其纳入任务中，从而使今后在任务本身中考虑到类似类型的分歧。我们注意到，这一整体工作流程也让人想起先前提出的一些方法，这些方法旨在利用工作者提供的反馈来改进众包任务中的说明 [56, 68]。不过，我们的工作流程利用审议过程，引导参与者提出更有效的解决方案，以应对观察到的分歧，而不是针对说明不足的问题。

3.4 系统原型实现

本节将介绍我们用于开展实验的原型[1] 的一些技术细节。为了构建适用于我们工作流程的系统原型，我们创建了两个主要组件：（1）一个用于收集基于区间的标量评分的**标注** 应用，以便测量模糊性和分歧（图 3）；（2）一个**审议** 应用，用于收集基于区间的评分，并将参与者配对到同步审议会话中（图 4）。

我们的**标注**应用程序遵循 Goldilocks [13], 的一般设计，并以静态 Web 应用程序的形式实现，输入的标注数据和输出的标注者响应均直接通过 Amazon Mechanical Turk（AMT）存储。我们使用自定义的 JavaScript 工具包[2] 与 AMT 交互，并协调实验条件和数据存储。我们的**审议**应用程序借鉴了以往同步审议系统的设计 [12, 74]。其前端与 AMT 的交互方式类似于我们的标注应用程序；审议过程中，前端会持续轮询新消息。我们通过一个额外的 Python 后端协调参与者匹配并将其分配到同步讨论室，该后端也会存储讨论内容。讨论参与者的匹配由一名人工操作员负责；该操作员会同步监控一个内部仪表板（第 4.3 节）。

4 实验

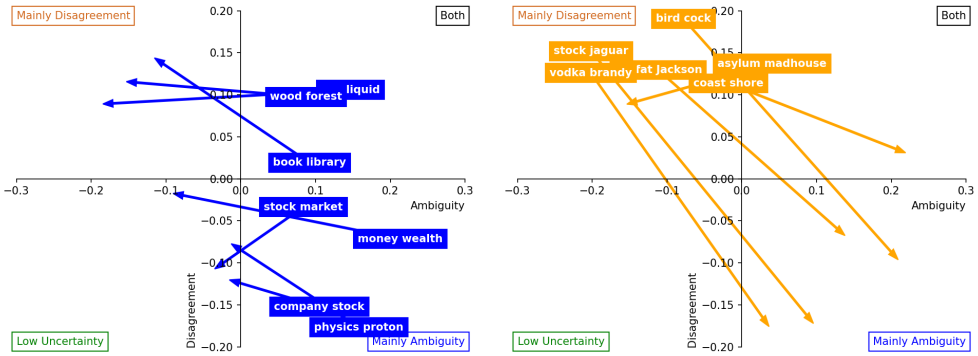
为评估干预措施对群体判断不确定性的影响，我们开展了标注实验，收集群体判断不确定性的测量结果，既包括实施任何干预措施之前的数据，也包括实施每项干预措施之后的数据。对于每个经过标注的实例，我们使用基于范围的标注应用程序，在以下条件下测量模糊性和分歧：基线——未实施干预；上下文——将上下文作为每个实例的一部分；审议——将审议干预提供的额外指南作为任务的一部分。

4.1 任务

在我们的实验中，我们选择了两项基于标注的任务，它们经常会在群体判断中产生不确定性：词语相似度（**wordsim**）和毒性评分（**毒性**）。这两个任务领域都曾用于先前研究，也是判断过程中包含多个不确定性来源的任务示例。

“**wordsim** ”领域由基于 WordSimilarity-535 测试集 [26] 的示例构成，任务是以 0 至 10 分的量表判断词对的相关性。选择这一领域，是因为其中包含多种来源，会导致群体和个人的不确定性。首先，“相关性”这一概念本身并没有明确的定义，

[1]代码可用：https://github.com/Social-Futures-Lab/targeted-interventions-code [2]代码可用：https://github.com/jmchn1994/amt-shim-template



(a) 主要表现出歧义的项目样本-模糊性（蓝色）以及之后新的不确定性应用上下文干预。

(b) 一组主要表现出分歧（橙色）的项目，以及应用审议干预后其新的不确定性。

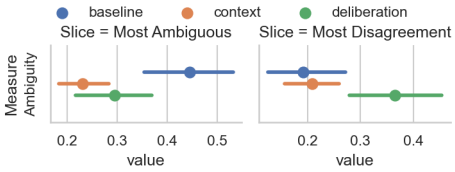
图 5. 示意图展示了少量项目的不确定性如何在不确定性空间中发生变化。橙色标示的项目主要表现出分歧，蓝色标示的项目主要表现出模糊性。箭头指向实施针对性干预后项目在不确定性空间中的新位置。分数经过重新缩放，使原点 (0, 0) 代表所有项目的平均模糊性和平均分歧。正值表示测量结果中的不确定性得分高于平均水平。

而 **wordsim** 任务本身也较为模糊，这可能导致不同的人对相关性有不同理解，并体现为不同的思想流派，例如从词语的含义、用法、一般性和出现模式等不同方面比较词语的相关性。此外，该任务涉及的许多词都有多个词义。由于没有提供上下文来消除歧义，标注者还必须自行决定如何处理不同词义可能带来的模糊性。为启动基于范围的标注流程，我们利用现有的相似度标注，选取了沿范围均匀分布且方差最低的 5 个词对示例作为种子。随后，我们从剩余项目中随机抽取 50 个词对，并将其分为 5 组，每组 10 个，由此构建标注数据集。

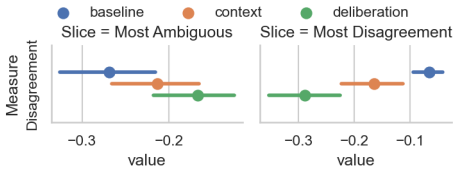
该**毒性** 领域由从维基百科讨论页 [66] 收集的评论组成，其任务是使用带有七点评分语义差异量表标签的连续评分量表，对每条评论的毒性进行判断。判断毒性本身就是一项伴随显著不确定性和分歧的任务。我们注意到，先前研究表明，每位标注者的背景以及评论发布时的具体情境，都会极大影响标注者是否会认为同一条帖子更具毒性 [72]。因此，标注中自然会产生分歧和模糊性。由于有些评论可能长达多个段落，大幅增加标注工作量，我们首先对数据集进行筛选，只保留评论和父评论长度均不超过 280 个字符的实例。随后，为启动基于范围的标注流程，我们利用数据集来源中已有的毒性标注，选取了 5 条种子评论示例；这些示例在范围内均匀分布，且方差最低。接着，我们从其余项目中随机抽取 50 条评论，分为 5 组，每组 10 条，以此创建标注数据集。

4.2 获取语境

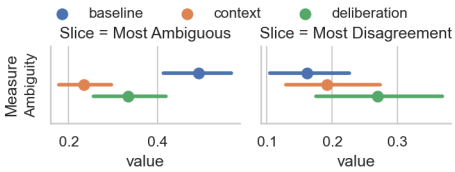
通常，获取语境的过程取决于群体的具体目标和任务。由于这一过程独立于工作流程本身，而且我们并未试图评估



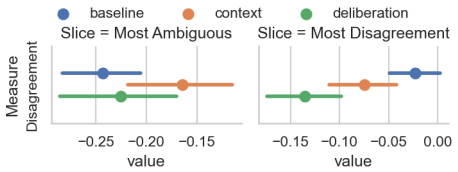
(a) *wordsim* 任务中每个切片的模糊性



(b) *wordsim* 任务中每个切片的分歧



(c) *toxicity* 任务中每个切片的模糊性



(d) *toxicity* 任务中每个切片的分歧

图 6. 各任务领域的点图，展示了基线、上下文和审议干预条件下的模糊性与分歧度量。对于每项度量，我们考察数据集中的两个切片：按模糊性排名前 10% 的实例 M_a （“最模糊”）和按分歧排名前 10% 的实例 M_d （“分歧最大”）。误差线表示 95% 置信区间。

所获取语境的质量，因此我们使用已补充语境的任务数据集来模拟获取语境的过程。在基线条件下进行标注时，标注者无法获知每个项目的语境；而在上下文条件下，语境则向他们开放。

对于 **wordsim** 任务，我们借鉴了先前研究 [40], 的做法，该研究使用包含目标词的例句来提供上下文。对于数据集中的每个词，我们通过选取一个例句来构建其上下文；该例句中词语的形式与其在**wordsim**词对中的形式相同。例句在可用时取自 WordNet [60]；若没有可用例句，则使用在线词典服务[3]。如果存在多个词义，则随机选取一个词义对应的例句。优先采用较短的例句，并对较长的句子进行人工简化。随后，将词对中每个词对应的例句拼接起来，构成该词对的上下文。

对于**毒性**任务，我们的数据集来源 [66] 已包含上下文信息，其形式为每条评论的父评论。我们通过附加与该项目相关的父评论，并加上表明其为父帖子的标签，向标注者提供上下文。

4.3 开展审议

我们使用群体任务开展审议，以制定审议干预措施所需的指南。在任务开始时，每位参与者首先完成培训环节，学习如何使用标注界面。完成该环节后，参与者会被安排进入等候室，在那里可能被分配到评估环节或审议环节。在评估环节中，参与者使用基于范围的标注界面，为所标注的实例提供判断。在审议环节中，参与者会与 1-2 名伙伴配对，并被要求使用实时同步讨论界面（图 4）讨论其范围标注中观察到的

[3]https://www.merriam-webster.com/

分歧，并协作制定供未来标注者使用的指南。任何参与者都可以提出或更新指南，且参与者只有在提出一条指南后才能离开讨论。评估环节和审议环节的分配以半自动方式完成：当参与者处于等候室时，审议系统会向该参与者提供一组可用的环节选项。随后，审议主持人可以从这些选项中进行选择，将其中一个分配给该参与者。

审议完成后，我们收集了每个项目的最终指南提案。随后，我们通过删除相似的提案，手动去除了重复内容。我们还对提案进行了小幅修改，使其在每个任务领域中的表述保持一致。收集到的提案随后被纳入审议条件标注实验的任务说明中，其中向 **毒性** 任务添加了 5 条新指南，向 **wordsim** 任务添加了 6 条新指南。

4.4 招募

我们从 Amazon Mechanical Turk（AMT）招募众包工作者，使用基于 Goldilocks [13] 的标注界面，分别在基线、上下文和审议条件下进行标注。对于每个领域中的每种条件，我们招募了 25 名工作者（共 150 人）。每位参与者需为所部署的每项任务标注 10 个项目。在每个领域内，我们确保一名工作者不能参与超过 1 项标注任务（如果尝试参与第一项以外的任何任务，则显示通知并阻止其继续操作），从而确保同一任务领域中不同条件之间使用独立的工作者群体。参与者完成训练任务后可获得 1.0 美元的基础报酬，完成全部标注后还可获得额外 1.0 美元。每完成一项标注，参与者在 **wordsim** 领域可获得 0.3 美元（总计 3.0 美元），在 **毒性** 领域可获得 0.5 美元（总计 5.0 美元）。测得两个领域的时薪中位数分别为 13.5 美元和 15.9 美元。

此外，我们还招募了另外一批 AMT 工人，参与各个领域实例的审议环节，以制定审议条件中使用的指南。对于每个领域和任务组，我们招募了 4 名讨论参与者（共 40 名）。我们通过资格条件确保参与审议环节的工人没有参与标注。工人参与一项为时一小时的讨论任务（包括 10 场讨论和 10 场标注环节），报酬为 20 美元。对于积极参与讨论且超过规定的 10 场讨论的工人，我们额外给予 4 美元奖金。

4.5 模拟实验

收集到 3 种条件下各自的标注实验数据后，我们便能够模拟为每个实例选择针对性干预的结果。在我们的模拟实验中，我们使用基线条件下收集的模糊性 M_{a} 和分歧 M_{d} 分数来决定该实例采用哪种干预。

在我们的实验中，我们选择了 0.1 的阈值，该阈值针对的是模糊性评分或分歧分数排名前 10% 的实例。为开展模拟，我们根据从基线条件中收集的 M_{a} 和 M_{d} 评分对实例进行排序。$\bar{\ }$ $\bar{\ }$我们利用这些评分确定模糊性和分歧分数的截断阈值（$M_{\mathrm{a}}, M_{\mathrm{d}}$）。$\bar{\ }$然后，对于数据集中的每个实例，我们首先检查其模糊性评分。如果 $M_{\mathrm{a}}(x) \geq M_{\mathrm{a}}$，我们便从该实例的上下文条件中抽取标注值，实施上下文干预，然后继续处理下一个实例。否则，我们检查分歧分数；如果$M_{\mathrm{d}}(x) \geq \bar{M_{\mathrm{d}}}$，则从该实例的审议条件中抽取标注值。如果两个不确定性指标都未超过阈值，我们将保持基线条件中的标注值不变。

4.6 结果

为评估我们的工作流程，我们重点关注两个主要方面：评估每项干预对其所针对的不确定性类型的影响，以及评估根据每个示例的不确定性测量值动态选择针对性干预，是否能够比统一应用干预更高效地降低不确定性。

具体而言，我们评估以下假设：

• **H1-a（干预措施有效）**：干预措施能够有效减少其所针对的不确定性来源：上下文在减少模糊性方面最有效，而审议在减少分歧方面最有效。

• **H1-b（干预措施具有针对性）**：干预措施无法有效减少其未针对的不确定性类型。

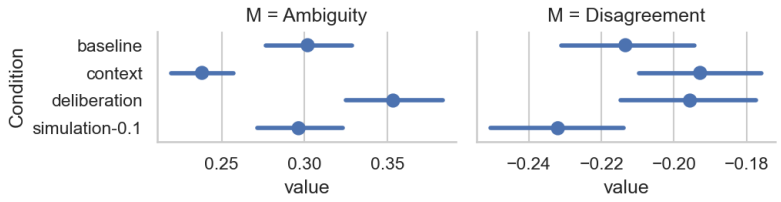
• **H2（高效降低不确定性）**：仅依据未施加任何干预时收集的不确定性测量结果来制定决策流程，可以为每个实例选择更优的干预措施，相较于对所有实例统一施加同一种干预，更高效地降低不确定性。

4.6.1 定向干预的有效性。本节将考察，在我们的两个任务领域中，关于干预有效性及其定向特性的假设是否得到支持。为检验这些假设，我们根据基线标注期间测得的主要不确定性来源，从每项任务中提取两个实例子集（切片）。对于每个领域，我们选取模糊性测量值最高的前 10% 实例，作为“最模糊”切片；并选取分歧测量值最高的前 10% 实例，作为“分歧最大”切片。随后，对于每组实例，我们追踪其在每项干预（上下文和审议）后重新标注时的不确定性。我们在图6中展示这些测量结果。

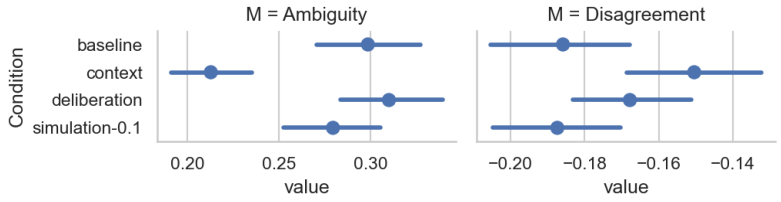
考察各领域中“最模糊”实例的子集后，我们发现，只有上下文干预条件在降低 **wordsim** 和 **toxicity** 两个任务领域中的模糊性方面具有统计显著性（$p < 0.001$；根据图基诚实显著差异检验，仅在基线条件与上下文条件之间观察到显著差异）。对于“分歧最大”实例的子集，在分歧方面也发现了类似结果：只有审议与基线条件的配对比较显示分歧显著降低（$p < 0.001$）。这支持**H1-a**，表明干预措施能够有效降低其所针对的那类不确定性。

我们还考察了干预措施对另一种（非目标）不确定性来源的影响。在 **wordsim** 和 **toxicity** 两个领域中，我们没有观察到非目标条件与基线条件之间存在统计显著的差异。未观察到显著性并不意味着非目标条件对不确定性来源毫无影响，但这确实表明其效果不及针对性干预，因此对**H1-b**提供了部分支持。值得注意的是，在**wordsim**领域“分歧最大”实例的子集中，尽管审议显著降低了分歧，但也显著增加了模糊性。考虑到任务的性质，我们推测，审议形成的指南使参与者在判断词语相关性时考虑了更多因素（词义、间接关系）；由于缺乏其他上下文，他们因此认为这些实例更加模糊。

4.6.2 决策过程效率。将视角提升一个层级，考察在整个数据集中的所有实例上统一应用每种干预的情况（图7），我们发现，上下文能够降低两个领域中的模糊性（对于**wordsim** 和**毒性**领域，分别为$p = 0.0027 < 0.01$ 和 $p < 0.001$）。然而，这似乎也伴随着轻微的



(a) wordsim 任务领域的比较



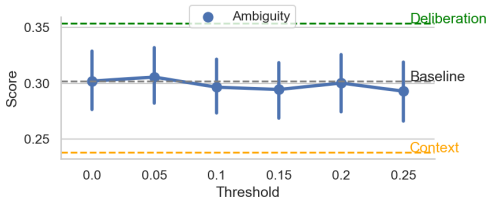
(b) 毒性任务领域的比较。

图 7. 各领域的点图，展示了在所有实例上应用统一干预（上下文或审议）后测得的模糊性和分歧，以及通过模拟针对每个实例筛选不同干预措施所得的结果simulation-0.1。误差线表示 95% 置信区间。

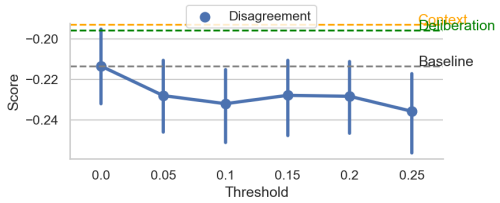
代价，在两种情况下还提高了平均意见分歧（对于**毒性**为$p = 0.026 > 0.01$，在我们的统计检验中均不显著[4] $p < 0.01$；对于**wordsim**为 $p > 0.01$，不显著）。这表明，对所有实例一概应用相同的干预可能会带来权衡，甚至可能增加其原本无意解决的不确定性来源。在考察审议条件时，我们发现，将其应用于整个数据集时，对任一不确定性来源都没有统计显著效应，但两项测量结果的平均值均略有上升。这表明，尽管审议对于分歧最大的实例可能有用，但广泛应用审议可能会产生不利影响。这一结果总体上与关于审议的先前研究一致；相关研究表明，审议可能只有在项目本身的模糊性已经较低时才有效 [74]，并且应主要用于具有较高分歧的困难案例。

接下来，我们比较模拟决策过程的结果：根据实例的不确定性主要源于模糊性还是分歧，为其分配不同的干预措施。与基线相比，我们发现，在两个领域中，我们的模拟过程（模拟-0.1）都使模糊性和分歧指标的均值有所降低。然而，这一下降在统计上并未达到显著水平。不过，未能获得显著结果并不意外，因为我们的模拟筛选方法仅对基线标注期间测得模糊性和分歧最高的前 10% 实例实施干预（最多仅影响 20% 的实例），其余所有实例仍保留原有标注。我们还注意到，提高决策阈值会使结果偏向上下文条件——模糊性显著下降更多，但代价是分歧增加。有趣的是，我们观察到两个任务领域对模拟决策过程的反应各不相同：**wordsim** 通过减少分歧使不确定性降幅最大（-8.7%），而**toxicity** 则使模糊性降幅更大（-6.4%）。我们

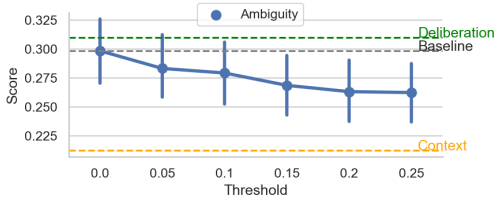
[4]我们预先将显著性水平设定为



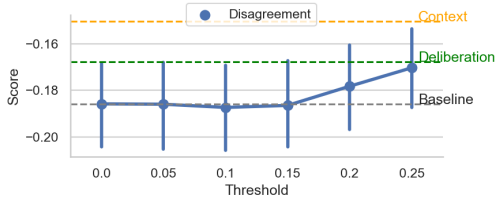
(a) wordsim任务中的总体模糊性



(b) wordsim 任务中的总体分歧



(c) 毒性任务的总体模糊性



(d) 毒性任务上的总体分歧

图8。展示在不同阈值下应用模拟干预措施的图表：0%（未应用干预措施）、5%、10%、15%、20%和25%。对于所有图表，数值越低表示来自相应来源的不确定性越低。每张图中均提供了三条参考线，用于表示以下来源的不确定性测量值平均值：基线（灰色）、上下文（橙色）和审议（绿色）。误差线表示模拟结果周围的95% 置信区间；参考线的置信区间未显示（请参见图7）。

推测，这可能是因为在**toxicity**判断中，分歧更难解决。归根结底，虽然我们无法通过一轮有针对性的干预，以统计显著的方式证明**H2** 成立，但我们确实观察到了一些差异，这或许能帮助我们避免在一律添加上下文或对所有实例进行审议之间做出权衡。

5 讨论

本节首先考察用于选择干预措施的阈值变化所产生的影响，并讨论阈值（其影响每轮的不确定性降低）如何与我们的工作流程的迭代改进式应用协同发挥作用。随后，我们将讨论通过审议产生的指南的一些定性观察，以及这些观察可能如何关联到我们在两个任务领域之间发现的差异。接着，我们将讨论我们的工作流程如何协调同时涉及模糊性和分歧的情境，并讨论除实验中所采用的众包标量评分标注之外，我们的工作流程如何推广到不同的任务和模态。最后，我们将讨论所探讨的两种干预措施——上下文和审议——存在的一些局限，以及未来可能解决其中部分局限的研究方向。

5.1 干预筛选阈值与迭代改进

在第 4.6.2 节中，我们发现，通过模拟一种决策过程，对模糊性和分歧程度最高的前 10% 实例分别应用干预措施，我们能够观察到这两种不确定性都有所降低，但降低幅度在统计上并不显著。由于受影响的实例最多占 20%，因此一个随之而来的问题是：如果改变这一阈值，允许对更多（或更少）实例应用干预措施，会发生什么？为了探讨这个问题，

我们调整了模拟参数，以模拟在其他阈值下的决策过程，如图 8 所示）。

通过这些模拟，我们可以观察到，所测试的两个任务领域对所选针对性干预的敏感程度有所不同。对于*wordsim* 领域，我们发现，应用针对性干预措施能够减少总体分歧，但对总体模糊性的改善相对有限。根据第 4.6.1 节的结果，我们知道，上下文干预能够有效降低最模糊实例的模糊性。这表明，审议干预可能加剧了高分歧案例的模糊性，从而抵消了上下文干预所带来的模糊性降低。我们推测，在这一领域，新增的指南促使人们更全面地看待“词语相似度”；标注者意识到，原本他们确信无疑（因而彼此存在分歧）的案例实际上存在模糊性，而且若非指南的帮助，他们不会考虑这些其他解释。另一方面，对于毒性领域，我们发现情况几乎相反：针对性干预措施降低了总体模糊性，但对总体分歧影响甚微（甚至可能使其增加）。这表明，在这一领域，更多上下文或许降低了人们对在线评论所处情境的理解模糊性，但也可能引发了新的分歧：不同标注者对毒性含义的理解各不相同 [72]。

虽然这一模拟结果本身很有意思，但我们注意到，在实际应用中，无法通过这种方法找到“最优”阈值，因为每项干预措施都必须应用于所有实例，过程效率低下。相反，我们认为，在实践中，优化阈值可能并非降低不确定性的最有效方式；更好的方法是采用**迭代改进** [35] 的方案，让工作流程额外迭代运行，在上一轮干预措施实施后，针对数据和任务继续执行。先前研究表明，我们工作流程中使用的一些不确定性干预措施（如审议），可能只对模糊性较低的实例有效，否则甚至会适得其反 [12, 74]。事实上，我们在图7中也观察到了这一点：若对所有实例一概实施审议，两个领域的总体分歧都会略有增加。然而，即便不加区分地实施审议无法降低分歧，针对性地实施审议仍然可以做到这一点（图8d）。这表明，与其优化阈值，更有效的方法是迭代运行工作流程：每次迭代后，原本存在模糊性（因而不适合审议）的实例可能变得不那么模糊，从而有望在下一轮将审议作为有效干预措施应用于这些实例。在迭代构建工作流程时，筛选阈值可以被视为控制每轮不确定性降低速率的一种方式（有些类似于“学习率”）。较低的阈值更为保守，对总体不确定性的影响较小，但更可能避免干预措施相互抵消彼此的益处；较高的阈值则反映出对干预措施更乐观的看法，虽然会提高某一轮未能降低不确定性的可能性，但一旦奏效，每一步带来的影响也更大。

5.2 所生成指南的效用

从结果中我们发现，在整个数据集上应用审议可能会导致分歧增加，尽管我们也观察到，对于分歧最大的那些案例，审议能够减少分歧。为探究这一现象，我们对审议过程中形成的若干指南进行了定性考察，以了解它们为何未能有效扩展到更多实例。

对于*wordsim* 领域，我们发现，审议促成了指南的制定，其中列出了判定“相似”的更多标准，例如：“反义词（明/暗、善/恶）是相似的。”、“词语之间的因果 [sic] 关系会让它们更相似。”以及“属于某种

自然演进的词语更相似。”然而，尽管这些指南很可能为讨论并制定这些指南时涉及的词对提供了更一致的标准，但在应用时仍可能出现分歧——例如，某一对词语是否应被视为因果配对或自然演进？对于毒性 领域，我们发现，审议促成了新指南的制定，例如：“针对政策而非个人的陈述不被视为具有毒性。”、“贬低或居高临下的陈述可能具有毒性。”与*wordsim*中的情况一样，这些指南总体上也相当狭窄（“针对政策的陈述”），或者在上下文有限时可能含糊不清（“居高临下的陈述”）。

虽然这并非对制定指南的有效性进行的全面探讨，但我们注意到，这有助于解释为什么根据我们特定的审议方案制定的指南在某些情况下可能无法很好地推广。我们还注意到，尽管我们在评估 Judgment Sieve 时采用了特定的审议设计，但其实现只是概念验证，并不代表针对我们的任务领域优化的最有效审议方案。在实践中，采用审议的群体可以选择其他设计，以提高匹配质量或引入专家主导的流程。

5.3 同时处理模糊性与分歧

正如我们在实验中观察到的，尽管模糊性和分歧在很大程度上是两种不同类型的不确定性，但一个实例同时具有高度模糊性和高度分歧的情况也并不少见。在这种情况下，应选择重点关注哪一方面？在我们的针对性干预模拟版本中，我们选择优先解决模糊性，然后再处理分歧。这一决定参考了先前研究的发现：我们使用的审议干预（以自包含同步在线讨论的形式进行）可能无法解决模糊性较高情况下的分歧 [74]。不过，这一特定选择未必总是最佳方案。在某些情况下，优先开展讨论可能更有成效。例如，围绕新兴议题的内容审核决策，可能既需要就指南达成共识（正如各平台针对 COVID-19 相关虚假信息活动所采取的调整），也需要收集能够支持最终决策的证据（例如当前的科学共识、评估内容是否与更大规模的虚假信息活动相关等）。在这些情况下，先开展审议或同时进行审议，可以帮助厘清后续所需的上下文，最终引导出更有效的上下文收集过程。未来一个值得探索的方向，是开发融合上下文收集与审议过程的方法，让群体能够根据需要在两者之间来回切换。

5.4 将我们的方法推广到不同任务和模态

在本研究中，我们针对短文本任务中的特定判断模态评估了 Judgment Sieve，即采用连续评分量表进行连续标量评分。然而，更广泛地说，在涉及群体人类判断的许多其他场景（例如专家参与的群体判断）和判断模态（例如单标签或多标签分类）中，有针对性地降低不确定性也可能带来益处。本节将讨论我们认为有机会推广工作流程的两个方面：通过针对特定任务的迭代流程支持**复杂任务**，以及支持**标量评分以外的模态**。

虽然 Judgment Sieve 是围绕在众包环境中降低群体判断的不确定性而构建的，但分类和量化不确定性并应用针对性干预措施这一更高层次的概念，也适用于其他类型的任务和场景。例如，在教育环境中，受 Judgment Sieve 启发的工作流程可能包括由专家级助教在评分时使用类似工具，测量分数分配方面的分歧，或某些类型答案所存在的模糊性。随后，这些测量结果可用于群体讨论

，进而促成评分标准细化（以降低不确定性），或在今后更新作业问题（以减少模糊性）。同样，在线社区也可以将受 Judgment Sieve 启发的流程纳入其内容审核流程，以便对决策中的不确定性进行审查，并在必要时加以处理。面向社区使用的 Judgment Sieve 迭代版本还可能随着社区规范不断演变，提供更长期的稳定性 [95]。

与类别标注或多标签标注等其他模态一样，Judgment Sieve 可以调整为使用针对这些模态设计的不确定性分类和度量指标。例如，在类别场景中可以使用软标签 [15]，从而利用离散度（如方差）来衡量单标注者分布的模糊性，并利用散度（如 KL散度、Wasserstein距离）来衡量标注者间分歧。我们也注意到，尽管其泛化潜力很大，但 Judgment Sieve 预计仍最适用于降低更复杂或更具主观性的判断场景中的不确定性；相比之下，以感知为主的任务预计较少受益于不确定性的分解式视角。

5.5 上下文的注意事项

虽然通常上下文可以减少模糊性，但在我们的实验中，我们确实观察到一些情况：上下文会增加不确定性。这主要发生在上下文出乎意料时。例如，在考察 wordsim 领域中添加上下文后模糊性增加的情况时，我们发现了诸如“bank, money”变得更加模糊的实例。然而，这并不意味着上下文干预失败；相反，我们观察到，引入的上下文（“He sat on the**河岸**”这一用法示例）很可能并非标注者在最初判断时预期的词义。尽管不确定性有所增加，但在实践中，这一结果很可能是可取的，因为它反映了不确定性的真实增加。

从更宏观的层面来看，额外的上下文在实际中能够降低不确定性的程度确实有限；最终，我们很可能会发现，收集更多上下文对消歧的益处微乎其微，因而出现边际收益递减。因此，如果要进行进一步迭代，或许应跟踪不确定性降低的幅度，并相应地权衡成本。

5.6 通过审议扩展任务规范的局限

最后，我们还要指出，当前审议过程的设计在扩展方面存在局限。在当前流程中，审议会产出额外的指南，并将其纳入说明。尽管任务发布者可以进行去重和重新组织等工作，但随着任务规范日益精确，用于界定任务本身的说明最终可能变得难以驾驭 [92]。在内容审核等当前的复杂任务中，我们可以看到这种情况：付费承包人员会接受大量培训，以便按照确切的审核指南执行工作。如果指南变得过于复杂，其解决分歧的能力就会下降，因为人们会难以理解或找到相关指南。

解决过于复杂的任务规范的一种潜在方案，或许存在于法律案例构建领域。在该领域中，决策边界可能复杂到无法实际应用，而判断往往依赖于参考既往判决，而不是援引成文法或指南。这为未来探索除审议之外减少分歧的替代方案提供了可能性。

6 结论

本文提出了一种新的工作流程，根据与模糊性和分歧相关的测量结果，针对每个实例实施针对性干预，从而更高效地降低群体判断中的不确定性。通过实验，我们发现，补充上下文和开展审议这两种干预措施，最能有效降低各自针对的不确定性类型。

我们还观察到，按项目动态筛选干预措施，有望避免对所有项目统一实施干预所带来的权衡。

参考文献

[1] Lora Aroyo 和 Chris Welty。2013。群体真实情况：在众包关系抽取真实标准中利用分歧。WebSci2013。ACM 2013，2013（2013）。[2] Shubham Atreja、Libby Hemphill 和 Paul Resnick。2022。人民的意愿是什么？对错误信息的审核偏好。arXiv abs/2202.00799（2022）。[3] Stephanie Alice Baker、Matthew Wade 和 Michael James Walsh。2020。疫情期间应对错误信息的挑战：内容审核与伤害概念的局限性。《澳大利亚国际媒体》177（2020），103–107。[4] Michael S. Bernstein、Joel Brandt、Robert C. Miller 和 David R. Karger。2011。两秒内汇聚众包：实现实时众包驱动的界面。载于第24届 ACM 用户界面软件与技术年度研讨会论文集（UIST ’ *11* 33 42 10 1145 2047196 2047201 ）。计[5] 2020 算机协会，美国纽约州纽约市，–。https://doi.org/ . / 。Lucas Beyer、Olivier J. H’enaff、Alexander Kolesnikov、Xiaohua Zhai 和 Aäron van den Oord。。《ImageNet 的研究已经完成了吗？》arXiv abs/2006.07159（2020）。[6] Umang Bhatt、Javier Antorán、Yunfeng Zhang、Q. Vera Liao、Prasanna Sattigeri、Riccardo Fogliato、Gabrielle Melançon、Ranganath Krishnan、Jason Stanley、Omesh Tickoo、Lama Nachman、Rumi Chunara、Madhulika Srikumar、Adrian Weller 和 Alice Xiang。2021。不确定性作为透明度的一种形式：衡量、传达和利用不确定性。计算机协会，美国纽约州纽约市， 401 413 10 1145 3461702 3462571 –。[7] 2021 27 2021 44 61 https://doi.org/ . / 。Abeba Birhane。。《自动化模糊性的不可能性》。《人工生命》（），–。[8] Flora Blangis、Slimane Allali、Jérémie F Cohen、Nathalie Vabres、Catherine Adamsbaum、Caroline Rey-Salmon、Andreas Werner、Yacine Refes、Pauline Adnot、Christèle Gras-Le Guen 等。2021。高收入国家儿童身体虐待诊断指南的差异：系统综述。《JAMA Network Open》4卷11期（2021），e2129068–e2129068。[9] Jonathan Bragg、Mausam 和 Daniel S. Weld。2018。Sprout：面向众包的众包任务设计。载于第31届 ACM 用户界面软件与技术年度研讨会论文集（UIST ’18）。计算机协会，美国纽约州纽约市， 165 176 10 1145 3242587 3242598 –。https://doi.org/ . / 。[10] Hancheng Cao、Vivian Yang、Victor Chen、Yu Jin Lee、Lydia Stone、N’godjigui Junior Diarrassouba、Mark E. Whiting 和 Michael S. Bernstein。2021。我的团队将继续前行：通过团队互动区分高存续能力与低存续能力的团队。《ACM 人机交互汇刊》4卷CSCW3期，文章230（2021年1月），27页。https://doi.org/10.1145/ 3432929 [11] Joseph Chee Chang、Saleema Amershi 和 Ece Kamar。2017。Revolt：用于标注机器学习数据集的协作式众包。载于2017年 CHI 计算机系统中的人为因素会议论文集。2334–2346。[12] Quanze Chen、Jonathan Bragg、Lydia B. Chilton 和 Daniel S. Weld。2019。Cicero：用于准确众包的多轮上下文论证。《2019年 CHI 计算机系统中的人为因素会议论文集》（2019）。[13] Quan Ze Chen、Daniel S. Weld 和 Amy X. Zhang。2021。Goldilocks：利用相对不确定性的、一致的众包标量标注。《ACM 人机交互汇刊》5卷CSCW2期，文章335（2021年10月），25页。https: // doi.org/ 10 1145 3476076 。/。John Joon Young Chung、Jean Y Song、Sindhu Kutty、Sungsoo Hong、Juho Kim[14] 和 Walter S Lasecki。2019。高效引出方法：估算群体集体答案。《ACM 人机交互汇刊》3卷CSCW（2019），1–25。[15] Katherine M. Collins、Umang Bhatt 和 Adrian Weller。2022。向每位标注者引出软标签并据此学习。载于人类计算与众包会议。[16] Corinna Cortes 和 Neil D. Lawrence。2021。会议同行评审中的不一致性：重新审视2014年 NeurIPS 实验。arXiv abs/2109.09774（2021）。[17] Stephen Crowder、Collin Delker、Eric Forrest 和 Nevin Martin。2020。《计量学统计学导论》。Springer。[18] Todd Davies 和 Reid Chandler。2013。在线审议设计：选择、标准与证据。arXiv 预印本 arXiv:1302.5177（2013）。[19] A. Philip Dawid 和 Allan Skene。1979。使用 EM 算法对观察者错误率进行最大似然估计。《英国皇家统计学会杂志 C辑（应用统计）》28（1979），20–28。[20] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、K. Li 和李飞飞。2009。ImageNet：大型层级图像数据库。2009年 IEEE 计算机视觉与模式识别会议（2009），248–255。[21] Djellel Difallah、Elena Filatova 和 Panos Ipeirotis。2018。机械土耳其机器人工作者的人口统计特征与动态变化。载于第十一届 ACM 网络搜索与数据挖掘国际会议论文集。135–143。

[22] Djellel Eddine Difallah、Gianluca Demartini 和 Philippe Cudré-Mauroux。2012。机械作弊：众包平台上的垃圾信息方案与对抗技术。载于 CrowdSearch。[23] Ryan Drapeau、Lydia Chilton、Jonathan Bragg 和 Daniel Weld。2016。Microtalk：运用论证提升众包准确率。载于 AAAI人类计算与众包会议论文集，第4卷。32–41。[24] Anca Dumitrache、Lora Aroyo 和 Chris Welty。2018。为医学关系抽取众包真实情况。《ACM 智能交互系统汇刊》。 8, 2, 11 2018 20 10 1145 3152889 文章（7月），页码。 2020
[25] https://doi.org/ . / Jenny Fan 和 Amy X. Zhang。。《数字陪审团：以公民参与为导向的平台治理方法》。《2020年 CHI 计算机系统中的人为因素会议论文集》（2020）。[26] Lev Finkelstein、Evgeniy Gabrilovich、Y. Matias、Ehud Rivlin、Zach Solan、Gadi Wolfman 和 Eytan Ruppin。2002。将搜索置于上下文中：重新审视这一概念。《ACM 信息系统汇刊》20卷1期（2002年1月），116–131。https://doi.org/10. 1145/503104.503110[27] Tommaso Fornaciari、Alexandra Uma、Silviu Paun、Barbara Plank、Dirk Hovy 和 Massimo Poesio。2021。超越非黑即白：通过软标签多任务学习利用标注者分歧。载于2021年北美计算语言学协会分会会议：人类语言技术论文集。计算语言学协会，线上，
2591 2597 10 18653 1 2021 204 –。 2011
[28] https://doi.org/ . /v / .naacl-main。Craig R Fox 和 Gülden Ülkümen。。《区分不确定性的两个维度》。Fox, Craig R. 和 Gülden Ülkümen（2011），《区分不确定性的两个维度》，载于《判断与决策研究论文集》，Brun, W.、Kirkebøen, G. 和 Montgomery, H. 编。奥斯陆：Universitetsforlaget（2011）。[29] Ujwal Gadiraju、Besnik Fetahu 和 Ricardo Kawase。2015。培训工作者以提升众包微任务的表现。载于欧洲技术增强学习会议。Springer，100–114。[30] Ujwal Gadiraju、Jie Yang 和 Alessandro Bozzon。2017。清晰度是一项值得追求的品质：任务清晰度在微任务众包中的作用。载于第28届 ACM 超文本与社交媒体会议论文集（HT '17）。计算机协会，美国纽约州纽约市，
5 14 10 1145 3078714 3078715 –。
[31] https://doi.org/ . / 。Timnit Gebru、Jamie H. Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna M. Wallach、Hal Daumé 和 Kate Crawford。2021。数据集数据表。《ACM 通讯》64（2021），86–92。[32] R. Stuart Geiger、Kevin Yu、Yanlai Yang、Mindy Dai、Jie Qiu、Rebekah Tang 和 Jenny Huang。2020。垃圾进，垃圾出？社会计算领域的机器学习应用论文是否报告了人工标注训练数据的来源？《2020年公平性、问责制与透明度会议论文集》（2020）。[33] Tarleton Gillespie。2018。《互联网守护者：平台、内容审核以及塑造社交媒体的隐秘决策》。耶鲁大学出版社。[34] Mitchell L. Gordon、Kaitlyn Zhou、Kayur Patel、Tatsunori Hashimoto 和 Michael S. Bernstein。2021。分歧解卷积：使机器学习性能度量指标与现实相符。载于2021年 CHI 计算机系统中的人为因素会议论文集（CHI '21）。计算机协会，美国纽约州纽约市，文章
388, 14 10 1145 3411764 3445423 页码。https://doi.org/ . / 。Shinsuke Goto、
[35] Toru Ishida 和 Donghui Lin。。《理解众包工作流程：迭代与并行流程的建模和优化》。载于 AAAI人类计算与众包会议。[36] Kevin A. Hallgren。2012。计算观察数据的评定者间信度：概述与教程。《心理学定量方法教程》8卷1期（2012），23–34。[37] Danula Hettiachchi、Mike Schaekermann、Tristan J. McKinney 和 Matthew Lease。2021。可变投入众包的挑战，以及可见金标准如何提供帮助。《ACM 人机交互汇刊》5卷CSCW2期，文章332（10月 2021 26 10 1145 3476073），页码。https://doi.org/ . / Martin Hilbert。。《认知偏差综述：
[38] 嘈杂的信息处理如何导致人类决策偏差》。《心理学公报》138卷2期（2012），211–37。[39] Stephen C. Hora。1996。概率引出中的随机不确定性与认知不确定性：以危险废物管理为例。《可靠性工程与系统安全》54（1996），217–223。[40] Eric Huang、Richard Socher、Christopher Manning 和吴恩达。2012。通过全局上下文和多个词原型改进词表示。载于第50届计算语言学协会年会论文集（第1卷：长篇论文）。计算语言学协会，韩国济州岛，
873 882 12 1092 –。https://aclanthology.org/P - E. Hullermeier 和 W. Waegeman。。
[41] 《机器学习中的偶然不确定性与认知不确定性：概念与方法简介》。arXiv: Learning（2019）。[42] Oana Inel 和 Lora Aroyo。2017。利用群体与机器的多样性提升命名实体识别性能。载于 ESWC。[43] Matthew Ingram。 [不详]。脸书移除那张越南战争照片为何如此重要。Fortune（[不详]。
https://fortune.com/2016/09/09/facebook-napalm-photo-vietnam-war/]Panagiotis G. Ipeirotis、Foster Provost 和 Jing Wang。2010。亚马逊土耳其机器人上的质量管理。载于 ACM 人类计算研讨会论文集（HCOMP '10）。计算机协会，

纽约州纽约市，美国，64 67 10 1145 1837885 1837906 –。https://doi.org/。/。
[45] Jialun Aaron Jiang、Morgan Klaus Scheuerman、Casey Fiesler 和 Jed R. Brubaker。2021。理解国际社会对在线有害内容严重程度的看法。《PLoS ONE》16（2021）。[46] V K. Chaithanya Manam、Dwarakanath Jampani、Mariam Zaim、Meng-Han Wu 和 Alexander J. Quinn。2019。TaskMate：一种提升众包中说明质量的机制。载于《2019年万维网大会论文集》。1121–1130。[47] Sanjay Kairam 和 Jeffrey Heer。2016。人群分化：刻画众包标注任务中的不同解释。载于《第19届ACM计算机支持的协同工作与社交计算大会论文集》。1637–1648。[48] Armen Der Kiureghian 和 O. Ditlevsen。2009。偶然性还是认识论？这重要吗？《Structural Safety》31（2009），105–112。[49] Travis Kriplean、Jonathan T. Morgan、Deen Freelon、Alan Borning 和 Lance Bennett。2011。ConsiderIt：改进结构化公共审议。载于《CHI '11计算机系统中的人为因素扩展摘要》（CHI EA '11）。ACM，纽约州纽约市，美国，1831 1836 10 1145 1979742 1979869 –。https://doi.org/。/。
[50] Ji-Ung Lee、Jan-Christoph Klie 和 Iryna Gurevych。。用于隐式培训非专家标注者的标注课程体系。arXiv abs/2106.02382（2022）。[51] Elisa Leonardelli、Stefano Menini、Alessio Palmero Aprosio、Marco Guerini 和 Sara Tonelli。2021。同意存在分歧：使用标注者分歧标注冒犯性语言数据集。载于《 2021 年自然语言处理实证方法会议论文集》。计算语言学协会，线上及多米尼加共和国蓬塔卡纳，10528 10539 10 18653 1 2021 822 –。https://doi.org/。/v / .
[52] emnlp-main。Dangwei Li、Zhang Zhang、Xiaotang Chen 和 Kaiqi Huang。。用于真实监控场景中行人检索的丰富标注行人数据集。《IEEE图像处理汇刊》28，4（2018），1575–1590。[53] E Allan Lind、John Thibaut 和 Laurens Walker。1973。对抗式与非对抗式诉讼程序中的证据发现与呈现。《密歇根法律评论》71，6（1973），1129–1144。[54] Angli Liu、Stephen Soderland、Jonathan Bragg、Christopher H. Lin、Xiao Ling 和 Daniel S. Weld。2016。用于关系抽取的有效众包标注。载于《北美计算语言学协会与人类语言技术会议 2016论文集》。[55] VK Chaithanya Manam 和 Alexander J Quinn。2018。Wingit：高效完善不清晰的任务说明。载于第六届AAAI人类计算与众包会议。[56] V. K. Chaithanya Manam、Dwarakanath Jampani、Mariam Zaim、Meng-Han Wu 和 Alexander J. Quinn。2019。TaskMate：一种提升众包中说明质量的机制。《2019年万维网大会论文集》（2019）。[57] Emily Megan Marshman、Ryan Thomas Sayer、Charles Henderson、Edit Yerushalmi 和 Chandralekha Singh。2018。改变助教评分实践的挑战：要求学生展示理解证据。《加拿大物理学杂志》96（2018），420–437。[58] Aiden R. McGillicuddy、Jean-Grégoire Bernard 和 Jocelyn Cranefield。2020。控制在线社区中的不良行为：对内容审核工作的考察。载于《国际交互科学会议》。[59] Stefano Menini、Alessio Palmero Aprosio 和 Sara Tonelli。2021。辱骂具有上下文依赖性，那么自然语言处理呢？上下文在辱骂性语言标注与检测中的作用。arXiv abs/2103.14916（2021）。[60] George A. Miller。1995。WordNet：英语词汇数据库。《Commun. ACM》38，11（1995年11月），39–41。https: //doi.org/ 10 1145 219717 219748 。/。Vikram Mohanty、David Thames、Sneha Mehta 和 Kurt Luther。2019。[61] Photo Sleuth：结合人类专业知识与人脸识别来识别历史肖像。载于《第24届国际智能用户界面会议论文集》（IUI '19）。计算机协会，纽约州纽约市，美国，547–557。https://doi.org/10.1145/3301275.3302301[62] Jethro Mullen 和 Charles Riley。 [不适用。]在舆论哗然后，脸书将恢复那张标志性的越南战争照片。《CNN商业》（[不适用）。https://money.cnn.com/2016/09/09/technology/facebook-censorship-vietnam-war-photo/index.html [63] Alexandra Papoutsaki、Hua Guo、Danaë Metaxa-Kakavouli、Connor Gramazio、Jeff Rasley、Wenting Xie、Guan Wang 和 Jeff Huang。2015。从零开始众包：新手需求方进行数据收集的一项实用实验。载于《人类计算与众包会议》。[64] R. Passonneau 和 Bob Carpenter。2013。标注模型的益处。《计算语言学协会汇刊》2（2013），311–326。[65] Ellie Pavlick 和 Tom Kwiatkowski。2019。人类文本推理中固有的分歧。《计算语言学协会汇刊》7 2019 677 694 10 1162 _ _00293 （），–。
[66] https://doi.org/。/tacl a John Pavlopoulos、Jeffrey Sorensen、Lucas Dixon、Nithum Thain 和 Ion Androutsopoulos。。毒性检测：上下文真的重要吗？载于《第58届计算语言学协会年会论文集》。计算语言学协会，线上，4296–4305。https://doi.org/10.18653/v1/2020.acl-main.396

[67] Vinodkumar Prabhakaran、Aida Mostafazadeh Davani 和 Mark Diaz。2021。发布标注者级标签及数据集中的相关信息。载于《第15届语言标注研讨会（LAW）与第3届意义表示设计研讨会（DMR）联合论文集》。计算语言学协会，多米尼加共和国蓬塔卡纳， 133 138 10 18653 1 2021 1 14 –。https://doi.org/。/v / .la w-。Vivek Pradhan、Mike Schaekermann[68] 和 Matthew Lease。2021。探寻模糊性：通过三阶段工作流程设计阐明面向众包工作者的标注指南。arXiv abs/2112.02255（2021）。[69] David L. Rosenhan、Sara L. Eisner 和 Robert J. Robinson。1994。笔记记录有助于陪审员回忆。《法律与人类行为》18（1994），53–61。[70] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael S. Bernstein、Alexander C. Berg 和 Li Fei-Fei。2015。ImageNet大规模视觉识别挑战赛。《国际计算机视觉杂志》115（2015），211–252。[71] Maarten Sap、Dallas Card、Saadia Gabriel、Yejin Choi 和 Noah A Smith。2019。仇恨言论检测中的种族偏见风险。载于《计算语言学协会年会》。https://www.aclweb.org/anthology/P19-1163.pdf[72] Maarten Sap、Swabha Swayamdipta、Laura Vianna、Xuhui Zhou、Yejin Choi 和 Noah A Smith。2021。持有态度的标注者：标注者信念与身份如何使有害语言检测产生偏差。arXiv预印本 arXiv:2111.07997（2021）。[73] Mike Schaekermann、Graeme Beaton、Minahz Habib、Andrew Lim、Kate Larson 和 Edith Law。2019。通过结构化裁定理解医疗数据分析中的专家意见分歧。《Proc. ACM Hum.-Comput. Interact.》3，CSCW，文章 76 2019 23 10 1145 3359178 （11月），页码。https://doi.org/。/ Mike Schaekermann[74]、Joslin Goh、Kate Larson 和 Edith Law。2018。可解决与不可解决的意见分歧：关于众包工作中工作者审议的研究。《Proc. ACM Hum.-Comput. Interact.》2，CSCW，文章154（2018年11月），19页。https://doi.org/ 10 1145 3274423 。/ Arjun Singh、Sergey Karayev、Kevin Gutowski 和 Pieter Abbeel。2017。Gradescope[75]：用于可扩展评估手写作业的快速、灵活且公平的系统。载于《第四届（2017）ACM规模化学习会议论文集》（L@S ' *17* 81 88 10 1145 3051457 3051466 ）。计算机协会，纽约州纽约市，美国，–。https://doi.org/。/。Rion Snow[76]、Brendan O'Connor、Daniel Jurafsky 和吴恩达。2008。便宜且快速——但质量好吗？评估自然语言任务中的非专家标注。载于《2008年自然语言处理实证方法会议论文集》。计算语言学协会，檀香山，夏威夷，254–263。https: //aclanthology.org/D 08 1027 - Robert Soden 、 Laura Devendorf 、 Richmond Y. Wong 、 Yoko Akama 和 Ann Light[77]。2022。HCI 中的不确定性模式。《Found. Trends Hum. Comput. Interact.》15（2022），317–426。[78] Thamar Solorio、Ragib Hasan 和 Mainul Mizan。2014。维基百科中的傀儡账号检测：用于身份关联的真实世界欺骗性写作语料库。载于《第九届语言资源与评测国际会议论文集》（LREC'14）。欧洲语言资源协会（ELRA），冰岛雷克雅未克，1355–1358。http://www.lrec-conf.org/proceedings/lrec2014/pdf/1007_Paper.pdf[79] George Stoica、Emmanouil Antonios Platanios 和 Barnab'as P'oczos。2021。Re-TACRED：解决TACRED数据集的不足。载于《AAAI人工智能会议》。[80] Nicolas Suzor 和 Darryl Woodford。2013。不断变化的社区规范下的同意与合法性评估：EVE Online 案例研究。Suzor, Nicolas P. & Woodford, Darryl（2013）不断变化的社区规范下的同意与合法性评估：EVE Online案例研究。《虚拟世界研究杂志》6，3（2013），1–14。[81] Swabha Swayamdipta、Roy Schwartz、Nicholas Lourie、Yizhong Wang、Hannaneh Hajishirzi、Noah A. Smith 和 Yejin Choi。2020。数据集制图：利用训练动态对数据集进行映射与诊断。载于《2020年自然语言处理实证方法会议论文集》（EMNLP）。计算语言学协会，线上， 9275 9293 10 18653 1 2020 746 –。https://doi.org/。/v / . emnlp-main[82]。R Peter Terrebonne[83]。1981。普通法的严格演化模型。《法律研究杂志》10，2（1981），397–407。[83] Craig Thorley、Lara Beaton、Phillip Deguara、Brittany Jerome、Dua Khan 和 Kaela Schopp。2020。模拟陪审团评议期间遇到的错误信息可能扭曲陪审员对审判的记忆并使其裁决产生偏差。《法律与犯罪心理学》25（2020），150–164。[84] Amos Tversky 和 Daniel Kahneman。1974。不确定条件下的判断：启发式与偏差。《科学》185（1974），1124–1131。[85] Jeroen Vuurens、Arjen P de Vries 和 Carsten Eickhoff。2011。你能接受多少垃圾信息？一项通过提高准确率分析众包结果的研究。载于《ACM信息检索众包研讨会论文集》（CIR'11）。21–26。[86] Warren E. Walker、Poul Harremoës、Jan Rotmans、Jeroen P. van der Sluijs、M. B. A. Asselt、Paul Janssen 和 Martin Krayer von Krauss。2003。定义不确定性：基于模型的决策支持中不确定性管理的概念基础。《综合评估》4（2003），5–17。

[87] Dongsheng Wang、Prayag Tiwari、Mohammad Shorfuzzaman 和 Ingo Schmitt。2021。利用众包中的标签分歧对加权数据集进行深度神经学习。《计算机网络》196 (2021), 108227. https: //doi.org/ 10 1016 2021 108227 . /j.comnet. . Zeerak Waseem。 . 你是种族主义者，还是我想多了？标注者对[88] Twitter 仇恨言论检测的影响。载于第一届自然语言处理与计算社会科学研讨会论文集。计算语言学协会，奥斯汀，得克萨斯州， 138 142 10 18653 1 16 5618 – 。https://doi.org/ 。 /v /W - 2016 [89] Chris Welty、Lora Mois Aroyo 和 Praveen Kumar Paritosh。 . 评估众包仪器的计量学框架。载于*HCOMP-2019*: *AAAI*人类计算会议。[90] Mark E. Whiting、Allie Blaising、Chloe Barreau、Laura Fiuza、Nik Marda、Melissa Valentine 和 Michael S. Bernstein。2019。结局非得如此吗？理解团队分裂的一致性。《*ACM*人机交互汇刊》 3, 209 2019 23 10 1145 3359311 CSCW，文章（11月），页码。[91] https://doi.org/ 。 / Janyce M. Wiebe、Rebecca F. Bruce 和 Thomas P. O'Hara。 . 主观性分类金标准数据集的开发与使用。载于计算语言学协会第*37*届年会论文集。计算语言学协会，学院公园市，马里兰州，美国，246–253。 1999 https://doi.org/10. 3115/1034678.1034721[92] Meng-Han Wu 和 Alexander J. Quinn。2017。让众包参与者困惑：亚马逊土耳其机器人上的任务指令质量。载于人类计算与众包会议。[93] Shuicheng Yan、Huan Wang、Thomas S. Huang、Qiong Yang 和 Xiaoou Tang。2007。基于不确定标签的排序。*2007 IEEE*国际多媒体与博览会会议(2007), 96–99。[94] Hao-Yu Yang、Junling Yang、Yue Pan、Kunlin Cao、Qi Song、Feng Gao 和 Youbing Yin。2019。学会保持不确定性：利用贝叶斯神经网络处理胸部X光片中的不确定标签。载于*CVPR*研讨会。[95] Amy X Zhang、Bryan Culbertson 和 Praveen Paritosh。2017。利用粗粒度话语序列刻画在线讨论。载于第十一届*AAAI*网络与社交媒体国际会议。[96] Yuhao Zhang、Victor Zhong、Danqi Chen、Gabor Angeli 和 Christopher D. Manning。2017。位置感知注意力和监督数据提升槽位填充效果。载于*2017*年自然语言处理实证方法会议论文集。计算语言学协会，哥本哈根，丹麦，35–45。https://doi.org/10.18653/v1/D17- 1004