arXiv:2305.02679v2 [cs.CL] 6 May 2023

# 相邻词如何影响人们对显著性解释的理解

**AlonJacovi**[1]* **Hendrik Schuff**[2], 3***Heike Adel**[2] **NgocThangVu**[3]
**YoavGoldberg**[1,41]巴伊兰大学[2]博世人工智能中心[3]斯图加特大学[4]艾伦人工智能研究所alonjacovi@gmail.com
{hendrik.schuff,heike.adel}@de.bosch.com
thang.vu@ims.uni-stuttgart.de yoav.goldberg@gmail.com

## 摘要

词级显著性解释（“词语上的热力图”）常用于传达基于文本的模型中的特征归因。近期研究发现，词长等表层因素会扭曲人们对所传达的显著性评分的理解。我们开展了一项用户研究，探究标记一个词的相邻词会如何影响解释接受者在显著性解释情境中对该词重要性的感知。我们发现，相邻词会显著影响该词的重要性评分。具体而言，我们发现这种影响会因相邻方向（左侧与右侧）以及短语和搭配的先验语言学和计算测量（相对于无关相邻词）而异。我们的结果引发了这样的疑问：是否应继续以词为单位传达基于文本的显著性解释，并为未来关于替代性显著性解释方法的研究提供参考。




图 1：用户研究示意图。我们请普通人对词重要性解释（灰色）之后词语的感知重要性进行评分。随后，我们根据不同测量下词语之间的关系，分析相邻词的重要性对这种解读的影响（橙色）。

## 1 引言

在为单个词分配重要性评分的可解释性方法背景下，我们关注短语层面特征对某个词的感知重要性的影响：文本通常以各种粒度层次自然构建并被理解，而这些层次超越了词语层面（Chomsky，1957；Xia，2018）。例如（图 1），词语“*York*”的作用受到包含它的短语“*New York*”的语境化影响。假设某个解释分别将重要性归因于“*New*”和“*York*”，那么“*New*”的重要性评分会如何影响解释接受者对“*York*”重要性的理解？我们的研究探讨了这一问题。

<escape>当前自然语言处理中的特征归因解释主要在词级或子词级别进行（Madsen 等人，2023；Arras 等人，2017；Ribeiro</escape>等人，2016；Carvalho 等人，2019）。以往研究探讨了词语和句子层面的特征对人们主观解读文本显著性解释的影响（Schuff 等人，2022）——研究发现，词长和频率等特征会影响用户对解释的感知（例如，用户可能会认为较长的词更重要）。

人工智能系统的解释要成功地将预期信息传达给解释接受者，并非易事（Miller, 2019; Dinu 等人, 2020; Fel 等人, 2021; Arora 等人, 2021）。对于特征归因解释（Burkart 和 Huber, 2021; Tjoa 和 Guan, 2021）而言，这类解释在自然语言处理中通常以基于词语重要性的解释形式出现（Madsen 等人, 2023; Danilevsky 等人, 2020），我们必须了解解释接受者如何解读归因输入对模型输出所起的作用（Nguyen 等人, 2021; Zhou 等人, 2022）。研究表明，假设解释接受者会“按照……来”解读解释，往往是错误的，

*两位作者对本研究的贡献相同。

倾向”（Gonzalez 等人，2021；Ehsan 等人，2021）。

这项研究分为两个阶段（图 1）。首先，我们在一种场景中收集普通人对重要性的主观自我报告评分：他们看到的是某事实核查自然语言处理模型以颜色编码呈现的词语重要性解释（第 2 节，图 2）。随后，我们拟合一个统计模型，根据二元组构式的各种先验测量（例如词语之间的句法关系，或它们在语料库中的搭配程度），将相邻词的重要性与词语评分关联起来（Kolesnikova，2016）。

我们观察到以下方面存在显著影响（第4节）：1. 左邻接与右邻接；2. 两个词语重要性之间的差异；3. 这些词语之间的短语关系（常见短语或无关联）。随后，我们根据相关文献推断这些影响可能的成因（第5节）。我们还在另一个英语语言领域中复现了Schuff 等人（2022）（第3节）。我们公开了收集的数据和分析代码。[1]

我们得出结论，普通人对英语词语重要性解释的理解**可能会受到相邻词重要性的偏差影响**，而阅读方向和语言中的短语单位可能会调节这种影响。未来关于特征归因的研究应探讨更有效的信息传达方法（Mosca 等人，2022；Ju 等人，2022），此类解释的实现也应注意，不要假定人类用户会客观地理解词级重要性。

## 2 研究说明

我们的分析分为两个阶段：收集普通人对词语重要性的主观解读，并检验各种属性是否会对收集到的评分产生显著影响，尤其是相邻词语的属性对评分词的影响。

### 2.1 收集感知重要性

我们请普通人评定特征重要性解释中某个词的重要性（图 2）。本研究的设置基于 Schuff 等人(2022)，主要区别在于文本领域。我们通过 Amazon Mechanical Turk 众包平台招募了共计 100 名参与者。[2]

[1]https://github.com/boschresearch/human-interpretation-saliency。[2]我们从英语国家中选取英语母语评审者，并分析第一项实验中64名参与者和第二项实验中36名参与者的回答。详细信息见附录 A。

| 测量 | 示例 | 描述 |
|---|---|---|
| 一阶成分 | 高度发达，多于，例如 | 最小多词成分子树成分树中。 |
| *Noun* 短语 | 东京马拉松、滑雪选手、英国 | 多词名词短语在成分树中。 |
| 频率 | 联合体、族、一种物种 | 原始的、未归一化的频率-频率。 |
| 泊松斯特林 | 一位美国人，例如，一个物种 | 泊松—斯特林二元词组评分。 |
| $\varphi^2$ | Massar Egbari，冰块曲棍球，Udo Dirkschneider | 皮尔逊相关系数的平方相关系数。 |

表 1：短语度量的示例子集。

**解释**。 我们采用文献中更常见的词语重要性解释可视化格式，即颜色编码（例如， Ar-ras 等人，2017； Wang 等人，2020； Tenney 等人，2020 ； Arora 等人， 2021 ）。我们使用来自两个来源的重要性值：随机化和 SHAP 值[3] (Lundberg 和 Lee， 2017)，针对 `facebook/bart-large-mnli`[4] (Yin 等人，2019；Lewis 等人，2020)，作为事实核查模型。

**任务**。 我们向参与者说明，模型正在执行一项合理的任务：判断给定句子是否为事实（Lazarski 等人，2021）。源文本取自 150 个维基百科句子的样本，[5] 目的是选择一个多词词块出现率较高的领域中的文本。

**流程**。 我们询问解释接受者：“你认为单词 [...] 对模型有多重要（1-7）？”并收集点量表答案，同时提供可选的评论栏。对 150 个句子中的每个句子，我们都会随机抽取一个单词，重复这一流程。

### 2.2 测量邻词效应

理想情况下，一个词的重要性评分应完全由其显著性强度解释。然而，先前的研究表明事实并非如此。在这里，我们关注参与者的回答是否以及在多大程度上可以由相邻词的属性解释，超出仅凭评分词自身的显著性所能解释的范围。

[3]由于我们的数据中观察到的最大 SHAP 值为 0.405，因此我们对所有 SHAP 值进行归一化，使其 $0.405^{-1}$ 覆盖完整的颜色范围。[4] https://huggingface.co/facebook/bart-large-mnli[5]有关 维基百科句子 集合的信息，请参见kaggle.com/datasets/mikeortman/wikipedia-sentences。

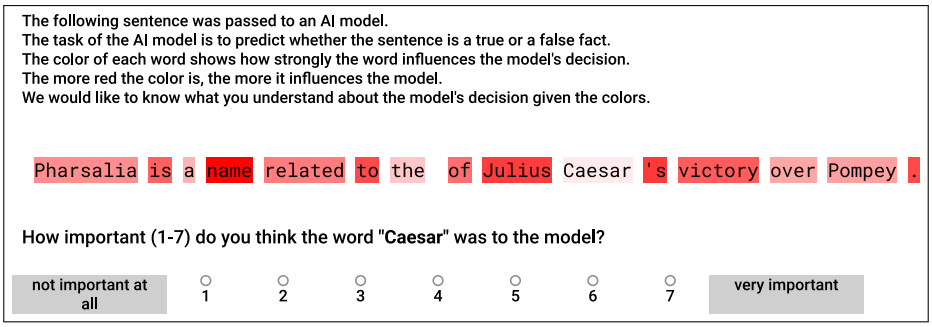



图 2：评分界面的屏幕截图。

仅由评分词的显著性解释。

**建模。** 我们使用有序广义加性混合模型（广义加性混合模型）分析收集到的评分。[6] 其关键特性是，它将有序响应变量（即我们设置中的重要性评分）建模为连续潜在尺度上协变量的平滑函数之和，同时还考虑随机效应。[7]

**先前模型术语。**我们纳入了Schuff 等人(2022)测试的所有协变量，包括评分词的显著性、词长等，以便在检验我们新增的短语级变量时对这些因素进行控制。我们沿用Schuff 等人对所有先前主效应和随机效应的控制方法。[8]

**新邻居项术语。** 以下变量决定了我们新增的模型术语，并构成分析的基础：左侧或右侧邻接；评分词的显著性（颜色强度）；两个词之间的显著性差异；以及两个词之间的关系是弱关系还是强关系。我们根据上述变量的交互作用，引入了四个新的双变量平滑项（图 3）。

我们将关系紧密的二元组称为一个块。为了对块进行可靠的测量，我们系统地测试了两类不同的二元组关系测量方法（表 1）：句法方法，通过依存句法分析；以及统计方法，通过语料库中的词语搭配。遵循Frantzi 等人(2000）的做法，我们结合使用句法和统计测量，将处于搭配第 0.875 百分位以上的一阶成分视为块（我们的观察结果对统计测量方法的 $\varphi^2$ 选择具有稳健性和百分位数均具有稳健性；参见附录C）。

## 3 复现先前结果

我们的研究与 Schuff 等人 (2022考察词级和句子级特征对重要性感知影响的实验相似。因此，本研究适合尝试复现先前的观察结果，以确认这些结果在不同语言领域中是否依然成立：与 Schuff 等人研究中的中等长度维基百科文本与短篇餐馆评论相比，以及 SHAP值与集成梯度（Sundararajan 等人，2017）。

结果是肯定的：我们复现了先前报告的显著影响，包括词长、显示索引 （即被评分实例在 150 个句子中的位置）、大写以及依存关系 对随机化解释和 SHAP值解释的影响（详情见附录 A）。这一结果进一步印证了先前的观察：尽管界面提供了客观的可视化，用户仍极易对显著性解释产生偏差。

## 4 邻居效应分析

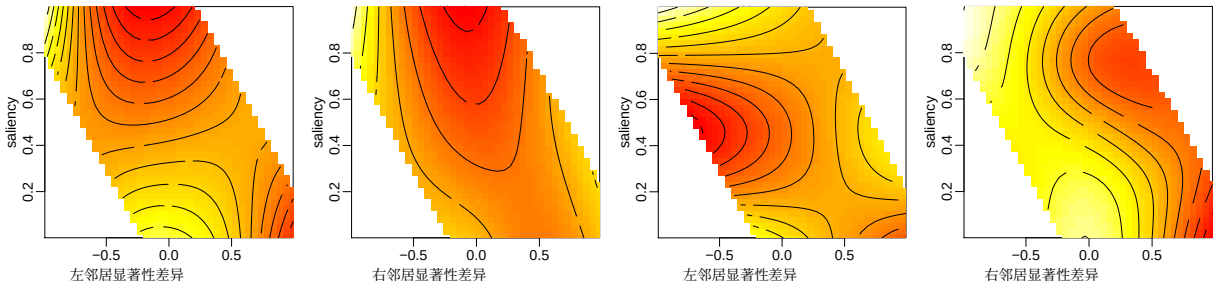
下面介绍我们使用（a）随机显著性值和（b）SHAP 值进行的两项实验的结果。

### 4.1 随机化解释

关于我们额外引入的邻居项，图 3 展示了四种所述函数（左/右 ×分块/无分块）的估计值。表2 列出了所有平滑项和参数项，以及Wald检验的结果（Wood，2013a，b）。附录 A 包含更多结果。

**非对称影响。**图3a 与图3b以及图3c与图3d揭示了左邻居和右邻居影响之间的定性差异。

[6]附录中的介绍性描述 B。 [7]随机效应可以控制例如个体参与者评分行为中的系统性差异，例如某位参与者倾向于给出总体上高于其他参与者的评分。[8]我们在建模时排除了两两交互项，因为这样可以提高稳定性，同时不损失表达能力。




(a) 左，无分块。（∗） （b）右侧，无分块。（∗） (c) 左侧，分块。（∗） (d) 右侧，分块。

图3：L 左侧和右侧的邻近词。(∗) 表示统计显著的平滑项。颜色按每个图形进行归一化。 e.

| Term | (e)自由度 | 参考自由度 | F | $p$ |
|---|---|---|---|---|
| s(显著性) | 11.22 | 19.00 | 580.89 | <**0.0001** |
| s(显示索引) | 3.04 | 9.00 | 22.02 | <**0.0001** |
| s(词长) | 1.64 | 9.00 | 16.44 | <**0.0001** |
| s(句长) | 0.00 | 4.00 | 0.00 | 0.425 |
| s(相对词频) | 0.00 | 9.00 | 0.00 | 0.844 |
| s(归一化显著性排名) | 0.59 | 9.00 | 0.37 | 0.115 |
| s(词位置) | 0.58 | 9.00 | 0.18 | 0.177 |
| te(左侧差异、显著性)：无分块 | 3.12 | 24.00 | 1.50 | **0.002** |
| te(左侧差异、显著性)：分块 | 2.24 | 24.00 | 0.51 | **0.038** |
| te(右侧差异、显著性)：无分块 | 2.43 | 24.00 | 0.47 | **0.049** |
| te(右侧差异、显著性)：分块 | 0.00 | 24.00 | 0.00 | 0.578 |
| 大写 | 2.00 | | 3.15 | **0.042** |
| 依存关系 | 35.00 | | 2.92 | <**0.0001** |

表2：随机显著性实验中单变量平滑项（上）和参数项（下）的（有效）自由度、参考自由度及Wald检验统计量。

我们通过计算显著差异区域来定量确认这些差异（Fasiolo 等人，2020；Marra 和 Wood，2012）。图4a和4b 展示了组块情形下（显著）差异和概率的相应图示。总的来说，我们得出结论：左侧和右侧词邻居的影响存在显著差异。

**组块影响**。 我们研究了与评分词位于同一分块内的邻词与不在同一分块内的邻词之间的差异。我们在图 3 中发现了定性差异，同时也发现了具有统计显著性的差异（图 4c和4d）。

**显著性调节邻词差异**。 图 3 显示，邻词显著性差异的影响（x轴）受到评分词显著性（y轴）的调节。我们通过比较评分词显著性为0.25和0.75时的函数，并使用单维差异图（Van Rij 等人，2015），从统计角度验证了这一观察结果（图4e）。

**综合效应**。 我们发现了两种普遍的对立效应：同化和对比。[9]

[9]我们借用了心理学中的这些术语（第5节）。

我们将同化 定义为这样一种情况：根据邻词的显著性高低，词语被感知到的显著性会被认为更重要（或不那么重要）。我们发现，当左侧 邻词与显著性中等（0.25–0.75）的评分词构成一个分块时，会产生同化效应。

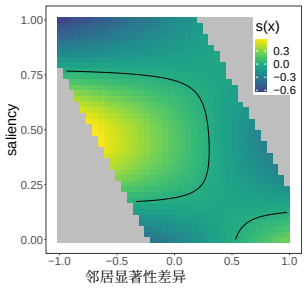
我们将对比 定义为：根据某个词的邻词显著性更高（或更低），人们对该词显著性的感知会降低（或提高）。我们发现，来自左侧和右侧邻居的对比效应并不会与评分词构成一个分块。[10]

### 4.2 SHAP值解释

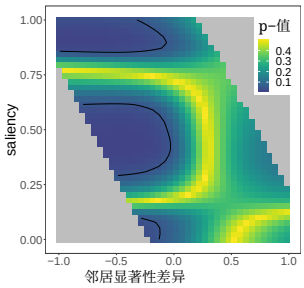
**共享结果**。 我们的 SHAP值实验证实了我们的观察结果：(i) 左侧/右侧邻词的影响不对称（图11a 和11b）、(ii) 组块影响（图11c和11d）、(iii) 显著性的调节效应（图11e），以及(iv) 同化和对比效应（图10d）。

**变体结果**。 值得注意的是，我们的 SHAP 值结果与随机化显著性结果在左右方向的影响方面有所不同。在随机显著性实验中，我们观察到分块内左侧邻居带来的同化效应（图 3c），以及分块外左侧和右侧邻居带来的对比效应（图 3a和 3b）。在 SHAP 值实验中，我们观察到分块内右邻居带来的同化效应（评分词显著性较低）和对比效应（评分词归一化显著性中等）（图 10d）。我们推测，这种差异可归因于 SHAP 值之间的相互依赖关系，如附录 12 B中图 所示。

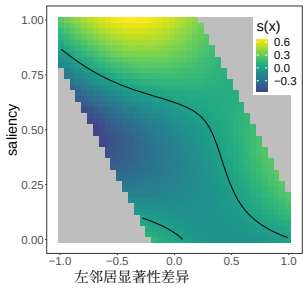
[10]请注意，尽管图3d显示出对比效应，但颜色归一化放大了该图中的细微差异，而 Wald检验并未 检出显著效应。



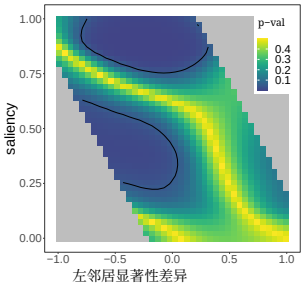

(a) 右/左差异为块（轮廓标记 0）。



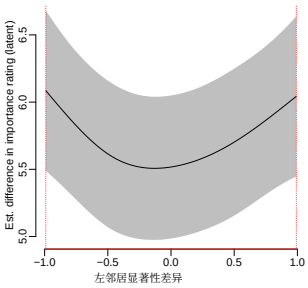

(b) 右/左差异（分块）$p$ 值（轮廓标记 0.05）。




(c) 有块/无块差异对于左邻居（轮廓标记 0).




（d）有块/无块差异（左） $p$值（轮廓标记 0.05)。




(e) 额定值之间的差异显著性 and left 邻近词语显著性。

图4：差异图。等高线指的是等高线。(e) 中的红色 x 轴标示显著差异。

### 4.3 主要结论

总体而言，我们发现 (a) 左右方向的影响并不相同，(b) 强烈的二元组关系可能会使左侧邻居引发的对比效应转变为同化效应，(c) 极端显著性可能会抑制同化效应，(d) 随机化解释和 SHAP 值解释中均可观察到偏置效应。

## 5 心理学理论基础

当然，同化效应是直观的——它仅仅意味着，当二元组关系较强时，邻词的重要性会从邻词“泄漏”到评分词上。但文献中是否有对所观察到的同化效应和对比效应的先例？它们彼此之间有何联系？

心理学研究启动项（例如接触特定词语）如何影响人类判断，这属于以下两类效应之一：同化（评分会被“拉向”启动项）和对比（评分会被“推离”启动项）效应（尤其参见Bless 和 Burger，2016）。

Förster 等人（2008）展示了整体加工（例如，关注整体结构）与局部加工（例如，关注结构细节）如何分别导致同化与对比。我们认为，我们的一些观察结果可以用他们的模型来解释：多词短语的邻词可能引发整体加工，从而导致同化（例如，在随机化解释实验中，左侧邻居）；而其他邻词（在随机化解释实验中，右侧邻居和无关的左侧邻居）则会引发局部加工，从而导致对比。未来的研究可以考察在特定语境中引发整体加工的属性。

## 6 结论

我们在英语维基百科领域开展了一项用户研究，研究场景是普通人观察以颜色编码重要性呈现的常见词重要性解释。在这一场景中，我们发现，当解释对象理解某个词被赋予的重要性时，其他词语的重要性可能会以意想不到的方式影响他们的理解。

通常认为，在传达特征归因解释中某个组成部分的重要性时，解释接受者会如实理解所呈现的重要性。但我们发现事实并非如此：解释接受者对输入部分的情境化理解——对我们而言，就是作为短语一部分的单词——可能会影响他们对解释的理解。

## 局限性

原则上，本研究观察到的效应只能应用于我们的用户研究所处的情境（英语文本、讲英语的众包工作者、以颜色编码的词级显著性等，详见论文）。因此，对于自然语言处理研究中一种相当合理且常见的情境，本研究仅作为一项存在性证明，表明普通人理解特征归因解释时，可能会受到输入中归因部分以外的上下文影响。在其他情境下为自然语言处理系统设计和实现解释技术，或为性质相似的其他系统采取行动时，应当调查这些效应能否推广到实际应用情境（我们计划为此发布完整的复现代码），或者采取审慎措施，预先考虑这些效应可能会推广，同时也不排除它们不会推广的可能性。

## 致谢

我们感谢 Diego Frassinelli 和匿名评审者提供的宝贵反馈与有益意见。A. Jacovi 和 Y. Goldberg 获得了欧盟“地平线 2020”研究与创新计划下欧洲研究理事会（ERC）的资助，资助协议编号为 802774（iEXTRACT）。N.T. Vu 获得了卡尔蔡司基金会的资助。

## 参考文献

Siddhant Arora、Danish Pruthi、Norman M. Sadeh、William W. Cohen、Zachary C. Lipton 和 Graham Neubig。2021。解释、编辑与理解：重新思考用于评估模型解释的用户研究设计。*CoRR*，abs/2112.09669。

Leila Arras、Franziska Horn、Grégoire Montavon、Klaus-Robert Müller 和 Wojciech Samek。2017。“文本文档中哪些内容是相关的？”：一种可解释的机器学习方法。*PloS one*，12(8):e0181142。出版方：公共科学图书馆，旧金山，加利福尼亚州，美国。

Herbert Bless 和 Axel M Burger。2016年。社会启动中的同化与对比。《心理学当前观点》，12:26–31。社会启动。

Nadia Burkart 和 Marco F. Huber。2021年。监督式机器学习可解释性综述。《人工智能研究杂志》，70:245–317。

Diogo V. Carvalho、Eduardo M. Pereira 和 Jaime S. Cardoso。2019年。机器学习可解释性：方法与度量综述。《电子学》，8(8):832。

NoamChomsky.1957. 句法结构. De Gruyter Mouton，柏林，波士顿。Marina Danilevsky、Kun Qian、Ranit Aharonov、Yannis Katsis、Ban Kawas 和 Prithviraj Sen。2020。自然语言处理领域可解释人工智能现状综述。载于《亚太计算语言学协会第*1*届会议暨第*10*届国际自然语言处理联合会议论文集》，*AACL/IJCNLP 2020*，中国苏州，*2020*年*12*月*4*日至*7*日，第447–459页。计算语言学协会。Jonathan Dinu、Jeffrey P. Bigham 和 J. Zico Kolter。2020。对特征归因解释中常见可解释性假设的挑战。*CoRR*，abs/2012.02748。ArXiv：2012.02748。Dagmar Divjak 和 Harald Baayen。2017。序数广义加性混合模型：理解人类评分的新视角。载于每一次冒险，都是新的开始：献给 *Laura A. Janda* 的研究，第39–56页。Slavica 出版社。Upol Ehsan、Samir Passi、Q. Vera Liao、Larry Chan、I-Hsiang Lee、Michael J. Muller 和 Mark O. Riedl。2021。可解释人工智能中的“人”：人工智能背景如何塑造人们对人工智能解释的看法。*CoRR*，abs/2107.13509。Matteo Fasiolo、Raphaël Nedellec、Yannig Goude 和 Simon N Wood。2020。现代广义加性模型的可扩展可视化方法。计算与图形统计学杂志，29(1)：78–86。Thomas Fel、Rémi Cadène、Mathieu Chalvidal、Matthieu Cord、David Vigouroux 和 Thomas Serre。2021。关注方差！基于 Sobol 敏感性分析的高效黑盒解释方法。载于《神经信息处理系统进展第*34*卷：*2021*年神经信息处理系统年会》，*NeurIPS 2021*，*2021*年*12*月*6*日至*14*日，线上，第26005–26014页。Jens Förster、Nira Liberman 和 Stefanie Kuschel。2008。全局与局部加工风格对社会判断中同化与对比的影响。人格与社会心理学杂志，94(4)：579。Katerina T. Frantzi、Sophia Ananiadou 和 Hideki Mima。2000。多词术语的自动识别：C-value/NC-value 方法。国际数字图书馆杂志，3：115–130。Ana Valeria Gonzalez、Anna Rogers 和 Anders Søgaard。2021。信念偏差与解释之间的相互作用。载于计算语言学协会论文集：*ACL/IJCNLP 2021*，线上会议，*2021*年*8*月*1*日至*6*日，载于第 ACL/IJCNLP 2021 卷的 *ACL*论文集，第2930–2942页。计算语言学协会。

Yiming Ju、Yuanzhe Zhang、Kang Liu 和 J un Zhao。2022。 无需连接规则的文本分类层次化解释生成。 *CoRR*，abs/2210.13270。

Olga Kolesnikova. 2016. 用于搭配检测的词语共现度量综述。*ComputacionySistemas*，20:327–344。

Eric Lazarski、Mahmood Al-Khassaweneh 和 Cynthia Howard。2021。 使用自然语言处理进行事实核查：一项综述。 *Designs*，5(3)。

Mike Lewis、Yinhan Liu、Naman Goyal、Mar- jan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020。 BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。收录于 计算语言学协会第*58*届年会论文集，第7871–7880页，线上。计算语言学协会。

Scott M Lundberg 和 Su-In Lee。2017。一种解释模型预测的统一方法。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展第*30*卷，第4765–4774页。Curran Associates, Inc. Andreas Madsen、Siva Reddy 和 Sarath Chandar。2023。神经自然语言处理的事后可解释性：综述。*ACM* 计算调查，55(8)：155:1–155:42。Giampiero Marra 和 Simon N. Wood。2012。广义加性模型分量置信区间的覆盖性质。斯堪的纳维亚统计学杂志，39(1)：53–74。Tim Miller。2019。人工智能解释：社会科学视角。人工智能，267：1–38。Edoardo Mosca、Defne Demirtürk、Luca Mülln、Fabio Raffagnato 和 Georg Groh。2022。Grammar-SHAP：一种高效、模型无关且具备结构感知能力的自然语言处理解释器。载于第一届自然语言监督学习研讨会论文集，第10–16页，爱尔兰都柏林。计算语言学协会。Giang Nguyen、Daeyoung Kim 和 Anh Nguyen。2021。特征归因方法的有效性及其与自动评估分数的相关性。载于神经信息处理系统进展第*34*卷：*2021*年神经信息处理系统年会，*NeurIPS 2021*，*2021*年*12*月*6*日至*14*日，线上，第26422–26436页。Peng Qi、Yuhao Zhang、Yuhui Zhang、Jason Bolton 和 Christopher D. Manning。2020。Stanza：一种适用于多种人类语言的 Python 自然语言处理工具包。载于第*58*届计算语言学协会年会论文集：系统演示。

Marco Túlio Ribeiro、Sameer Singh 和 Carlos Guestrin。2016。“我为什么要相信你？”：解释任意分类器的预测。载于第*22*届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，美国加利福尼亚州旧金山，*2016*年*8*月*13*日至*17*日，第1135–1144页。ACM。

Hendrik Schuff、Alon Jacovi、Heike Adel、Yoav Gold- berg 和 Ngoc Thang Vu。2022。 人类对文本中基于显著性的解释的理解。收录于*FAccT '22*：*2022 ACM*公平、问责与透明度会议，大韩民国首尔，*2022*年*6*月*21*日至*24*日，第611–636页。ACM。

MukundSundararajan、AnkurTaly 和 Qiqi Yan。2017年。深度网络的公理化归因。载于第*34*届国际机器学习会议论文集，*ICML 2017*，悉尼，新南威尔士州，澳大利亚，*2017*年*8*月*6*日至*11*日，第70卷机器学习研究论文集，第3319–3328页。PMLR。

Ian Tenney、James Wexler、Jasmijn Bastings、Tolga Bolukbasi、Andy Coenen、Sebastian Gehrmann、Ellen Jiang、Mahima Pushkarna、Carey Radebaugh、Emily Reif 和 Ann Yuan。2020。语言可解释性工具：面向自然语言处理模型的可扩展交互式可视化与分析。Erico Tjoa 和 Cuntai Guan。2021。可解释人工智能（XAI）综述：迈向医疗领域的 XAI。*IEEE*神经网络与学习系统汇刊，32(11):4793—4813。Jacolien Van Rij、Martijn Wieling、R Harald Baayen 和 Dirk van Rijn。2015。itsadug：使用广义加性混合模型解读时间序列和自相关数据。Junlin Wang、Jens Tuyls、Eric Wallace 和 Sameer Singh。2020。基于梯度的自然语言处理模型分析具有可操纵性。收录于计算语言学协会论文集：*EMNLP 2020*，线上会议，*2020*年*11*月*16*日至*20*日，第 EMNLP 2020 卷，载于*ACL*论文集，第247–258页。计算语言学协会。Simon N Wood。2013a。关于扩展广义加性模型平滑项的p值。*Biometrika*，100(1):221–228。Simon N Wood。2013b。回归模型中随机效应的简单检验。*Biometrika*，100(4):1005–1010。Simon N Wood。2017。广义加性模型：*R*语言入门。CRC出版社。Xiufang Xia。2018。一种记忆新词的有效方法——词汇块。语言研究理论与实践，8:14941498。Wenpeng Yin、Jamaal Hay 和 Dan Roth。2019。零样本文本分类基准测试：数据集、评估与蕴含方法。

收录于*2019*年自然语言处理实证方法会议与第九届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP 2019*），中国香港，*2019*年*11*月*3*日至7日，第3912–3921页。计算语言学协会。

Yilun Zhou、Serena Booth、Marco Túlio Ribeiro 和 Julie Shah。2022。特征归因方法是否正确归因了特征？载于第三十六届*AAAI*人工智能会议、*AAAI 2022*、第三十四届人工智能创新应用会议、*IAAI 2022*、第十二届人工智能教育进展研讨会、*EAAI 2022*（线上），*2022*年*2*月*22*日至*3*月*1*日，第9623–9633页。AAAI出版社。

## A 用户研究详情

本节详细介绍我们的用户研究设置。

### A.1 界面

图5 展示了评分界面的截图。图6 展示了注意力检查的截图。

### A.2 注意力检查

我们为每位参与者设置三个注意力检查，并参照 Schuff 等人（2022），在研究最后三分之二的部分随机安排这些检查。

### A.3 参与者

我们通过 Amazon Mechanical Turk 从英语国家招募了 76 名众包工作者参与随机化解释研究，并招募了 36 名众包工作者参与 SHAP值解释研究。我们要求工作者至少完成 5,000 个人类智能任务，且通过率达到 95%。我们通过三项隐藏式注意力检查筛选评分者；只有正确回答所有检查的评分者才能纳入研究（但无论是否纳入，都会获得全额报酬）。在 76 名工作者中，有 64 名通过筛选，也就是说，我们在参与者层面排除了 15.8% 的回答。在 36 名工作者中，所有人都通过了筛选。参与者平均获得 8.95 美元的时薪。我们不会收集参与者的任何个人身份信息。

## B 统计模型详情

在本节中，我们简要介绍所使用的统计模型（即广义加性混合模型），并提供分析的其他结果。

| **示例** |
| --- |
| 新发病原体研究所是一个与佛罗里达大学相关的跨学科研究机构。Luca Emanuel Meisl（生于1999年3月4日）是一名奥地利足球运动员，目前效力于FC Liefering。黑喉簇舌巨嘴鸟（Aulacorhynchus atrogularis）是一种近雀形目鸟类，分布于厄瓜多尔中部至玻利维亚西部。Christopher Robert Coste（生于1973年2月4日）是一位作家，也曾是美国职业棒球大联盟的捕手。WGTA于2014年11月3日向美国联邦通信委员会（FCC）交还了其许可证。 |

表 3：研究中使用的维基百科句子示例。

### B.1 广义加性混合模型简介

关于广义加性混合模型，我们参照Schuff 等人(2022）（附录）中非常简短的介绍。简而言之，有序广义加性混合模型可描述为一种广义加性模型，它还考虑了随机效应，并通过连续潜变量对有序评分进行建模，再利用估计的阈值将该变量划分为有序类别。有关更多细节，Divjak 和 Baayen(2017) 在语言学语境下对有序广义加性模型作了实用介绍，而Wood (2017) 则提供了一本关于广义加性（混合）模型的详细教材，涵盖实现和分析细节。

### B.2 我们分析中的模型详情

我们控制了所有主效应（词长、句长等）以及 Schuff 等人(2022) 使用的所有随机效应。由于纳入交互项会导致模型不稳定，我们排除了两两交互项。

此外，我们还纳入了四个新的双变量平滑项。每个术语都对显著性（即评分词的颜色强度）与相邻（左侧或右侧）词语相对于评分词的显著性差值进行张量积建模。对于左右两侧，我们分别对以下邻词的平滑项进行建模：(i) 与评分词处于同一词汇块内的邻词；(ii) 不处于同一词汇块内的邻词。图3 展示了估计得到的四个（双变量）函数。

### B.3 数据预处理

继 Schuff 等人 （2022）的做法，我们排除了完成时间少于一分钟的评分（完成速度快得不合乎情理），以及长度超过 20 个字符的词语。我们实际上排除了 1.8% 的评分。

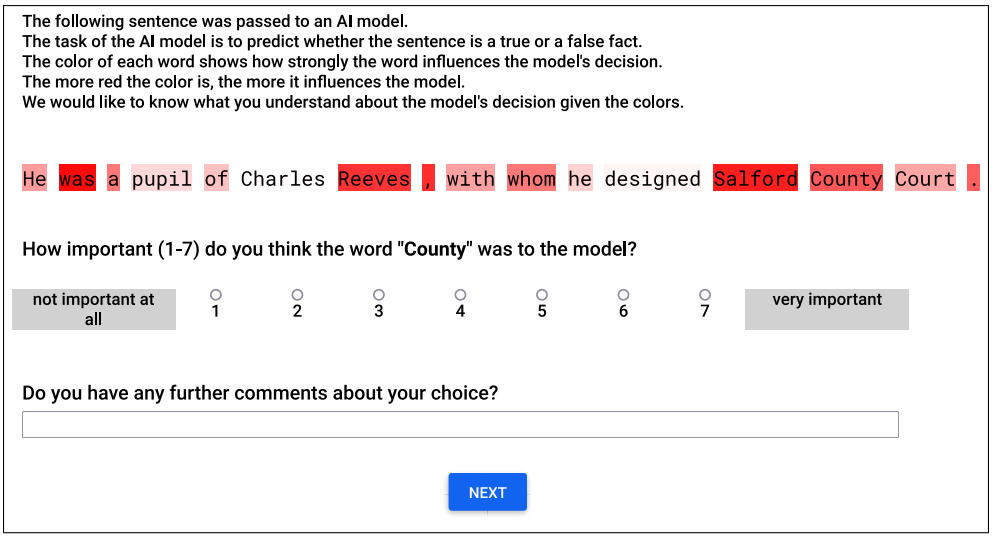



图5：评分界面的屏幕截图。

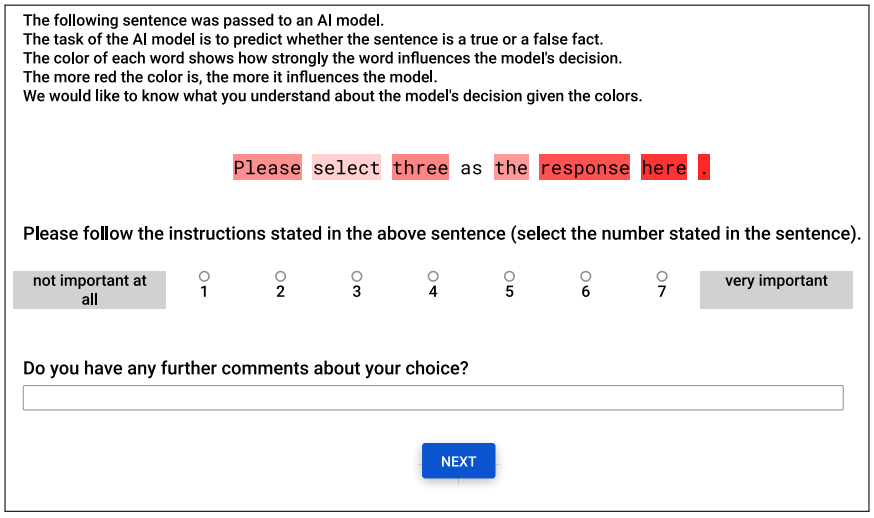



图6：注意力检查评分界面的屏幕截图。

为了分析左侧邻居和右侧邻居，我们还必须确保只纳入同时存在左、右邻居的评分。因此，我们进一步排除了对句子中最左侧或最右侧词语进行评分的记录。这排除了 11.7% 的评分。总计而言，我们因此使用 9489 条评分来拟合我们的模型。

### B.4 短语块测量

我们探索并结合了两种识别多词短语（或“分块”）的方法。

**句法测量（成分）**。我们首先基于句子的句法树应用二元分块测量。我们使用 Stanza（Qi 等人，2020）（版本 1.4.2）为每个句子生成句法树。我们评估评分词及其相邻词（左侧/右侧）是否在尽可能低的层级共享一个成分。具体而言，我们（a）从评分词开始，在句法树中向上移动一级；（b）从相邻词开始，在句法树中向上移动一级。如果此时我们到达句法树中的同一节点，则评分词与其相邻词共享一个一阶成分。如果到达的是不同节点，则二者不共享成分。将一级共享成分的类型限定为名词短语后，可得到一个进一步的类别。我们在表4（上半部分）中提供了从数据中提取的一级共享成分及相应名词短语成分的示例。

**统计测量（共现分数）**。此外，我们还探讨了数值型关联度量，并计算了 NLTK 的 *Bigra-* 中所有可用的二元搭配度量

*mAssocMeasures*模块[11]。计算基于维基百科句子（脚注5）中的 700 万条 2018 年维基百科句子。表4 （下半部分）提供了各项指标的描述，以及在我们的数据中评分最高的示例。我们将示例分为构成成分与非构成成分两类，以突出成分过滤器对于将内容有意义地分类为分块或无分块的必要性。

### B.5 详细结果

如第 4 节所述，在我们的随机化解释实验中，我们观察到左侧/右侧邻居、有分块/无分块邻居以及评分词显著性级别的不同影响。

**左侧邻居与右侧邻居**。 图 7 展示了有分块邻居（图 7a 和 7b）及无分块邻居（图7c和 7d）中左侧邻居与右侧邻居之间的差异图（及相应的 p值）。

**有分块与无分块**。 相应地，图 8展示了左侧邻居（图 8a 和 8b）以及右侧邻居（图8c 和 8d）中有分块邻居与无分块邻居之间的差异图（及相应的 p值）。

**不同显著性级别之间的差异**。 图 9显示，对于左侧邻居（图 9a）和右侧邻居（图 9b），显著性差异的影响在评分词的不同显著性级别（0.25 和 0.75）之间存在显著差异。

我们在表5中报告了随机化解释实验的详细Wald检验统计量。

### B.6 SHAP 值结果

我们还报告了 SHAP 值实验结果的详细信息。图 11 展示了左/右、分块/无分块以及评分词显著性水平的差异图。表 6报告了 SHAP 值解释实验的详细 Wald检验统计量。图 12 展示了随机化解释中的显著性评分分布如何均匀随机，与 SHAP值的分布形成对比。

### B.7 复现 Schuff 等人(2022)

我们证实了 Schuff 等人(2022)此前的研究结果，并发现**词长**、**显示索引**、**大写**和**依存关系**均有显著影响。表 5 列出了随机化显著性实验的详细统计数据，表6列出了 SHAP 实验的详细统计数据。

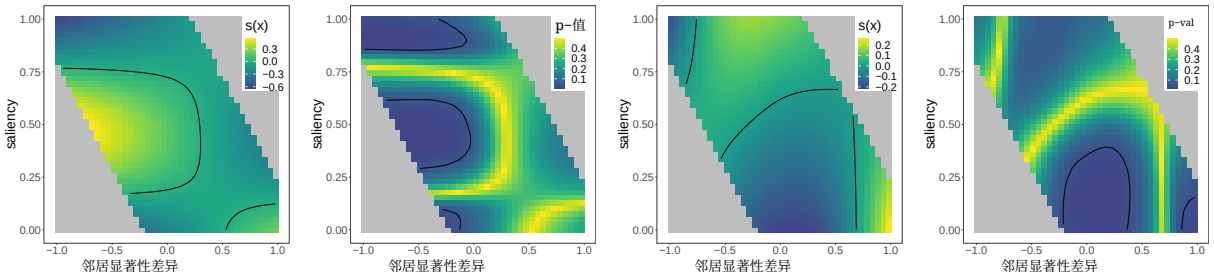
## C 对评估参数的稳健性。

为确保我们的结果并非由阈值与共现测量的特定组合造成的假象，我们研究了以下情况下结果的变化：(i) 在 $\{0.5, 0.75, 0.875\}$ 范围内改变阈值；(ii) 在 {Jac- card, 类似互信息，$\varphi^2$, 泊松-斯特林} 范围内改变共现测量。我们发现了显著的交互作用，并观察到所有设置下相似的交互模式，以及显著差异的区域（左/右、有分块/无分块以及不同显著性水平）。我们在图 13 至 18 中提供了有代表性的图表。此外，表 7 和 8 表明，改变阈值或共现测量后得到的模型统计数据与表 5中报告的结果基本一致。我们选择 $\varphi^2$ 和 87.5% 的阈值，因为没有其他模型的偏差解释率更高；并且在不同测量和阈值下对随机抽取的有分块/无分块示例进行比较时，该设置得到的结果最佳。

[11] https://www.nltk.org/_模块/nltk/metrics/ association.html

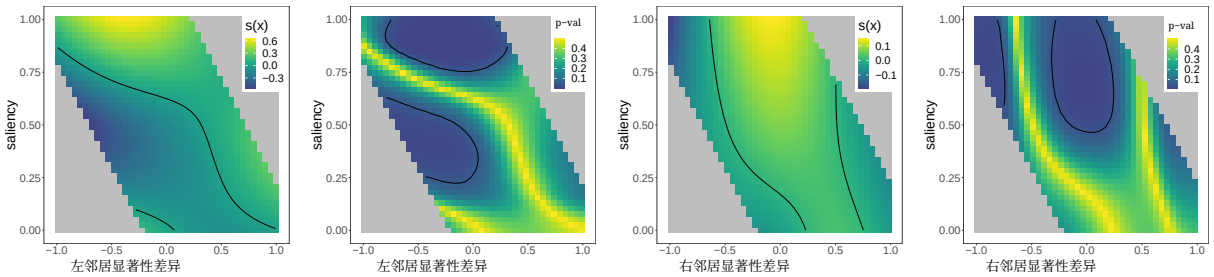
| 测量 | 成分示例 | 无成分示例 | 描述 |
|---|---|---|---|
| 一阶成分 | 高度发达，超过，<br>例如，DVD 组合、400 万 | — | 最小的多词成分子<br>成分树中的树。 |
| 名词短语 | 东京马拉松、滑雪选手、该<br>英国、一位退休的、该城市 | — | 多词一阶名词短语中的<br>成分树。 |
| 互信息 | 以及，多于，冰球，<br>联合王国，一种物种 | 是一个、的、在.....中、是一种、它曾经<br>是 | 双词互信息变体（根据<br>NLTK 实现）。 |
| 频率 | 联合、族、物种，<br>一位美国人，例如 | .....的，在.....中，是一个，朝.....，在.....<br>上 | 原始未归一化频率。 |
| 泊松 斯特林 | 一位美国人，例如，一个物种，<br>也，以及该族 | 是一个，.....的，在.....中，是一位，它曾经，<br>一直是 | 泊松-斯特林二元语法评分。 |
| 杰卡德 | Massar Egbari、冰球、空<br>军、多于、Udo<br>Dirkschneider | 教师/学生<br>教学/学习，是一个，已经是，<br>它曾经是，的 | 二元组杰卡德指数。 |
| $\varphi^2$ | Massar Egbari、冰球、Udo<br>Dirkschneider、空军、新<br>新西兰 | 教师/学生<br>教学/学习，是一个，已经是，<br>足球运动员 who，是一<br>个 | 皮尔逊相关系数的平方<br>。 |

表4：我们测试的短语度量列表。数值度量的示例根据最高共现分数选取，而（布尔型）名词短语和成分示例则是任意选取的。对于数值度量，我们提供的示例包括：(a) 与其邻词构成一个成分，以及 (b) 不构成成分。这些示例凸显了将数值评分与成分过滤器结合的必要性。



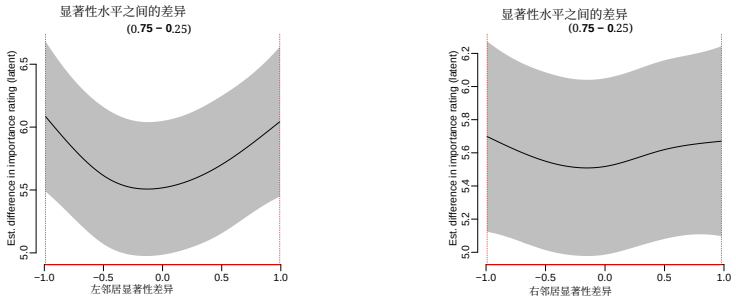

(a) 右侧与左侧之间的差异（分块）。等高线标示零值。(b) 差异（分块） $p$ 值。等高线标示 0.05。(c) 右侧与左侧之间的差异（无分块）。等高线标示零值。(d) 差异（无分块） $p$ 值。等高线标示 0.05。

图7：随机化解释实验中（无）词汇块邻居的差异和 $p$ 值。



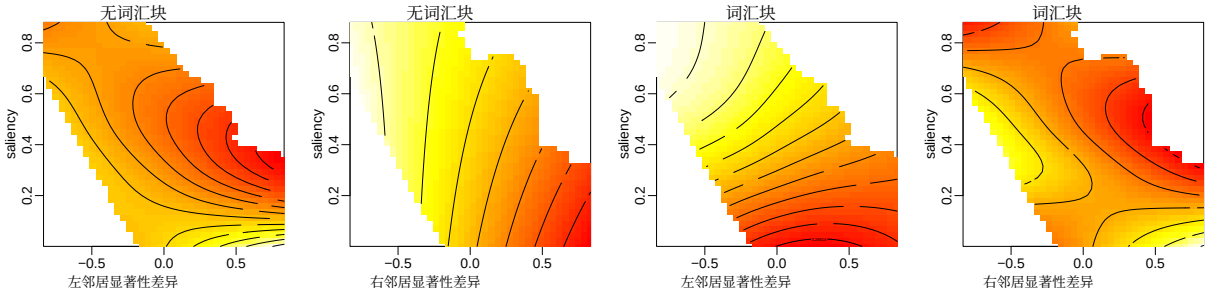

(a)（左侧邻居）分块与无分块之间的差异。等高线标示零值。(b)（左侧邻居）差异 $p$ 值。等高线标示 0.05。(c)（右侧邻居）分块与无分块之间的差异。等高线标示零值。(d)（右侧邻居）差异 $p$ 值。等高线标示 0.05。

图 8：我们随机化解释实验中左侧和右侧邻居的差异和 $p$ 值。



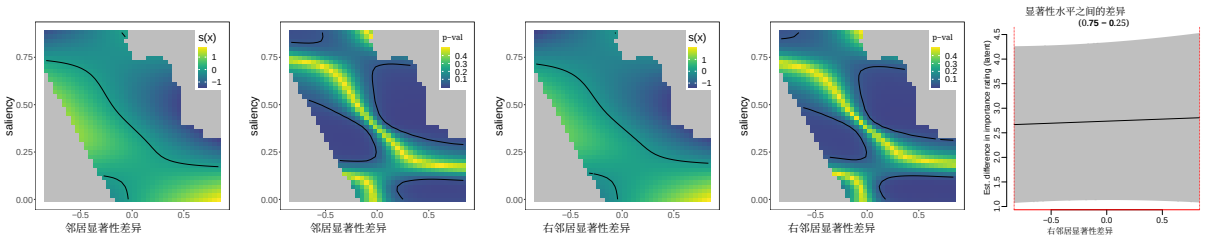

(a) 左。 (b) 正确。

图9：示例性高（0.75）和低（0.25）词语显著性级别之间显著性差异影响的差异图。红色 x 轴区域表示存在显著差异。



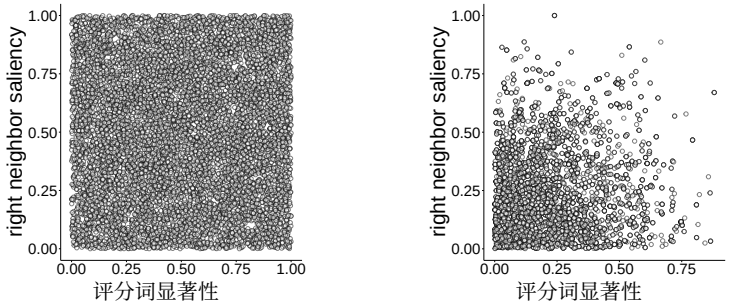

(a) 左侧，无分块。 (b) 右侧，无分块。 (c) 左侧，分块。 (d) 右侧，分块。（∗）

图10：在我们的 SHAP值实验中，左侧和右侧的邻居。（∗）标记统计显著的平滑项。颜色按图分别归一化。请注意，前三幅图对应不显著效应，其各自的颜色映射覆盖的数值范围较小。




(a) 右/左差异 (b) 右/左差异 (c) 分块/无分块差异（分块）（ $p$ ）值（等高线标记 0.05）。(d) 右邻居的分块/无分块差异（等高线标记 0）。(e) 随机化显著性与右邻居显著性之间的差异。（等高线标记 0）。

图 11：我们的 SHAP值实验结果差异图。等高线指等高线。 (e) 中的红色 x 轴标记显著差异。




(a) 随机化显著性。 (b) SHAP值显著性。

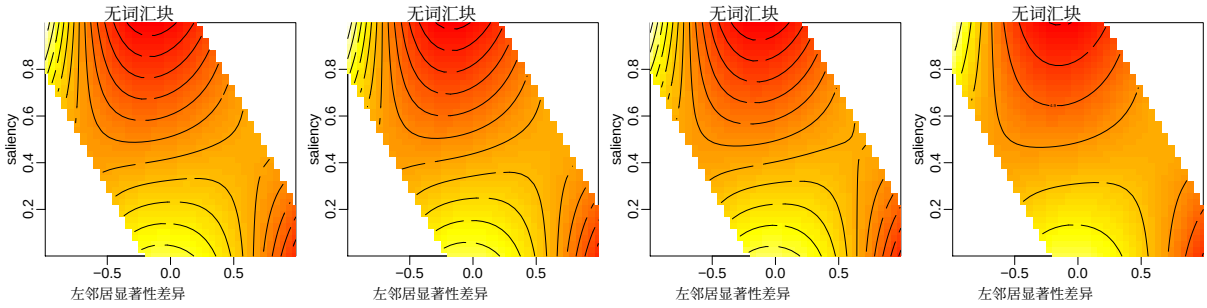
图12：在我们的随机化解释（左）和 SHAP值实验（右）中，评分词显著性与右侧邻居显著性的分布对比。

| Term | (e)df | 参考自由度 | F | $p$ |
|---|---|---|---|---|
| s(显著性) | 11.22 | 19.00 | 580.89 | <**0.0001** |
| s(显示索引) | 3.04 | 9.00 | 22.02 | <**0.0001** |
| s(词长) | 1.64 | 9.00 | 16.44 | <**0.0001** |
| s(句长) | 0.00 | 4.00 | 0.00 | 0.425 |
| s(相对词频) | 0.00 | 9.00 | 0.00 | 0.844 |
| s(归一化显著性排名) | 0.59 | 9.00 | 0.37 | 0.115 |
| s（词位置） | 0.58 | 9.00 | 0.18 | 0.177 |
| te(左侧差异、显著性)：无分块 | 3.12 | 24.00 | 1.50 | **0.002** |
| te(左侧差异、显著性)：分块 | 2.24 | 24.00 | 0.51 | **0.038** |
| te(右侧差异、显著性)：无分块 | 2.43 | 24.00 | 0.47 | **0.049** |
| te(右侧差异、显著性)：分块 | 0.00 | 24.00 | 0.00 | 0.578 |
| s(句子 ID) | 0.00 | 149.00 | 0.00 | 0.616 |
| s(显著性,句子 ID) | 16.13 | 150.00 | 0.14 | 0.191 |
| s(工作者 ID) | 62.19 | 63.00 | 30911.89 | <**0.0001** |
| s(显著性,工作者 ID) | 62.11 | 64.00 | 16760.88 | <**0.0001** |
| 大写 | 2.00 | | 3.15 | **0.042** |
| 依存关系 | 35.00 | | 2.92 | <**0.0001** |

表5：随机显著性实验结果详情。使用 $t = 87.5\%$和 $\varphi^2$ 测量时，单变量平滑项（顶部）、随机效应项（中部）和参数固定项（底部）的（有效）自由度、参考自由度和Wald检验统计量。

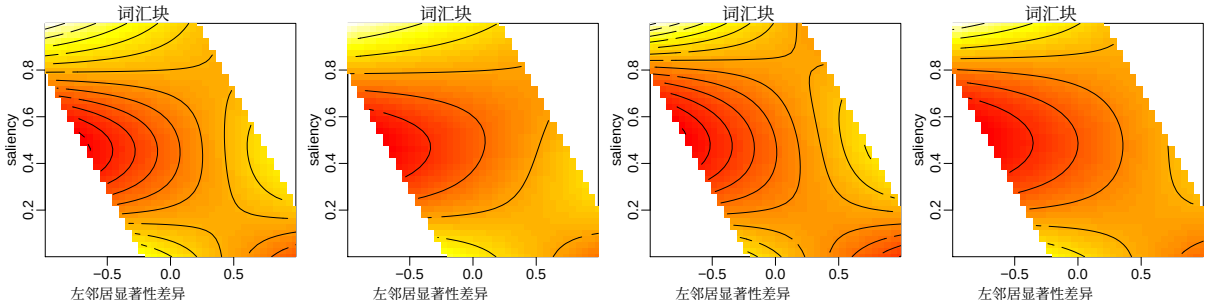
| Term | (e)df | 参考自由度 | F | $p$ |
|---|---|---|---|---|
| s(显著性) | 6.71 | 19.00 | 18.85 | <**0.0001** |
| s（显示索引） | 1.88 | 9.00 | 6.45 | <**0.0001** |
| s(词长) | 2.04 | 9.00 | 4.43 | <**0.0001** |
| s（句长） | 0.00 | 4.00 | 0.00 | 0.98 |
| s(相对词频) | 0.00 | 9.00 | 0.00 | 0.64 |
| s(归一化显著性排名) | 0.89 | 9.00 | 1.99 | **0.002** |
| s(词位置) | 0.42 | 9.00 | 0.12 | 0.19 |
| te(左差异、显著性)：无分块 | 0.00 | 24.00 | 0.00 | 0.37 |
| te(左侧差异、显著性)：分块 | 0.00 | 24.00 | 0.00 | 0.49 |
| te(右侧差异、显著性)：无分块 | 0.99 | 24.00 | 0.20 | 0.06 |
| te(右侧差异、显著性)：分块 | 3.24 | 24.00 | 1.09 | 0.01 |
| s(句子 ID) | 0.00 | 149.00 | 0.00 | 0.52 |
| s(显著性,句子 ID) | 11.31 | 150.00 | 0.10 | 0.14 |
| s(工作者 ID) | 34.77 | 35.00 | 14185.28 | <**0.0001** |
| s(显著性,工作者 ID) | 62.11 | 64.00 | 16760.88 | <**0.0001** |
| 大写 | 2.00 | 0.35 | 0.71 | |
| 依存关系 | 34.59 | 36.00 | 8468.22 | <**0.0001** |

表6：SHAP 实验结果详情。使用 $t = 87.5\%$ 和 $\varphi^2$ 测量时，单变量平滑项（顶部）、随机效应项（中部）和参数固定项（底部）的（有效）自由度、参考自由度及 Wald 检验统计量。



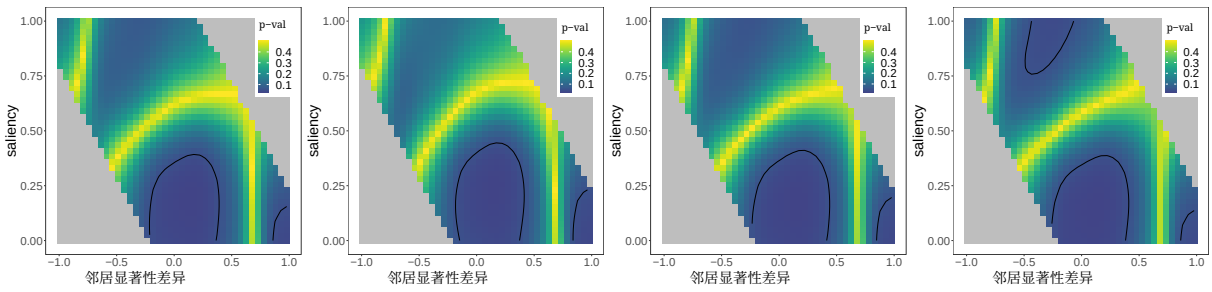

(a) $\varphi^2$。 (b) 杰卡德。 (c) 类似互信息。 (d) 泊松-斯特林。

图 13：在随机化解释实验中，针对外部块设置下左侧显著性差异的张量积交互，采用不同的共现度量。我们发现所有设置下的模式都相似。 $t = 87.5$ 在所有图中均保持一致。



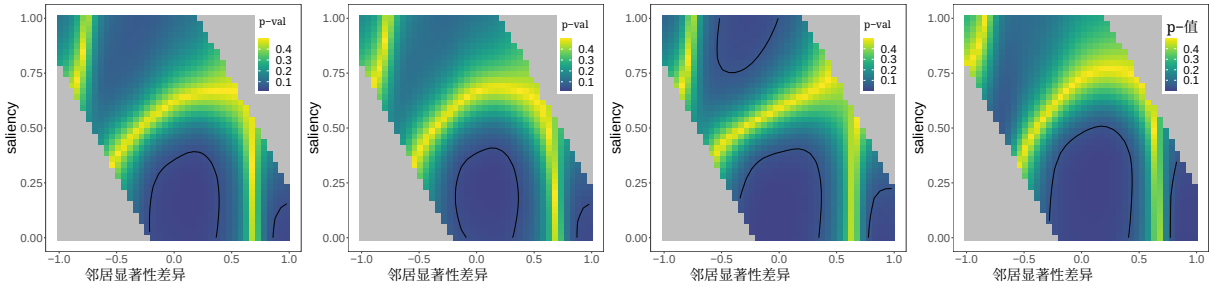

(a) $\varphi^2$。 (b) 杰卡德。 (c) 类似互信息。 (d) 泊松-斯特林。

图14：在块内设置中，针对随机化解释实验，在不同共现度量选择下的左侧显著性差异的张量积交互。我们在所有设置中都发现了相似的模式。 $t = 87.5$ 在所有图中均保持一致。



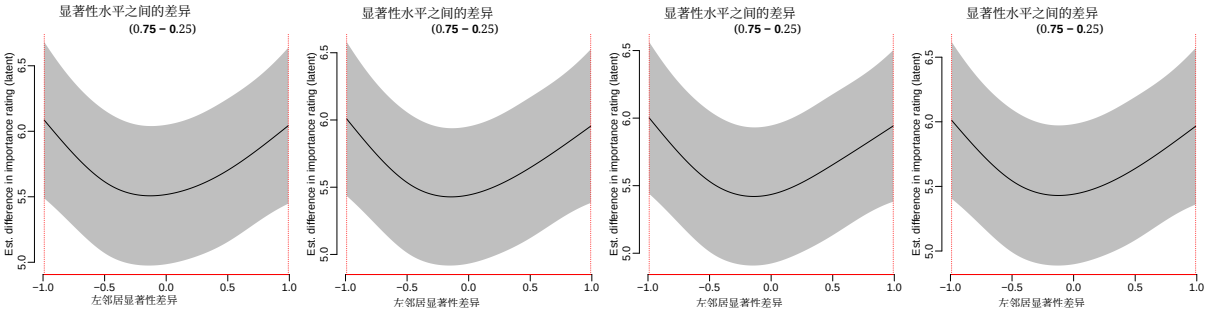

(a) $\varphi^2$。 (b) 杰卡德。 (c) 类似互信息。 (d) 泊松-斯特林。

图15：对于我们的随机化解释实验，采用不同的共现度量时，无词汇块邻词的右侧与左侧之间的 $p$ 值。我们发现所有设置下的模式都相似。所有图中的 $t = 87.5$ 均一致。



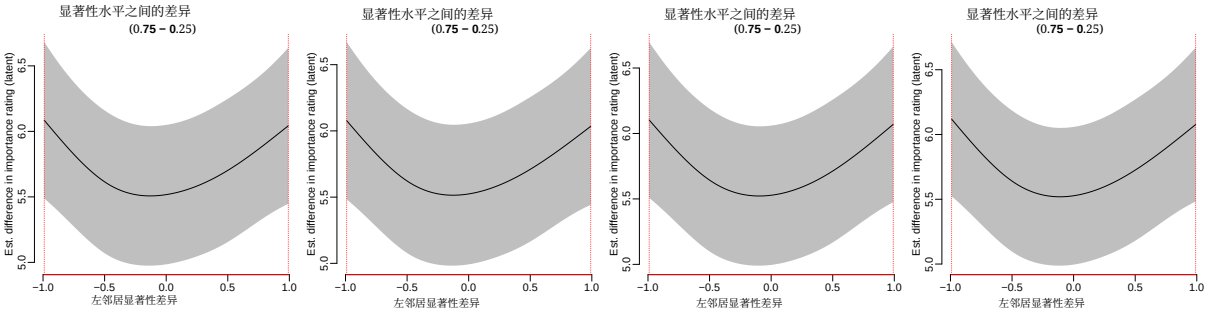

(a) $t = 87.5\%$ (b) $t = 75\%$ (c) $t = 50\%$ (d) $t = 25\%$

图 16：在随机化解释实验中，对于无词汇块邻居，在不同阈值选择下右侧与左侧之间差异的 $p$ 值。我们发现所有设置下的模式相似。所有图中均使用 $\varphi^2$ 测量。




(a) $t = 87.5\%$ (b) $t = 75\%$ (c) $t = 50\%$ (d) $t = 25\%$

图 17：在随机化解释实验中，不同阈值选择下，示例性高（0.75）和低（0.25）评分词显著性级别之间左侧显著性差异影响的差异图。我们发现所有设置下的模式相似。所有图中均使用 $\varphi^2$ 测量。




(a) $\varphi^2$。 (b) 杰卡德。 (c) 类似互信息。 (d) 泊松-斯特林。

图18：在我们的随机化解释实验中，针对不同共现度量选项，比较示例性的高（0.75）和低（0.25）评分词语显著性水平之间左侧显著性差异影响的差异图。我们发现所有设置下的模式都相似。 $t = 87.5$ 在所有图中都保持一致。

| Term | (e)df | 参考自由度 | F | $p$ |
|---|---|---|---|---|
| s(显著性) | 11.23 | 19.00 | 547.16 | < **0.0001** |
| s(display_index) | 3.10 | 9.00 | 20.93 | < **0.0001** |
| s(词_长度) | 1.61 | 9.00 | 16.47 | < **0.0001** |
| s(句子_长度) | 0.00 | 4.00 | 0.00 | 0.436 |
| s(相对_词语_频率) | 0.00 | 9.00 | 0.00 | 0.814 |
| s(归一化_显著性_排名) | 0.58 | 9.00 | 0.36 | 0.120 |
| s(词语_位置) | 0.59 | 9.00 | 0.18 | 0.173 |
| te(左侧差异、显著性)：无分块 | 2.90 | 24.00 | 1.21 | **0.003** |
| te(左侧差异、显著性)：分块 | 3.34 | 24.00 | 0.92 | **0.015** |
| te(右侧差异、显著性)：无分块 | 2.50 | 24.00 | 0.67 | **0.021** |
| te(右侧差异、显著性)：分块 | 0.00 | 24.00 | 0.00 | 0.836 |
| s(句子_id) | 0.00 | 149.00 | 0.00 | 0.601 |
| s(显著性,句子_id) | 17.35 | 150.00 | 0.15 | 0.178 |
| s(worker_id) | 62.19 | 63.00 | 30421.05 | < **0.0001** |
| s(显著性,worker_id) | 62.11 | 64.00 | 17591.01 | < **0.0001** |
| 大写 | 2.00 | | 3.01 | **0.049** |
| dependency_relation | 35.00 | | 2.93 | < **0.0001** |

表7：使用 $t = 25\%$和 $\varphi^2$测量进行随机化解释实验时，单变量平滑项（顶部）、随机效应项（中部）和参数固定项（底部）的（有效）自由度、参考自由度和Wald检验统计量。

| Term | (e)df | 参考自由度 | F | $p$ |
|---|---|---|---|---|
| s(显著性) | 11.21 | 19.00 | 584.57 | < **0.0001** |
| s(display_index) | 3.04 | 9.00 | 21.63 | < **0.0001** |
| s（词语_长度） | 1.63 | 9.00 | 16.66 | < **0.0001** |
| s(句子_长度) | 0.00 | 4.00 | 0.00 | 0.407 |
| s（相对_词_频率） | 0.00 | 9.00 | 0.00 | 0.813 |
| s(归一化_显著性_排名) | 0.56 | 9.00 | 0.32 | 0.130 |
| s(单词_位置) | 0.65 | 9.00 | 0.22 | 0.159 |
| te(左侧差异、显著性)：无分块 | 3.10 | 24.00 | 1.57 | **0.0010** |
| te(左侧差异、显著性)：分块 | 1.79 | 24.00 | 0.34 | 0.082 |
| te(右侧差异、显著性)：无分块 | 2.37 | 24.00 | 0.47 | **0.048** |
| te(右侧差异、显著性)：分块 | 0.64 | 24.00 | 0.05 | 0.249 |
| s(句子 ID) | 0.00 | 149.00 | 0.00 | 0.638 |
| s(显著性,句子 ID) | 17.14 | 150.00 | 0.15 | 0.164 |
| s(工作者 ID) | 62.19 | 63.00 | 30521.95 | < **0.0001** |
| s(显著性,工作者 ID) | 62.11 | 64.00 | 16749.25 | < **0.0001** |
| 大写 | 2.00 | | 3.23 | **0.039** |
| 依存关系 | 35.00 | | 2.94 | < **0.0001** |

表 8：在我们的随机化解释实验中，使用 $t = 87.5\%$和类MI度量时，单变量平滑项（上）、随机效应项（中）和参数固定项（下）的（有效）自由度、参考自由度和 Wald 检验统计量。