衡量解释偏差的影响：推荐系统自然语言解释研究

Krisztian Balog 谷歌
挪威斯塔万格
krisztianb
@google.com

Filip Radlinski 谷歌 英国伦敦
filiprad
@google.com

Andrey Petrov 谷歌 英国伦敦
apetrov
@google.com

arXiv:2303.09498v1 [cs.HC] 16 Mar 2023

摘要

尽管解释可能对决策产生影响，但目前缺乏量化其对用户选择影响的研究。本文提出了一种实验方案，用于衡量带有正面或负面偏差的解释在多大程度上会导致用户选择次优推荐。该方案的关键要素包括：偏好引出阶段，以便实现推荐个性化；从评论中手动识别并提取物品方面；以及一种通过结合正面和负面方面来引入偏差的受控方法。我们以两种不同的文本格式研究解释：物品方面列表和流畅的自然语言文本。通过一项涉及129名参与者的用户研究，我们证明解释会显著影响用户的选择，而且这一发现适用于不同的解释格式。

CCS 概念

• **以人为中心的计算** → **用户研究**；**HCI 中的实证研究**；• **信息系统** → **推荐系统**。

关键词

可解释推荐；自然语言理由说明；评估解释；解释类型；解释偏差

ACM 引用格式：

Krisztian Balog、Filip Radlinski 和 Andrey Petrov。2023。解释偏差影响的测量：推荐系统自然语言解释的研究。载于*2023*年 *CHI*计算机系统中的人为因素会议扩展摘要（*CHI EA '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，纽约州纽约，美国，8页。https://doi.org/10. 1145/3544549.3585748

1 引言

推荐系统已遍布现代社会，因此人们对所提供的推荐有多大程度的信任，引发了许多研究 [14, 20]。因此，人们高度关注如何为这些系统配备解释功能，以帮助用户做出明智的决定 [25, 31, 34]。在最早的研究之一中，Herlocker 等人 [17] 表明，解释能够提高人们采纳推荐的可能性。然而，研究推荐的最终满意度时发现，某些类型的解释会导致用户高估或低估推荐项目的真实价值 [3, 11]。尽管已知解释能够显著影响用户的决策过程 [31], , 但在衡量和理解其特效方面仍存在重大空白 [28]。大多数关于解释的研究都聚焦于一个主观问题：人们如何看待这些推荐 [1, 5, 11, 22, 25]。对于如何客观量化解释对人们选择的影响程度，关注则相对较少。这与说服力这一概念密切相关，即解释"使用户接受或忽略某些物品的能力" [11]。以往研究区分了促使人们高估或低估的解释类型 [3, 11]。相较之下，我们将定量而细致地研究自然语言解释的内容偏差如何影响用户选择。

与以往工作的一个关键区别在于，先前研究几乎只关注突出物品正面方面的解释，即支持某项选择，或许会通过指出该物品与用户此前表示喜欢的其他物品相似之处来做到这一点。解释也可以通过强调为什么不 应选择某物来帮助用户做出决策。我们首次在受控条件下开展研究，有意使解释偏向物品的正面或负面方面；研究情境中的物品，其相关性高低已有先验认知。因此，我们能够量化有偏的正面或负面解释在多大程度上会导致用户选择次优推荐。为此，我们受到 Musto 等人 [22] 的研究启发，他们"处理并分析评论，从中提取出一组经常在评论中讨论的特征（当然带有正面情感倾向），这些特征能够促使用户喜欢上推荐项目。"然而，我们既识别正面方面，也识别负面 方面——正如 Bilgic 和 Mooney [3],所指出的："解释的目标不应是向用户'推销'物品，而应是帮助用户做出知情判断。"

我们要探讨的首要研究问题是：当用户面对一组可供选择的推荐项时，如何量化解释对用户决策的影响程度，以及这种影响是正面的还是负面的。我们的主要贡献是设计了一套实验方案，用于定量研究解释对用户选择的影响。第二项贡献是对从129名受试者那里获得的结果进行初步分析。我们提供了基线统计数据，展示用户在面对有偏解释时，有多大可能最终选择相关性较低的推荐项。除此之外，我们还研究了解释格式的影响，直接比较了两种常见格式，即物品方面列表和自然语言文本。我们发现

允许为个人或课堂使用免费制作本作品的部分或全部内容的数字或纸质副本，前提是这些副本不得用于营利或商业用途，不得分发，并且每份副本都须附有本声明及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*CHI EA '23*，*2023*年*4*月*23*日至*28*日，德国汉堡 © 2023 版权所有者/作者所有。ACM ISBN 978-1-4503-9422-2/23/04。https://doi.org/10.1145/3544549.3585748

尽管这两种格式导致的整体行为相似，但与自然语言文本相比，方面列表中的解释偏差影响更为显著。在设计可解释推荐系统时，可以将这些因素纳入考量。总体而言，这项工作提供了一个衡量和优化解释的框架，以帮助引导用户做出明智的决策。

2 相关工作

关于可解释推荐的研究聚焦于不同维度，包括：（1）解释目标（例如，透明度、信任、有效性、说服力）；（2）风格（例如，基于内容、基于协同、基于知识/效用）；（3）范围（例如，用户模型、过程、推荐项目）；以及（4）格式（例如，文本、视觉）[25, 31, 34]。另一种分类方式是依据生成解释的方法：模型本身可解释的推荐模型与事后解释（即，理由说明）[4]。我们的方法可归类为基于内容的事后自然语言解释生成。我们的目标并非解释为何推荐某个物品（why），而是向用户提供额外信息（物品的相关特征），以帮助他们进行决策。

过去曾研究过多种图形和文本解释格式 [11, 17],，其中自然语言历来 [25] 以及近年来 [1, 5, 22, 23, 27]都是最受欢迎的形式。基于文本的解释形式多种多样，从标签或关键词 [3, 32] 到一个或多个句子 [1, 5, 22, 27, 30]。在本研究中，我们将简短的描述性短语（方面）和流畅的自然语言摘要都视为解释。我们的方法独特之处在于，摘要直接根据这些方面生成，从而能够直接比较这两种解释格式。

评论已被用于改进推荐，由此推动了评论感知推荐系统相关研究 [6,18],，使这类系统自然具备更高程度的可解释性和透明度 [16]。相比之下，我们的工作延续了利用评论生成事后推荐理由说明的研究脉络，并将推荐引擎视为黑箱 [5, 21–23, 27]。Muhammad 等人 [21] 突出展示物品中最重要、可能与用户相关的特征（优缺点），这些特征是根据用户自己的评论确定的。同样，Chen 和 Wang [7, 8] 提取用户对特定产品属性的情感倾向，并据此向用户提供给定推荐的替代选项，同时解释其中的权衡，也就是哪些属性会得到改善，哪些属性则会有所妥协。Chang 等人 [5] 采用多步骤流程，通过众包生成自然语言解释：（1）改进算法生成的标签簇；（2）综合评论文本，为这些标签簇撰写解释；（3）通过投票选出最佳解释。我们采用了类似的方法，但为众包工作者安排的任务简单得多：给定一组评论，他们只需从中提取正面和负面方面。随后，我们会将这些方面以列表形式呈现，或者运用最先进的神经语言建模技术，将它们转化为流畅的自然语言文本。Ni 等人 [23] 找出可用作理由说明的评论片段，并探索使用神经语言模型生成令人信服且多样化的理由说明。Penha 等人 [27] 从物品的评论中选出一个有帮助的句子，该句由分类器预测得出，并被用于生成关于某个物品或一对物品的基于模板的解释。与我们的工作最为接近的是 Musto 等人 [22], 的研究，他们提出了一个完全自动化的流程，用于生成自然语言理由说明：（1）提取一组刻画物品的方面；（2）确定其中最相关的方面；（3）提取并汇总讨论这些方面的评论句子。尽管解释生成流程存在相似之处，他们关注的是自动化的算法层面，而我们关注的是解释的影响 。我们的方面是简短的描述性短语，并不像 [22], 中的方面那样局限于名词；这些短语由我们手动提取和整理。最重要的是，我们在解释中也纳入具有负面情感的方面，而不只包含正面方面。

解释评估的目标可以是界面相关方面 [3, 8, 11, 17, 21], 、生成方法中可量化的属性（例如，自然语言的可读性 [10] 或事后解释的忠实度 [26]），也可以通过问卷 [1, 5, 11, 22, 25] 评估对质量的主观感知（涉及透明度、信任、有效性等）。与本研究最相关的是 [3]提出的（ [11]也采用了）评估方案，该方案通过比较两次评分的差异来衡量解释类型的说服力：首次评分依据解释给出，第二次评分则在"消费"物品后给出。研究通过向受试者提供关于物品的更多详情（描述或用户评论），来近似模拟消费过程。随后，受试者需要根据这些额外信息调整评分。我们的实验设计更贴近现实，因为我们不假设用户会消费物品，而是通过用户做出的物品选择来衡量解释的影响；在此之前，我们会了解他们最有可能选择什么。

3 研究设计

本节介绍我们设计的研究，旨在回答以下两个研究问题：

- **研究问题 1** 当用户看到一组物品推荐时，解释在多大程度上会影响他们的决策?
- **研究问题 2** 解释的呈现形式对结果有多大影响?

我们的实验围绕一组物品选择任务展开，参与者需要从三个作为推荐呈现给他们的物品中进行选择。这些物品根据初始的偏好引出阶段选出，因此我们知道用户最可能喜欢三个物品中的哪一个，以及最不喜欢哪一个。我们通过解释对随机选中的物品施加受控偏差量，从而衡量用户的选择会在多大程度上受到影响——特别值得关注的是，他们有多常会倾向于选择不同于统计意义上最可能选择的物品。采用被试内设计，参与者会接触不同的实验条件，其中偏差量和解释形式各不相同。

3.1 概述

本研究以电影领域为背景。观看电影是最主要的媒体消费和娱乐活动之一，

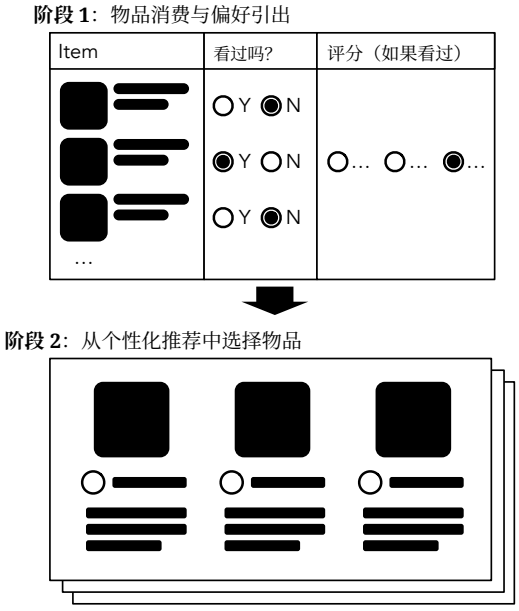

图1：我们的实验设计概览。

大多数人都能轻松理解这类内容，无需接受任何培训或具备先决知识。事实上，电影一直是可解释推荐研究最多的领域之一[1, 2, 9, 11, 12, 17, 22, 30, 32]。请注意，尽管这些说明专为电影量身定制，但所提出的设计具有通用性，也可应用于其他领域。

为了能够向参与者提供个性化推荐，该研究分为两个阶段，如图1所示。[1]在阶段 1中，参与者针对候选物品池中的400部电影，逐一说明自己是否看过；如果看过，则说明喜爱程度（采用由不喜欢、中性和喜欢构成的三点评分量表）。在阶段2中，参与者按照随机顺序完成总计12项物品选择任务，对应于各种实验条件（参见表1）。在每项任务中，系统会提供三部电影供其选择，这些电影均从他们尚未观看过的物品中选出。对于每部电影，都会显示片名、海报和剧情简介，之后还会（可选地）提供解释（见附录A.3中的图6）。物品选择任务设置在以下场景中：想象一下，您正在独自旅行，在途中，某项服务向您提供了一次免费电影租赁，但该服务只允许您从三个可能的选项中进行选择。在接下来的每项任务中，请根据所提供的信息，选择您更愿意观看的电影。

我们假设，推荐中附带的解释会影响用户的物品选择。为验证这一假设，我们将基线设置（为全部三件物品显示中性解释）与"偏差"设置进行对比：后者为其中两件物品显示中性解释，同时以受控方式让第三件物品的解释整体上更正面或更负面。接下来，我们将详细介绍实验设计的各个要素。

[1]此外，我们还要求参与者填写任务后问卷，以便日后对他们的偏好进行定性分析。

表1：实验条件。每位参与者都会体验所列的每种条件，且各体验一次（顺序随机）。

ID	Bias	解释	方面$^+$	方面$^-$
基线				
#1 #2 #3 #4	无偏差	-	-	-
		分项列出 流畅自然语言	2	2
一个随机物品受到更正面/负面的解释的偏向				
#5 #6	偏差 +	分项列出 流畅自然语言	3	1
#7 #8	偏差 -	分项列出 流畅自然语言	1	3
#9 #10	偏差 ++	分项列出 流畅自然语言	4	0
#11 #12	偏差 --	分项列出 流畅自然语言	0	4

3.2 物品收集

我们使用 MovieLens 25M 数据集 [15] 以及 Amazon Reviews 2018 数据集中的电影与电视子集 [23]。两者之间的映射由 [33] 发布的 Reviews2Movielens 数据集（v2 映射）提供。我们采用 [2] 中使用的分层抽样方法，并稍作修改，抽取共 400 部电影：（i）按收到的评分数量计算出的历来最热门的 150 部电影；（ii）对于 1992 年至 2016 年之间的每一年，以及最热门的 10 种类型（动作、冒险、纪录片、喜剧、犯罪、剧情、恐怖、爱情、科幻和惊悚）中的每一种，各随机抽取一部尚未纳入前 150 部电影的电影。[2] 我们要求每部电影至少有 100 条相关评论。

3.3 生成个性化推荐

研究参与者会看到一组物品选择任务。在每项任务中，他们需要从三个物品 x_a、x_b x_c 和 中做出选择，这些物品以随机顺序呈现。根据初始偏好引出阶段，这些物品经过筛选，使用户估计的偏好排序为 $x_a > x_b > x_c$。为此，我们采用由三种成熟推荐算法组成的集成，这些算法代表不同类别的协同过滤方法，并在先前一项关于电影推荐的用户研究中表现最佳 [2]：基于物品的 k 近邻 [29], 加权正则化矩阵分解 [19], 以及稀疏线性方法 [24]。这些算法的预测结果通过基于共识的投票系统（波达计数法）合并为集成推荐。预计该集成的表现会优于任何单一推荐算法。至关重要的是，只有用户尚未看过的电影才有资格获得推荐，因为展示用户已经看过的电影可能会影响对解释影响的测量 [1]。

假设每位参与者需要完成共计 $m = 12$ 个物品选择任务（参见表1），$m \times 3$ 个物品按以下方式抽取。[3]令 n 表示此人尚未看过的电影数量，

[2]如果无法做到，则从同一年随机抽取一部电影。[3]尚未看过的电影少于 3×12 部的参与者将被排除在研究之外。

	A桶					B桶					C桶				
项目排名	1	. . .	i	. . .	s	$s+1$	. . .	$s+i$	. . .	$2s$	$2s+1$	. . .	$2s+i$	. . .	n

图2：为项目选择任务采样物品，确保所有项目样本中的采样物品之间的距离最大。

有趣、傻气、甜美 精彩的喜剧 有几处小笑点，但还远远不够 平淡无奇的导演	一部好看、干净又刺激的电影 充满动作场面 适合全家观看的娱乐作品 对喜剧的可预测尝试	扣人心弦 一部糟糕的翻拍作品 情节可预测 低成本特效
据说这部电影有趣、傻气又甜蜜，而且是一部精彩的喜剧。不过，也有人说它有几个小笑点，但还不够，而且导演水平平平。	据说这部电影不错、干净，而且令人愉快——引用称其充满动作场面，是一部娱乐作品适合全家观看，但有些人觉得它是一次可预见的喜剧尝试。	据说这部电影扣人心弦，但另一些人——觉得它是部糟糕且情节老套的翻拍作品，特效制作成本低。

图 3：基于同一组方面的分项列出式和流畅自然语言解释示例。

表 2：根据提及的正面和负面方面数量，对解释情感进行受控生成。

	极端负面（偏差 --）	负面（偏差 -）	中性（无偏差）	正面（偏差 +）	极端正面（偏差 ++）
方面$^+$	0	1	2	3	4
方面$^-$	4	3	2	1	0

这些电影按推荐分数排序，其中 x_1 是排名最高的推荐项，x_n 是排名最低的推荐项。物品被分为三个大小大致相等（$s = \lfloor n/3 \rfloor$）的类别 A、B 和 C。为抽取 m组三个物品，集合 $i \in [1..m]$ 中选出的物品为：$x_a = x_i$、$x_b = x_{s+i}$和 $x_c = x_{2s+i}$。这样可以确保在所有物品选择任务中，展示给用户的三个物品在推荐分数方面彼此尽可能相距最远；视觉解释见图2。m 组物品（x_a、x_b和 x_c）被随机分配到 m 个实验条件中。

3.4 生成解释

为了量化解释对用户决策的影响程度，我们需要一种受控的解释生成方式，使解释默认保持中性，但也可以加入偏差，使整体情感倾向更加正面或负面。为了具体界定解释整体呈中性/正面/负面所需的条件，我们假设对于每个物品 $M = 4$，正面和负面方面均已识别。此处的方面是简短的自然语言文本（通常为 1–5 个词），用于表达人们可能喜欢或不喜欢一部电影的原因，例如“充满动作场面”“相当老套且不真实”或“对经典文学作品的精彩演绎”。随后，可以通过调整解释中提及的正面和负面方面的数量来控制其情感倾向，其中 $\text{Aspects}^+ = \text{Aspects}^- = M/2$ 表示中性解释（参见表 2）。方面从评论中提取，详见下文第 3.4.1 节。

我们考虑两种解释格式：分项列出和流畅自然语言文本。分项列出解释由一系列方面及其对应的情感倾向组成，并以赞或踩图标表示。另一种方式是，流畅自然语言格式将相同信息呈现为流畅自然语言描述；示例见图3。有关如何根据物品方面列表生成流畅自然语言解释，详见第3.4.2节。

3.4.1 方面提取。 通过众包手动提取正面和负面方面。参与者会看到10条正面评论（4–5星）和10条负面评论（1–2星），并被要求找出能够补全以下句子的方面：这部电影可能会受到喜欢（不喜欢），因为它是/有/包含 ___。目标并非面面俱到，而是着重保证数据质量（即优先考虑精确率而非召回率）。为此，论文作者会进一步检查提取出的方面，并手动筛除过于苛刻或冒犯性的措辞、过于笼统的方面（例如“优秀”“糟糕”），以及与电影本身无关的方面（涉及价格、配送、媒介等）。筛选后，正面和负面方面各少于4个的电影会从推荐池中移除（共31部）。更多详情见附录A.1。

3.4.2 流畅自然语言解释生成。 我们使用少样本性能达到最先进水平的大语言模型（PaLM [9], 62b参数模型），将选定的方面转化为流畅自然的句子。我们为每种方面组合分别创建了一个提示，其中包含三个手工编写的训练示例。

3.5 参与者条件分配

本研究通过网络平台进行众包。每位参与者都会按照随机顺序，恰好一次体验表1所示的12种实验条件。事实上，只有11种独特条件，因为#1和#2相同；但为了建立稳健的基线，这种“无解释”设置会呈现两次。另一种基线是以两种不同格式（#3–#4）呈现中性解释。在其余条件（#5–#12）中，向评定者展示的三件物品中有一件会被随机选为“带有偏差”的物品，方法是改变随该物品提供的解释中正面和负面方面的数量。其余两件物品的解释保持中性，即恰好包含两个正面方面和两个负面方面。这些条件使我们能够比较不同类型的解释，即无解释（#1–#2）、分项列出（#3、#5、#7、#9、#11）和流畅自然语言（#4、#6、#8、#10、#12）解释；也能够比较偏差程度，即无偏差（#1–#4）、中等偏差（#5–#8）和较大偏差（#9–#12）。

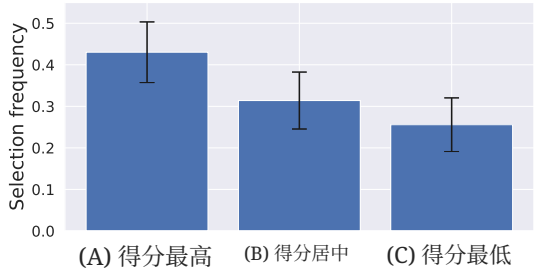

图 4：基线条件，无解释。

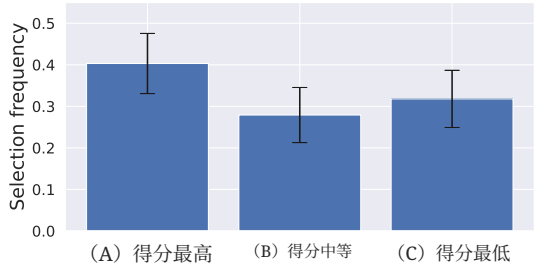

图5：基线条件，中性解释。

我们承认，先列出正面方面、再列出负面方面，或反过来，可能会产生影响。同时，为避免通过引入另一个维度而使设计进一步复杂化，我们通过平衡两种可能的设置来控制这一变量，即先列正面方面和先列负面方面。具体而言，当同时显示正面和负面方面时（#3–#8），我们采用类似拉丁方的设计，轮流使用预先设定的序列，使其在每位参与者内部以及所有参与者之间都完全平衡（见附录 A.2 中的表5）。

4 结果

现在，我们展示定量结果，说明推荐附带的偏差性解释如何直接影响用户通过物品选择所表达的偏好。

4.1 参与者

共有 129 名参与者参加了我们的研究，他们是领取标准合同工资的承包人员（符合其就业所在国家的生活工资法律）。他们均居住在美国，且以英语为母语。性别分布为：女性占 62.7%，男性占 35.7%，1.6% 的人选择不透露。年龄分布为：18–24 岁占 5.6%，25–34 岁占 29.4%，35–44 岁占 32.5%，45–54 岁占 19.8%，55–64 岁占 9.5%，65 岁以上占 3.2%。他们自我报告的每周平均观影时长为：5.7% 的人为 <2 小时，37.9% 的人为 2–5 小时，29% 的人为 6–10 小时，16.9% 的人为 11–16 小时，10.5% 的人为 >16 小时。参与者平均用 41 秒在阶段 1 对一批五部电影进行评分（看过/喜欢），并用 43 秒从阶段 2 的三条推荐中选择他们想看的电影。

4.2 解释存在与否的影响

请回想，在每个实例中，参与者都会获得三项推荐，而我们知道评定者最应该偏好哪一项。我们的首项检验旨在验证这些偏好是否遵循预期顺序，以及中性解释的存在是否会影响这一顺序。具体而言，设三项推荐分别为$x_a \in A$、$x_b \in B$ 和 $x_c \in C$，其中 A、B和 C 分别是包含得分最高、中等和最低物品的类别（参见图2）。设每个类别被选中的相对频率（对应于用户的物品选择）为其选择频率

分别记为 p_A、p_B 和 p_C。由此可得预期排序为 $p_A > p_B > p_C$。

不显示解释时，我们发现，所有参与者中，$p_A = 0.43 \pm 0.07$、$p_B = 0.31 \pm 0.07$ 和 $p_C = 0.26 \pm 0.06$ 的 95% 置信区间使用Goodman方法计算得出 [13]；见图4。因此，预期排序成立，得分最高和最低的类别之间的差异在统计上显著。

纳入中性解释后，我们观察到

$p_A^n = 0.4 \pm 0.07$、$p_B^n = 0.28 \pm 0.07$ 和 $p_C^n = 0.32 \pm 0.07$，如图所示。虽然得分最高的类别仍然获得最多的选择，但令人惊讶的是，B桶和C桶的相对排序现在颠倒了。此外，三个类别的选择频率彼此更加接近，导致置信区间出现重叠。这意味着，尽管实验设计周密，解释似乎仍会产生一些未受控制的特效。例如，得分中等和较低的物品可能并未被推荐系统很好地区分开来，因此这可能是数据噪声造成的随机效应。然而，没有解释的“表现良好”的基线设置以及相对较大的样本量（n=258）表明，情况并非如此。另一种可能是，突出正面和负面方面的中性解释会促使用户做出更多“冒险”的选择，尝试排名较低的推荐项。在后续分析中，我们将中性解释设置（图5）作为基线。不过，B桶和C桶的相对排序仍值得进一步探究。

4.3 偏向正面或负面的解释效果

通过调整正面和负面方面的数量，我们可以使解释呈现正面或负面偏差，程度既可以较弱（例如，三个正面方面和一个负面方面），也可以较强（例如，四个负面方面且没有正面方面）。表3展示了偏差对物品选择（行）的影响，具体取决于偏差的位置和导演（列）。为简单起见，我们不区分偏差程度（轻微或强烈的正面/负面偏差），也不区分解释类型（流畅自然语言或分项列出），而是报告汇总计数。

某个分组中的偏差会影响同一分组中的选择（表3中以灰色突出显示），也会影响其他分组。例如，对 C 中的物品施加正向偏差会增加对 C 的选择，但也会减少对 A 和 B 的选择。同样，对 A 中的物品施加负向偏差会减少对 A 的选择，同时增加对 B 的选择。这符合直觉，但也有少数例外，无法观察到预期的行为，例如，负面

表3：不同类别（行）的选择频率，取决于解释中偏差的有无及其方向（列）。灰色单元格背景表示偏差发生在与选择相同的类别中。绿色/红色箭头表示相对于无偏差设置，选择频率的变化。

选择	无偏差	A中的偏差		B中的偏差		C中的偏差	
		+/++	-/--	+/++	-/--	+/++	-/--
A	0.40 ± 0.07	0.49 ± 0.09▲	0.37 ± 0.09▼	0.37 ± 0.09▼	0.41 ± 0.09▲	0.36 ± 0.08▼	0.50 ± 0.09▲
B	0.28 ± 0.07	0.25 ± 0.08▼	0.34 ± 0.08▲	0.42 ± 0.09▲	0.30 ± 0.08▲	0.27 ± 0.08▼	0.29 ± 0.08▲
C	0.32 ± 0.07	0.26 ± 0.08▼	0.29 ± 0.08▼	0.21 ± 0.08▼	0.28 ± 0.08▼	0.38 ± 0.09▲	0.21 ± 0.07▼

表4：根据解释中偏差的存在与方向（列），分项解释（上方区块）与流畅自然语言解释（下方区块）中不同类别（行）的选择频率；单元格注释与表3相同。

选择	无偏差	A中的偏差		B中的偏差		C中的偏差	
		+/++	-/--	+/++	-/--	+/++	-/--
条目式解释							
A	0.40 ± 0.10	0.49 ± 0.13▲	0.35 ± 0.11▼	0.39 ± 0.12▼	0.47 ± 0.13▲	0.34 ± 0.12▼	0.55 ± 0.12▲
B	0.28 ± 0.09	0.26 ± 0.11▼	0.34 ± 0.11▲	0.45 ± 0.13▲	0.19 ± 0.11▼	0.26 ± 0.11▼	0.27 ± 0.11▼
C	0.32 ± 0.10	0.25 ± 0.11▼	0.31 ± 0.11▼	0.16 ± 0.10▼	0.33 ± 0.12▲	0.40 ± 0.12▲	0.19 ± 0.10▼
流畅自然语言解释							
A	0.40 ± 0.10	0.49 ± 0.12▲	0.40 ± 0.13	0.34 ± 0.12▼	0.36 ± 0.12▼	0.38 ± 0.12▼	0.45 ± 0.12▲
B	0.28 ± 0.09	0.24 ± 0.10▼	0.33 ± 0.12▲	0.39 ± 0.13▲	0.40 ± 0.12▲	0.27 ± 0.11▼	0.31 ± 0.12▲
C	0.32 ± 0.10	0.26 ± 0.10▼	0.27 ± 0.12▼	0.27 ± 0.12▼	0.24 ± 0.11▼	0.35 ± 0.12▲	0.23 ± 0.11▼

偏差在 A 中会增加对 B 的选择，但不会增加对 C 的选择。尽管这些异常仍有待未来研究，但显然，对解释施加偏差会极大地影响人们做出的选择。值得注意的是两个极端情况：（1）对最相关的推荐项施加负向偏差，会使人们对其认为最佳推荐项的选择率降低 9%；（2）对相关性最低的推荐项施加正向偏差，其选择率会提高一倍，较无偏差基线高出近 19%。请注意，带有正向偏差的相关性最低推荐项的选择频率（0.38 ± 0.09），达到了带有负向偏差的最相关推荐项的选择频率（0.37 ± 0.09）。

4.4 解释格式的影响：流畅自然语言与条目式解释

在我们的设计中，解释以两种方式呈现：一种是属性列表，另一种是提及相同属性的流畅文本。我们提出的最后一个问题是：我们的发现在多大程度上取决于特定的解释格式？表4按解释格式细分了先前的结果。主要发现如下。首先，在无偏差设置下，结果没有差异。其次，当解释存在偏差时，条目式解释的表现更"符合预期"：也就是说，当偏差与选择发生在同一分组中（灰色单元格）时，正面偏差总是会提高选择频率，而负面偏差总是会降低选择频率。流畅自然语言解释则并非如此。第三，我们发现，在给定分组中，正面偏差与负面偏差设置之间的差异，在条目式解释的情况下往往大得多。这在直觉上说得通，因为正面和负面方面通过可视化点赞/点踩图标明确呈现，而流畅文本则更容易掩盖差异。尽管存在这些差异，结果仍表明，两种解释格式下的变化模式一致：18个单元格中有14个单元格的变化箭头指向相同方向。

5 结论

我们设计了一套研究协议，可用于定量研究解释对用户在物品推荐中选择的影响。该设计的关键要素包括：通过偏好引出来生成个性化推荐；手动识别并提取要纳入解释的物品方面；通过组合正面和负面方面，以受控方式引入偏差；以及以两种不同的文本格式呈现解释。我们开展了一项用户研究，结果表明，解释确实会显著影响人们作出的物品选择，而且这些发现适用于两种解释格式。研究结果还带来了一些出乎意料的发现，值得在未来的工作中进一步探究。我们还计划对结果进行更详细的统计分析，并根据问卷后回复开展定性评估。此外，条目式解释与流畅自然语言解释在绝对影响程度上的差异表明，后者的具体措辞可能会起作用。衡量措辞上的细微差别是否会产生影响，是一个值得未来研究探讨的课题。最后，我们聚焦于电影推荐，但我们的方法也可推广至其他领域：在这些领域中，由于物品集合规模庞大，用户会依赖自动建议（例如书籍、音乐和食谱）。在其他领域重复这一实验将会很有意义。

参考文献

[1] Krisztian Balog 和 Filip Radlinski。2020 年。《衡量推荐解释质量：解释目标之间的冲突》。载于第 *43* 届 *ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR '20*）。329–338。[2] Krisztian Balog、Filip Radlinski 和 Shushan Arakelyan。2019 年。《用于个性化推荐的透明、可审查和可解释用户模型》。载于第 *42* 届 *ACM SIGIR*信息检索研究与发展会议论文集（*SIGIR'19*）。265–274。

[3] Mustafa Bilgic 和 Raymond J. Mooney。2005 年。《解释推荐：满意度与推广》。载于《Beyond Personalization 2005：2005 年智能用户界面国际会议推荐系统研究下一阶段研讨会论文集》。13–18。[4] Or Biran 和 Courtenay V. Cotton。2017 年。《机器学习中的解释与论证：综述》。载于 IJCAI 2017 可解释人工智能研讨会。[5] Shuo Chang、F. Maxwell Harper 和 Loren Gilbert Terveen。2016 年。《基于群体的个性化自然语言推荐解释》。载于第 10 届 ACM推荐系统会议论文集（RecSys '16）。175–182。[6] Li Chen、Guanliang Chen 和 Feng Wang。2015 年。《基于用户评论的推荐系统：最新进展》。用户建模与用户适应交互 25, 2（2015 年 6 月），99–154。[7] Li Chen 和 Feng Wang。2014 年。《产品推荐的情感增强解释》。载于第 23 届万维网国际会议论文集（WWW '14）。239–240。[8] Li Chen 和 Feng Wang。2017 年。《基于产品评论中特征情感的推荐解释》。载于第 22 届智能用户界面国际会议论文集（IUI '17）。17–28。[9] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vinodkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael Isard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。2022 年。《PaLM：借助 Pathways 扩展语言建模》。arXiv: 2204 02311 [ ]。cs.CL Felipe Costa、Sixun Ouyang、[10] Peter Dolog 和 Aonghus Lawlor。2018 年。《自然语言解释的自动生成》。载于第 23 届智能用户界面国际会议伴随论文集（IUI '18）。[11] Fatih Gedikli、Dietmar Jannach 和 Mouzhi Ge。2014 年。《我该如何解释？推荐系统中不同解释类型的比较》。Int. J. Hum. Comput. Stud. 72, 4（2014 年），367–382。[12] Azin Ghazimatin、Soumajit Pramanik、Rishiraj Saha Roy 和 Gerhard Weikum。2021 年。《ELIXIR：从用户对解释的反馈中学习以改进推荐模型》。载于2021年网络会议论文集（WWW '21）。3850–3860。[13] Leo A. Goodman。1965 年。《多项比例的同时置信区间》。Technometrics 7, 2（1965 年 5 月），247–254。[14] Jason L. Harman、John O'Donovan、Tarek Abdelzaher 和 Cleotilde Gonzalez。2014 年。《人类对推荐系统的信任动态》。载于第 8 届 ACM推荐系统会议论文集（RecSys '14）。305–308。[15] F. Maxwell Harper 和 Joseph A. Konstan。2015 年。《MovieLens数据集：历史与背景》。ACM Trans. Interact. Intell. Syst. 5, 4，文章 19（2015 年 12 月），19 页。[16] Xiangnan He、Tao Chen、Min-Yen Kan 和 Xiao Chen。2015 年。《TriRank：通过建模方面实现评论感知的可解释推荐》。载于第 24 届 ACM国际信息与知识管理会议论文集（CIKM '15）。1661–1670。[17] Jonathan L. Herlocker、Joseph A. Konstan 和 John Riedl。2000 年。《解释协同过滤推荐》。载于2000年 ACM计算机支持的协同工作会议论文集（CSCW '00）。241–250。[18] María Hernández-Rubio、Iván Cantador 和 Alejandro Bellogín。2019 年。《基于从用户评论中提取的物品方面观点的推荐系统比较分析》。用户建模与用户适应交互 29（2019 年 4 月），381–441。[19] Yifan Hu、Yehuda Koren 和 Chris Volinsky。2008 年。《面向隐式反馈数据集的协同过滤》。载于2008年第八届 IEEE数据挖掘国际会议论文集（ICDM '08）。263–272。[20] Johannes Kunkel、Tim Donkers、Lisa Michael、Catalin-Mihai Barbu 和 Jürgen Ziegler。2019 年。《让我来解释：个性化与非个性化解释对推荐系统信任的影响》。载于2019年 CHI计算机系统中的人为因素会议论文集（CHI '19）。1–12。[21] Khalil Ibrahim Muhammad、Aonghus Lawlor 和 Barry Smyth。2016 年。《关于推荐系统观点型解释的真实用户研究》。载于第 21 届智能用户界面国际会议论文集（IUI '16）。256–260。[22] Cataldo Musto、Pasquale Lops、Marco de Gemmis 和 Giovanni Semeraro。2019 年。《通过对用户评论进行基于方面的情感分析来论证推荐》。载于第 27 届 ACM用户建模、适应与个性化会议论文集（UMAP '19）。4–12。[23] Jianmo Ni、Jiacheng Li 和 Julian McAuley。2019 年。《使用远程标注评论和细粒度方面来论证推荐》。载于2019年自然语言处理实证方法会议暨第 9 届自然语言处理国际联合会议论文集。188–197。

[24] Xia Ning 和 George Karypis。2011。SLIM：面向 Top-N 推荐系统的稀疏线性方法。载于*2011 年 IEEE 第 11 届国际数据挖掘会议论文集（ICDM '11）*。497–506。[25] Ingrid Nunes 和 Dietmar Jannach。2017。决策支持与推荐系统中解释的系统性综述与分类体系。用户建模与用户适应交互27，3-5（2017），393–444。[26] Georgina Peake 和 Jun Wang。2018。解释挖掘：推荐系统潜因子模型的事后可解释性。载于第 *24 届 ACM SIGKDD* 国际知识发现与数据挖掘会议论文集（*KDD '18*）。2060–2069。[27] Gustavo Penha、Eyal Krikon 和 Vanessa Murdock。2022。用于语音产品搜索的基于成对评论的解释。载于*ACM SIGIR* 人类信息交互与检索会议（*CHIIR '22*）。300–304。[28] Alessandro Piscopo、Oana Inel、Sanne Vrijenhoek、Martijn Millecamp 和 Krisztian Balog。2022。SIGIR 2022 上首届推荐系统解释质量测量研讨会（QUARE 2022）报告。*SIGIR* 论坛56，2（2022 年 12 月）。[29] Badrul Sarwar、George Karypis、Joseph Konstan 和 John Riedl。2001。基于项目的协同过滤推荐算法。载于第 *10* 届万维网国际会议论文集（*WWW '01*）。285–295。[30] Nava Tintarev 和 Judith Masthoff。2012。评估推荐系统解释的有效性。用户建模与用户适应交互22（2012 年 10 月），399–439。[31] Nava Tintarev 和 Judith Masthoff。2015。解释推荐：设计与评估。载于推荐系统手册（第 2 版），Francesco Ricci、Lior Rokach、Bracha Shapira 和 Paul B. Kantor（编）。Springer 美国，第 10 章，353 382–。[32] Jesse Vig、Shilad Sen 和 John Riedl。2009。标签解释：使用标签解释推荐。载于第 *14* 届智能用户界面国际会议论文集（*IUI '09*）。47–56。[33] Yury Zemlyanskiy、Sudeep Gandhe、Ruining He、Bhargav Kanagal、Anirudh Ravula、Juraj Gottweis、Fei Sha 和 Ilya Eckstein。2021。DOCENT：从大型文档集合中学习自监督实体表示。载于欧洲计算语言学协会分会第 *16* 届会议论文集：主会场（*EACL '21*）。2540–2549。[34] Yongfeng Zhang 和 Xu Chen。2020。可解释推荐：综述与新视角。信息检索基础与趋势14，1（2020），1–101。

A 附录：研究设计

我们提供了有关研究设计和数据准备的更多细节。

A.1 方面提取

通过众包收集的数据已经过自动清理和预筛选；这包括修正大小写、删除末尾空白、筛除过长的方面，以及删除作为其他方面子串的方面。不过，数据还需要论文作者进一步手动清理和筛选，因为提取出的方面并非都符合模板（也就是说，人工参与者没有足够严格地遵循说明），有些方面可能过于严厉或冒犯，或者听起来太过个人化，不适合作为解释。在手动清理过程中，一些方面经过了轻微改写，并删除了近似重复项。最初收集的 9,635 组电影—方面配对中，自动预筛选后保留了 7,414 组，手动筛选结束后则保留了 5,948 组。

值得强调的是，推荐是个性化的，而附带的解释并非如此；也就是说，在相同实验条件下收到相同推荐的所有参与者，都会看到该物品对应的相同解释，以确保不存在未受控制的偏差。

A.2 解释

当正面和负面方面同时显示时（条件 #3–#8），我们通过轮换表5中所示的序列，使每位参与者的展示顺序以及所有参与者整体的展示顺序都完全平衡。这些序列遵循拉丁方设计，其中二进制值 P/N 由最低有效位决定。

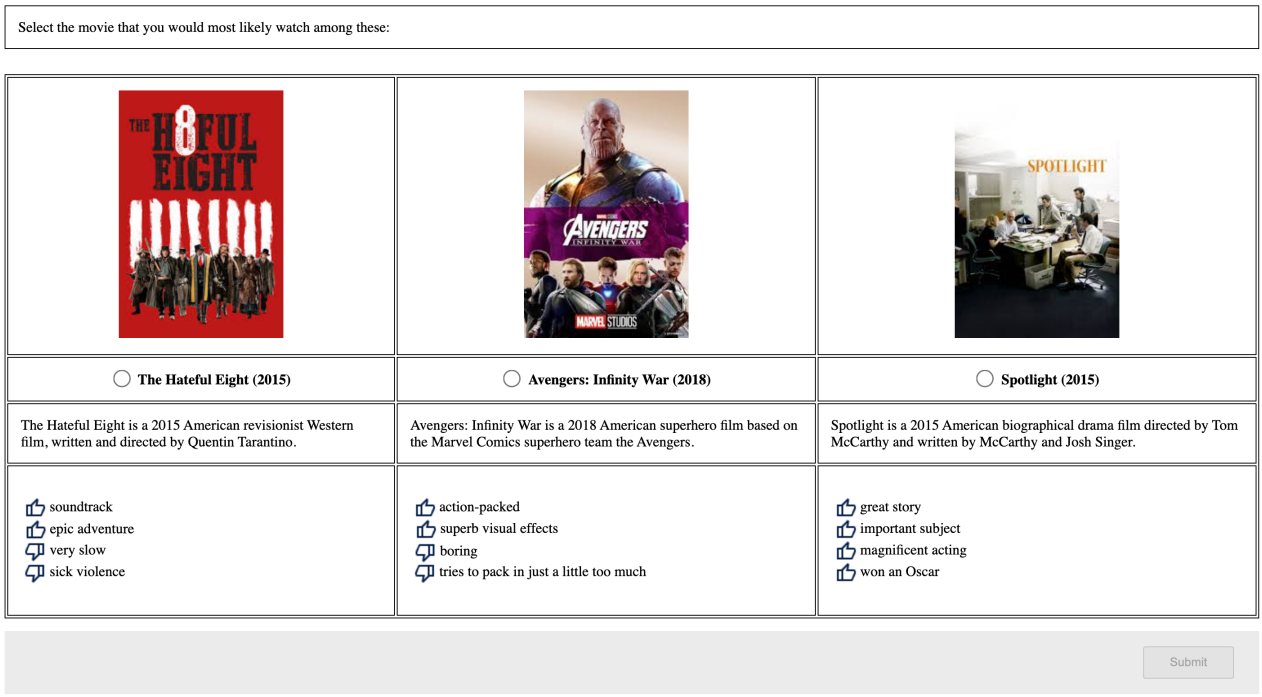

图 6：物品选择界面（阶段 2）。

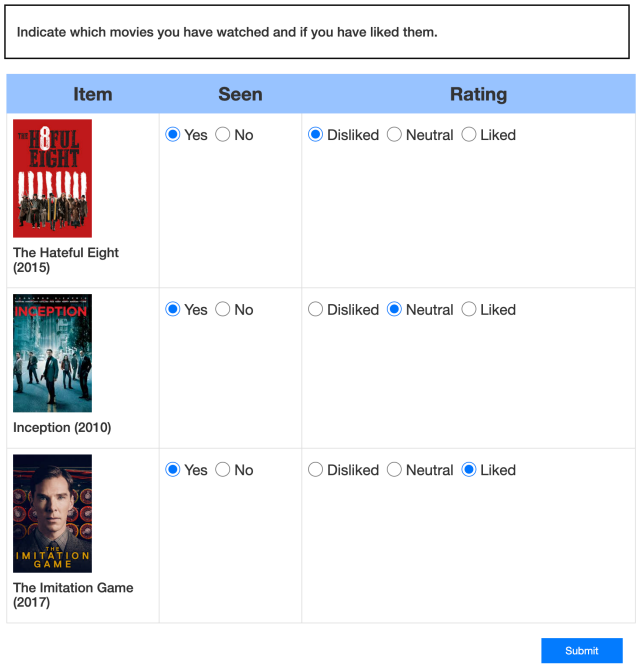

图 7：物品消费与偏好 elicitation 界面（阶段1）。

表5：决定解释中先列出正面还是负面方面的序列。

序列	条件					
	#3	#4	#5	#6	#7	#8
1	N	P	P	N	N	P
2	P	N	N	P	P	N
3	N	P	P	N	N	P
4	P	N	N	P	P	N
5	N	P	P	N	N	P
6	P	N	N	P	P	N

A.3 用户界面

图7和图6分别展示了阶段1和阶段2所使用的用户界面截图。