arXiv:2603.02427v1 [cs.HC] 2 Mar 2026

学习关注：调查数据中专心受访者与注意力不集中的受访者的无监督建模

Ilias Triantafyllopoulos

纽约大学斯特恩商学院，美国纽约州纽约市 ilias.triantafyllopoulos@nyu.edu

Panos Ipeirotis

纽约大学斯特恩商学院，美国纽约州纽约市 panos@nyu.edu

摘要

行为与社会科学调查的完整性取决于能否检测出提供随机或低投入作答的注意力不集中的受访者。传统防护措施（如注意力检测题）往往成本高、反应滞后且效果不一致。我们提出了一个统一的无标签注意力不集中检测框架，利用互补的无监督视角，通过几何重构（自编码器）和概率依赖建模（Chow-Liu 树）对回答一致性进行评分。我们引入了“百分位损失”目标函数，以增强自编码器对异常的鲁棒性；但我们的主要贡献在于识别出实现无监督质量控制所需的结构条件。在九个异质的真实世界数据集上，我们发现，检测效果与其说取决于模型复杂度，不如说取决于调查结构：具有连贯、相互重叠题项组的测量工具呈现出强协方差模式，使得即使是线性模型也能可靠地区分注意力集中的受访者与注意力不集中的受访者。这揭示了一个关键的“心理测量-机器学习对齐”：能够最大限度提高测量信度的相同设计原则（例如内部一致性），也能最大限度提高算法的可检测性。该框架为调查平台提供了一种可扩展、与领域无关的诊断工具，将数据质量与测量工具设计直接关联起来，从而无需增加受访者负担即可开展审查。

1 引言

行为、社会和政治学家高度依赖调查来了解人类情感和行为。然而，这类研究的效度始终受到“内容无响应性”的威胁——受访者因注意力不集中、疲劳或缺乏投入，无法持续如实反映其内在状态。研究人员将这一现象定义为回答在一致性上的丧失

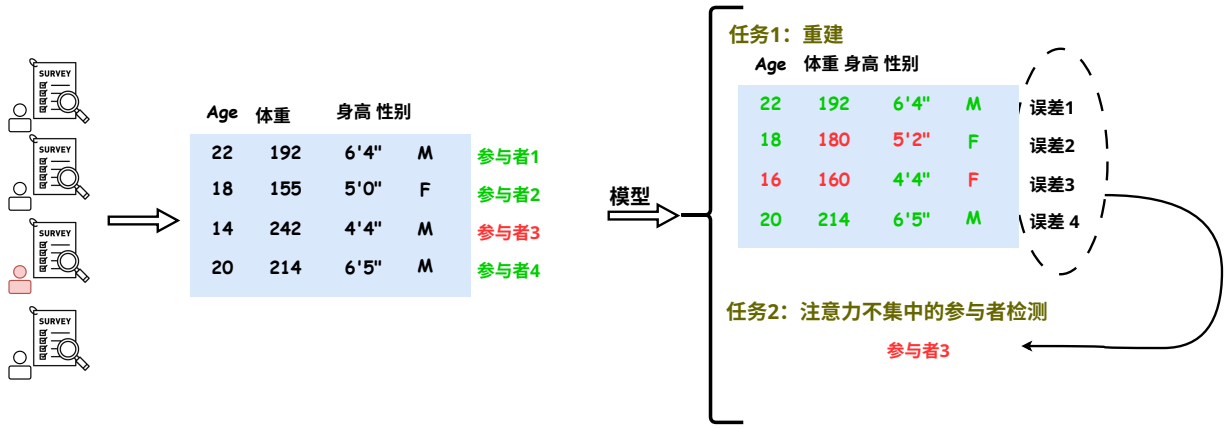


图1：用于说明所提方法的一个简单示例。多个参与者对各个问题的调查回答以结构化分类数据的形式表示（例如，年龄、身高、性别、体重）。模型首先学习重构每位参与者的回答（任务1），然后根据重构误差检测不连贯（难以建模）的回答。我们假设，存在许多此类不一致之处的参与者注意力不集中（任务2）。绿色表示专心受访者或重构良好的回答；红色表示注意力不集中的受访者或重构不佳的模式。这两项任务共同展示了无监督重构如何实现对低质量调查数据的自动检测。

对数据项的回答（Nichols等，1989年）。受访者可能因调查疲劳或不愿配合而随机作答，从而影响研究结果的效度（Meade和Craig，2012年）。随着人们日益依赖Amazon Mechanical Turk（MTurk）和Prolific等在线众包平台，这一问题愈发严重。尽管这些平台能快速收集多样化的数据，但参与者在研究期间更可能同时处理多项任务（Good- man等，2013年），注意力不集中的回答或欺诈性回答也更为常见，尤其是在参与者有机会作弊的情况下（Peer等，2022年）。

目前，用于缓解这些问题的技术仍主要采取被动应对的方式。研究人员依赖嵌入式注意力检测题（Curran，2016年）、个体回答变异指标（Weaver 和 Prelec，2013）等机制设计方法，例如贝叶斯真理血清。这些方法虽有用，却给研究过程带来了显著的“负担”：它们会增加受访者的认知负荷、延长调查完成时间，还可能引发测量反应性。

无监督机器学习提供了一种颇具前景的替代方案：直接对回答模式的“一致性”建模，以此检测注意力不集中，而无需带标签的训练数据或明确的“陷阱”问题。至关重要的是，这种方法解决了该领域有监督分类的一项根本局限：客观真实标签的缺失。收集受访者数据时，我们无法确切判断哪些人是随机作答，或是否提供了捏造的信息，因为我们没有真实标签可供比较。创建包含此类信息的数据集需要耗费大量成本和时间；即使能够实现，也很可能包含偏差和不一致之处，因为不同研究对注意力不集中的定义各不相同。因此，无监督学习技术不仅高效，还提供了一种必要的、无需依赖真实标签的替代方案。

在本研究中，我们不再局限于简单的算法基准测试，而是致力于构建一套结构理论

无监督数据质量。我们统一并评估了三类与领域无关的方法：(i) 非线性自编码器（AE），(ii) 线性（零层）自编码器，以及 (iii) 采用 Chow–Liu 树结构的生成式贝叶斯网络（Chow 和 Liu 1968 ）。据我们所知，此前没有研究对这些方法进行并列基准测试，以检测结构化调查数据中的注意力不集中。此外，我们通过实证刻画了一种关键现象：随着自编码器的重构能力增强，足以拟合所有输入，异常敏感度可能会降低，因为即使是不连贯的模式也开始能够被良好重构。为解决这一重构-检测权衡以及“污染学习”（模型学会重构噪声）这一挑战，我们提出了百分位损失（PL）目标函数，将一种稳健训练技术应用于调查领域。

这种无标签方法对于可扩展部署至关重要。首先，它避免了测量反应性，并降低了自动化风险。其次，它仍与缺乏嵌入式核查的遗留数据集兼容。第三，它降低了对特定情境规则的依赖，从而能够在异质样本中进行部署前筛查和收集后审计（Hong等人 2020）。

我们的贡献如下：

1. **基于“未清洗”数据的全面基准测试：**验证无监督检测需要原始数据集，其中不专心作答还未被滤除——这在数据通常已经预先清洗的公共存储库中十分罕见。我们整理了九个涵盖不同人群（青少年、MTurk工作者、代表性样本）和主题的异质真实世界数据集。这使我们能够在缺乏完美真实标签的情况下，建立首个严格且与领域无关的注意力不集中检测基准。

2. **心理测量-机器学习对齐：**利用这一多样化测试平台，我们证明无监督检测的成功主要取决于调查结构，而非数据集规模或模型复杂度。我们发现，各种用于检测注意力不集中的无监督模型的性能，与基于协方差结构构建的设计良好的调查问卷相关（例如，具有一致性且相互重叠的题项组）（克龙巴赫 1951，Clark和Watson 1995，2019年）。由此形成一条有力的设计原则：具有高“一致性”的工具本质上更容易通过算法进行质量控制。

3. **稳健百分位损失（PL）：**我们在本研究中为自编码器引入百分位损失。PL 不再最小化平均重构误差，而是最小化误差最低的 p 百分位样本的平均误差。它通过聚焦主体结构，同时降低对不连贯“噪声”案例的重视，解决了重构-检测权衡问题，并防止自编码器对异常过拟合。

4. **概率基线：**我们将 Chow–Liu 树应用于分类调查数据，并使用其逐行似然作为典型性得分。尽管它已广泛用于建模，我们仍展示了它作为一种可解释、无监督的注意力不集中检测器所具备的竞争力；它能补充基于重构的视角。

5. **可执行框架：** 我们将这些发现转化为关于如何设计调查问卷以及设置可嵌入组织调查平台的适应性流程管线的具体建议。

图 1 展示了我们的两阶段工作流：(1) 通过无监督建模重构或以概率方式解释典型模式；随后 (2) 根据重构误差或（负）似然对受访者排序，以检测注意力不集中。我们的训练方法完全不依赖标签，也不需要额外题项或注意力检测题，因此适用于现有数据集以及部署前的质量监控。

2 相关工作

2.1 问卷不认真作答检测

传统的注意力不集中检测方法依赖于嵌入式设计特征，例如注意力检测题、作答时间阈值，或直线作答等基于模式的指标（Kim 等人，2019）。贝叶斯真理血清（Weaver 和 Prelec，2013）则代表了另一类方法，通过机制设计激励受访者如实作答。更先进的统计与机器学习方法包括基于 EM 的真相发现（Dawid 和 Skene，1979）、无监督可信度建模（Yin 等人，2007），以及利用行为特征进行监督分类（Schroeders 等人，2022、Ozaki，2024）。然而，监督方法高度依赖稀缺且针对特定问卷的标签，而这些标签通常来自注意力检测题，只能捕捉部分注意力不集中现象。这限制了此类方法的可扩展性和泛化能力，因此有必要采用无监督、数据驱动的替代方法，直接根据作答结构推断受访者的一致性。

基于项目反应理论（IRT）的心理测量学方法通过个体拟合统计量在个体层面进行筛查，从而标记在拟合模型下不太可能出现的回答模式。一个典型例子是标准化对数似然指数 l_z，它根据估计的项目参数和受访者的估计潜在特质，评估其完整项目得分向量有多出人意料（Drasgow 等人，1985）。在此情境下，较大的负值表示粗心作答。后续研究提出了修正版本，例如 l_z^*，用于校正使用估计能力水平而非真实潜在特质时产生的统计偏差（Snijders，2001）。当完整的 IRT 校准在计算上或实践中难以实施时，较简单的“分组法”代理指标会通过检查受访者的回答是否与经验项目易度相符，来近似体现相同的思路（即样本中每个项目的平均作答率或平均分数）。例如，个体总分相关 r_{pt} 计算个体回答向量与按难度排序的项目均值向量之间的相关性（Donlon 和 Fischer，1968）。r_{pt} 接近零或为负的值表明，受访者可能在随机作答，或其作答方式与总体其他成员相比缺乏一致性。

2.2 用于异常检测和调查注意力不集中的自编码器

自编码器最初被用于无监督异常检测，因为它们能够重构典型输入，同时对异常输入产生更高的误差（Hawkins 等人，2002）。这一原理后来通过多种变体得到拓展，例如去噪自编码器（Vincent 等人，2010）、变分自编码器（Kingma 和 Welling，2013）和鲁棒深度自编码器（Zhou 和 Paffenroth，2017），这些方法都旨在提升对噪声的鲁棒性或显式建模离群值。这些创新确立了自编码器作为一种灵活基础，可用于不同领域的异常检测（Goodge 等人，2021、Rubio 等人，2020）。

直到最近，自编码器才被专门应用于不专心作答问题。Alfons 和 Welz (2024) 引入自编码器来分类不专心作答，但其评估仅限于合成数据。Welz 和 Alfons(2023年) 提出了 CODERS，该方法将自编码器重构误差与变点检测相结合，以识别受访者在调查过程中何时开始敷衍作答，并在一个真实数据集和多个合成数据集上进行了概念验证。我们的工作在三个关键方面有所不同。首先，我们针对分类调查数据调整自编码器，引入变量级损失加权和百分位损失，以应对重建—检测权衡。其次，我们不关注调查过程中敷衍作答的起始时点，而是着眼于识别整体上注意力不集中的受访者。第三，我们在九个多样化的真实世界调查数据集上验证了这一方法，这些数据集包含认真作答和不专心作答的参与者，并建立了大规模实证基准。由此，我们将自编码器定位为一种可扩展且不受领域限制的工具，可用于调查实践中的注意力不集中检测。

2.3 概率方法

近年来，研究者已利用概率模型来检测调查数据中的异常或注意力不集中的受访者。贝叶斯网络可以近似题目回答的联合分布，从而标记离群值（Lindskou 等，2021）。潜在类别分析和混合建模方法也通过设定一个与正常类别相区别的“离群值”潜在类别，以识别离群值。例如，混合项目反应理论模型将努力不足的受访者（其回答近乎随机或与内容无关）与专心受访者区分开来，从而改进参数估计和个体拟合度（Ulitzsch 等，2024）。这些概率框架还展现出实际应用价值，例如通过贝叶斯潜在类别联合模型检测自动化的“机器人”受访者（Roman 等，2022）。

2.4 调查设计与检测可靠性

重叠或连贯的题项组对注意力不集中检测的有效性有着重大影响。心理测量理论认为，冗余且相互关联的题项能够提供内置的一致性检查，从而识别粗心作答。当多个题项测量同一构念（例如，一组相关问题构成的题项组）时，注意力集中的受访者应以一致的方式作答，从而产生较强的共享方差和较高的内部信度。相比之下，粗心或随机作答者往往会因对相关题项作答不一致而破坏这种协方差模式。从实证角度看，预期相关结构的存在使离群值更容易被发现。例如，心理测量同义或反义题对（几乎重叠的题项）已被用于检查一致性：对于两个理应引出相似回答的问题，受访者若给出截然不同的答案，就很可能没有集中注意力（Meade和Craig，2012年，Goldammer 等人 2020）。如果每个题项都各不相同（测量完全不同的构念），这类检查的效力就会大大降低，因为不存在可供违反的预期模式。因此，研究人员强调要谨慎设计调查，构建范围充分且具有适当冗余度的连贯题项组，以提高数据的可靠性（Clark和Watson 2019年）；我们认为，这也有助于检测不专心作答者。

表1. 数据集摘要

数据集	样本	变量	特征	AFV
Robinson-Cimpian (2014)	14,765	98	619	6.32
Pennycook 等(2020)	853	188	708	3.75
Alvarez 等 (2019年)	2,725	39	196	5.03
Uhalt (2020)	308	60	337	5.62
O'Grady 等(2019年)	355	72	322	4.47
Buchanan和Scofield（2018年）	1,038	23	159	6.91
Moss 等人 （2023年）	2,277	51	332	6.51
Mastroianni和Dana （2022）	1,036	51	322	6.31
Ivanov等人(2021)	860	67	310	4.63

每个样本由若干变量组成。变量可以是问题，也可以是人口统计记录。每个变量都由特征构成，具体说明见第 4.2 节。最后一列还给出了每个变量的平均特征数（AFV）。

3 数据

我们介绍研究所用的数据集，概述每个数据集的主要特点，并说明数据的收集方式以及其中包含哪些注意力检测题。在表1中，我们汇总了所有数据集的统计信息，包括样本数量、变量数量、特征数量以及每个变量的平均特征数。我们使用 https://datasetsearch.research.google.com/查找公开可用的调查数据集，这些数据集包含注意力检测题，且主要采用结构化/分类回答（而非文本或其他形式的非结构化数据）。我们确定了九个数据集，它们在主题、受访者群体和质量控制机制方面存在显著差异，以确保评估能够反映该方法在不同情境下的稳健性。这种异质性涵盖：（1）受访者类型（如青少年（Robinson-Cimpian 2014）、MTurk工作者（Moss 等人 2023年）以及具有全国代表性的成年样本（Mastroianni and Dana 2022）；（2）调查主题（从政治态度到错误信息易感性）；以及（3）注意力检测题设计（从没有检测题（Robinson-Cimpian 2014）到嵌入多项核查（Pennycook 等人 2020）。除了考虑多样性，我们还采用了两项纳入标准：（a）数据集必须包含注意力检测题（对于 Robinson-Cimpian (2014) 的情况，则须包含能够明确检测受访者是否注意力不集中的知识检测题）；（b）必须保留未通过这些检查项的参与者回答。这些标准的限制性很强，因为大多数已发表研究只公开经过"清理"的数据集，其中已删除注意力不集中的受访者，这使得大规模评估注意力不集中检测颇具挑战性。

Robinson-Cimpian (2014)，调皮回答者：该数据集来自2012年戴恩县青少年评估（DCYA），这是一项面向14,765名美国高中生开展的匿名网络调查。该研究考察了"恶作剧型受访者"——即故意给出极端或不合情理回答（例如夸大报告健康行为）的青少年，这些回答会扭曲性取向、性别认同和残障等类别之间的群体间差异估计。研究考察的结果包括自杀意念、学校归属感和物质使用。与本研究中的其他数据集不同，这项调查没有嵌入注意力检测题；研究人员转而依据领域知识以及受访者对无关问题的回答来识别注意力不集中的案例。因此，该数据集颇具特色，因为它提供了一个大规模场景，在其中注意力不集中是通过推断得出的

其依据是极端或前后不一致的作答模式，而非明确的筛查题项。

Pennycook 等人 （2020），COVID-19 错误信息：本研究考察了注意力与认知反思如何影响 COVID-19 错误信息的传播。共有 853 名美国受访者评估了以社交媒体形式呈现的 30 条新闻标题（15 条真实、15 条虚假），其真实标签由事实核查来源核实。核心目标是检验错误信息的分享究竟是因为人们未能考虑准确性，还是因为他们确实相信虚假说法。为此，该研究设置了四种实验条件，同时改变问题的表述框架（准确性与分享意愿）以及回答选项的顺序。数据质量通过多项注意力检测题进行评估：两道选择题指令题项、一道要求选择特定数值（“3”）的嵌入式李克特量表题项，以及一道额外的自我报告题项，询问参与者是否随机作答。这种结构化条件与多样化注意力检测题的组合，使该数据集特别适合在不同操作定义下评估注意力不集中检测。

Alvarez 等人 （2019年），注意力不集中：本调查通过 Qualtrics 的 e-Rewards 样本库收集了 2,725 名加利福尼亚州成年人的回答，以研究注意力不集中如何影响政治态度测量。该研究考察了注意力不集中的受访者是否会引入噪声、满意化或错误报告，以及他们在人口统计学特征上是否与认真作答的参与者不同。数据质量通过三道陷阱问题进行监测：两道要求特定答案的选择题，以及一道要求填写“government”一词的开放文本题项。重要的是，完整数据集同时包含专心受访者与注意力不集中的受访者，使其成为在具有连贯、多题项政治构念的领域中评估检测方法的强基准。

Uhalt (2020), 注意力检查与回答质量：这项 Qualtrics 调查纳入了 308 名参与者，重点是使用李克特量表题项对人格特质进行自我评估。为评估作答质量，该工具嵌入了多项明确的注意力检测题，要求参与者选择特定选项（例如，“如果我在认真关注本调查，我会选择‘强烈同意’”）。这些检查项分布在整份问卷中，用以监测参与投入程度。该数据集提供了一个有用的测试案例，因为它将结构化的人格量表（注意力集中的回答应表现出内部一致性）与来自嵌入式核查的明确真实标签结合起来。

O’Grady 等 (2019年) 道德基础：本研究从美国商科本科生中收集了 355 份有效回答，以考察道德基础如何预测亲社会行为。参与者完成了道德基础问卷（MFQ），该问卷测量个体化基础（关怀、公平）和约束性基础（忠诚、权威、纯洁）。数据质量通过嵌入调查中的一道多项选择指令操纵检查（IMC）进行监测，以确保受访者仔细阅读说明。该数据集提供了一套结构化的态度测量题组，并包含一道用于识别注意力不集中的注意力检测题。

Buchanan和Scofield (2018年)，低质量数据：这项在线 Qualtrics 调查收集了 1,038 份有效回答，并考察了心理学研究中检测低质量数据的方法。研究利用作答时间、点击次数和分布异常等行为与统计指标，探讨了注意力不集中的回答、敷衍回答和自动化回答（例如问卷机器人）。研究还通过嵌入式注意力检测题监测数据质量，要求参与者“请对本题选择强烈同意”。该数据集的独特之处在于，将传统调查回答与基于元数据的质量指标相结合，因此适用于测试仅依赖回答模式的检测方法。

Moss 等人 (20 23)，伦理数据：研究1调查了2,277名常驻美国的活跃MTurk工作者，以

考察他们对MTurk的经济依赖、时间投入以及对公平的看法。该数据集具有异质性，涵盖了工作者体验的多个方面。我们通过两项筛查来监控数据质量：一项要求参与者选出对先前某个条目的正确概述，另一项则直接询问参与者是否同意以下陈述："我没有在阅读本调查中的问题。"这种组合既提供了对注意力不集中的间接测量，也提供了自我报告测量。

Mastroianni和Dana (2022)，态度变化： 研究1通过Prolific从具有全国代表性的美国成年人成年人样本中收集了1,036份答复，以考察社会态度长期变化的看法。参与者将自己对历史公众舆论趋势的估计与实际民调数据进行比较，研究尤其关注人们是否会系统性地高估自由化程度。研究在调查题目中嵌入了一道注意力检测题，要求受访者在两个空格中都填写数字"1"，以确保数据质量。该数据集提供了一个高质量、具有代表性的基准样本，并包含一道用于检测注意力不集中的指令操纵检查。

Ivanov等人 (2021) 种族怨恨： 这项MTurk调查于2020年4月收集了860份答复，以研究人们对新冠肺炎期间减少监禁的态度。研究考察了有关监狱健康风险的信息、种族怨恨和共情如何影响人们对释放在押人员的支持，以及支持程度是否因犯罪类型、年龄或在押人员的健康状况而异。研究通过两道嵌入式注意力检测题监测数据质量。该数据集提供了一种异质性的态度测量工具，包含多道检查项，可用于评估政治敏感情境下的注意力不集中。

4 方法

本研究探讨用于检测注意力不集中的调查受访者的无监督领域无关机器学习方法。我们比较了三类方法：(i)结合定向百分位数损失的非线性自编码器，(ii)没有隐藏层的基础线性自编码器，以及(iii)概率贝叶斯网络。所有方法均应用于结构化分类变量和经分类处理的数值变量。

4.1 数据预处理

在我们的实验中，我们仅关注分类变量。不过，我们也考虑数值变量：如果其不同取值少于 20 个，就将其视为分类变量；否则，根据其标准化值将其离散化为预定义类别。首先，我们对每个数值变量进行标准化 $z = (x - \mu)/\sigma$ 其中，$\mathbf{x}$ 表示原始值，μ 表示均值，σ 表示该变量的标准差。随后，我们将这些值归入六个离散区间，如表2所示。

诚然，许多问题都是开放式的，需要以文本作答。我们仅关注结构化的分类变量和经过分类的数值变量，因此在分析中排除了开放式回答。我们将对包含开放式回答的调查问卷进行分析，作为未来一项富有价值的研究方向。

表2. 数值变量分类

类别	标准化值范围
底部极端 Low	$z < -1.4$
	$-1.4 \leq z < -0.7$
正常	$-0.7 \leq z \leq 0.7$
High	$0.7 < z \leq 1.4$
顶部极端	$z > 1.4$
缺失数据	N/A

基于标准化值对数值变量进行分类。

4.2 数据准备

本研究使用的数据集是调查问卷，每位参与者都要回答一系列通常为多项选择题的问题。由于大多数问卷条目采用封闭式结构，数据本质上属于分类数据。我们所有的方法都适用于数值数据，因此会将分类变量转换为适合网络使用的数值表示。每个调查问题（变量）都采用独热编码，其中每个可能的回答选项都表示为一个单独的二元特征。具体而言，对于有 k 个可能答案选项的问题，会创建一个长度为 k 的向量（如果存在跳过该问题的用户，则为 k+1 ），其中只有所选答案对应的位置标记为 1（其余元素设为 0）。这种转换形成的特征空间，其输入维度远大于原始问卷条目的数量，因为每个分类变量都会扩展为多个二元特征。每个模型都将每份回答视为一组特征，并学习整个回答集合的底层分布。

4.3 非线性自编码器

自编码器是一类神经网络，旨在通过将输入编码到低维潜在空间，并从这种压缩表示中重建数据，以无监督方式学习数据的紧凑表示。自编码器由两部分组成：编码器用于学习数据在较低维度下的有效表示；解码器则学习如何从这种低维表示还原初始数据（图2）。

编码器： 编码器函数记为 f_θ，它将输入响应向量 $\mathbf{x} \in \mathbb{R}^d$ （其中 d为特征数量）映射为低维潜在表示 $\mathbf{z} \in \mathbb{R}^m$ （维度为 $m \ll d$）：

$$\mathbf{z} = f_\theta(\mathbf{x}) = \sigma(W_e\mathbf{x} + \mathbf{b}_e)$$

其中 $W_e \in \mathbb{R}^{m \times d}$ 是权重矩阵，$\mathbf{b}_e \in \mathbb{R}^m$ 是偏置项， $\sigma(\cdot)$ 是一种非线性激活函数，例如 ReLU 或 Sigmoid函数。

解码器： 解码器函数记为 g_ϕ，它根据潜在表示重构原始响应向量：

$$\hat{\mathbf{x}} = g_\phi(\mathbf{z}) = \sigma(W_d\mathbf{z} + \mathbf{b}_d)$$

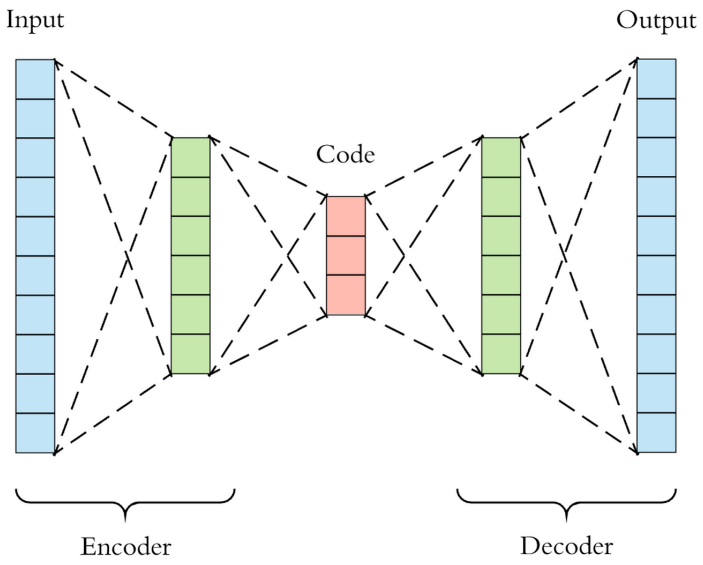


图2：简单自编码器的架构。自编码器由两部分组成：(a) 编码器，将信息编码为潜变量；(b) 解码器，将信息解码还原为初始输入。来源

其中 $W_d \in \mathbb{R}^{d \times m}$ 为解码器权重矩阵，$\mathbf{b}_d \in \mathbb{R}^d$ 为偏置项，$\hat{\mathbf{x}} \in \mathbb{R}^d$ 为重构响应向量。

潜变量的数量 m，以及编码器和解码器的深度、每层单元数和激活函数的选择，都是可以调整以获得最佳性能的超参数。为防止过拟合并增强泛化能力，我们对网络的权重应用 L2 正则化（也称为权重衰减）Krogh 和 Hertz(1991)。正则化后的目标函数如下：

$$\mathcal{L}_{\text{reg}} = \mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \sum ||W||_2^2$$

其中 λ 为 一种正则化超参数，用于控制对较大权重值的惩罚 s.

此外，训练期间还会在隐藏层中使用 dropout，以随机停用神经元Srivastava 等人(2014）。训练时，每个单元都以概率p_{drop}被丢弃：

$$\mathbf{h}^{(l)}\text{drop} = \mathbf{r} \odot \mathbf{h}^{(l)}, \quad r_i \sim \text{Bernoulli}(1 - p_{\text{drop}})$$

其中，$\mathbf{h}^{(l)}$ 是第 l 层的激活向量，$\odot$ 表示与丢弃掩码$\mathbf{r}$进行逐元素乘法。p_{drop} 也是一个需要调整的超参数。

损失函数。对于分类调查响应，训练自编码器时使用二元交叉熵（BCE）损失：

$$\text{BCE}(\mathbf{x}, \hat{\mathbf{x}}) = -\sum_{i=1}^{d} \Big[x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i) \Big].$$

如前所述，我们的数据集由变量（问题）组成，每个变量在独热编码后会扩展为多个二元特征。为尊重这一嵌套结构，我们修改

二元交叉熵损失，以按变量计算误差，并按特征数量的对数进行归一化，从而使具有较多回答选项的问题不会不成比例地主导总损失。形式上：

$$\mathcal{L}_{\text{base}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{1}{\log |F_v|} \sum_{f \in F_v} \text{BCE}(x_f, \hat{x}_f),$$

其中，V 是调查变量（问题）的集合， F_v 是变量 v 的特征（可能的回答）集合。

虽然这种基础损失有助于忠实地重构所有输入模式，但自编码器的设计目标是尽可能降低整体重构误差。这对于我们的不专心用户检测目标而言可能并非最优：我们主要关注的并非同样出色地重构每个回答，而是对数据中一致且占多数的模式进行建模，同时让不一致的模式以较高误差凸显出来。为此，我们引入百分位损失（*PL*）（Merrill 和 Olson（2020）），将其作为基础公式的扩展。PL 最初应用于计算机视觉领域，旨在对常见背景而非罕见异常进行建模。PL 背后的原理是，假设异常很少见，那么一个批次中第95百分位损失对应异常的概率在统计上极低。因此，PL 将学习重点放在批次中误差最低的子集上，从而避免模型对异常输入过拟合。

令l_j 表示样本 j 在 L_{base}下的总重构损失。对于包含 N 个样本的小批量，我们按升序排列这些损失：

$$\ell_{(1)} \leq \ell_{(2)} \leq \cdots \leq \ell_{(N)}.$$

给定选定的百分位数 $p \in (0, 100],$，我们仅保留误差最低的子集，其大小为 $k = \left\lfloor \frac{p}{100} \cdot N \right\rfloor$，并忽略误差最高的样本。百分位损失定义如下：

$$\mathcal{L}_{\text{PL}} = \frac{1}{k} \sum_{j=1}^{k} \ell_{(j)}.$$

在实践中，设置 $p < 100$ 会使学习重点转向准确重构数据分布中的大部分数据，同时有意降低对非典型回答的重视程度。这在我们的情境中能达到预期效果：产生一致模式的专心受访者会塑造模型的表示，而粗心作答或随机作答的注意力不集中的受访者则会产生明显更高的重构误差，因此更容易被标记为异常。

随机性检测协议： 如前所述，自编码器适用于捕捉数据集中各特征之间的内部相关性，并构建能够保留这些依赖关系的结构化潜在表示。鉴于自编码器能够学习有效回答模式的紧凑且有意义的表示，我们假设自编码器可以根据重构损失有效区分专心受访者和注意力不集中的受访者。

异常检测中的一个关键方法论区别，是归纳式 （在未来数据中检测新的异常）与转导式 （在当前数据集中检测异常）设定之间的区别。调查审核通常属于转导式：研究人员收集一批 N 回复，并旨在清理这批特定数据。因此，我们的主要评估是在

完整数据集上训练自编码器，并计算相同样本的重构误差。虽然这种“在测试集上训练”的做法对监督分类而言无效，但在无监督异常值检测中却很常见，因为自编码器经过正则化（通过瓶颈 $m \ll d$ 和百分位损失），能够捕捉多数流形。根据定义，即使异常存在于训练集中，也无法很好地拟合该流形（Zhou 和 Paffenroth，2017，Eduardo等人 2020）。

因此，我们按重构误差从高到低排序，将损失最高的受访者识别为潜在的不专心参与者。我们的假设基于异常检测和离群值检测领域的大量既有研究：在这些研究中，自编码器及类似的深度学习架构已成功用于识别偏离预期模式的数据点；应用领域包括计算机视觉（An和Cho 2015，Zhou 和 Paffenroth，2017）、时间序列（Xu等人 2018年）和表格数据（Eduardo等人 2020）。

我们的方法建立在这样一个假设之上：注意力不集中的受访者缺乏一致的回答模式，因而会产生自编码器无法重构的随机性。相比之下，观点连贯但独特的少数群体仍可被重构，因为其内部一致性使其处于学习到的流形之中。这一区别与早期的降维技术（如主成分分析（PCA）（Abdi和Williams 2010））相呼应：结构化的少数群体模式由较弱的成分捕捉，而纯随机回答则无法表示。因此，自编码器主要标记的是不一致性，而非合理的少数群体观点，这阐明了我们方法的一项重要适用边界。

设置：对于每个数据集（调查），我们在训练自编码器之前分别进行超参数调优。我们采用贝叶斯优化来高效探索超参数空间。该优化在 Keras 框架内使用 KerasTuner 包进行。表 3 概述了调优过程中探索的超参数及其各自的搜索范围。优化过程在 30 次试验中评估各种配置，使用 20% 的验证集划分，并将每个候选模型最多训练 300 个轮次，早停设为 10 个轮次。根据验证损失最小化选择表现最佳的配置。

4.4 基础线性自编码器

为建立简单基线，我们还实现了一个线性自编码器，可将其视为自编码器范式最基本的形式。其总体结构与非线性自编码器相同：编码器将输入回答映射到压缩的潜在空间，解码器再根据该表示重构原始输入。关键区别在于我们移除了所有非线性与中间层。实际上，这意味着编码器和解码器都简化为单一线性变换：

$$\mathbf{z} = W_e \mathbf{x}, \qquad \hat{\mathbf{x}} = W_d \mathbf{z},$$

其中 $W_e \in \mathbb{R}^{m \times d}$ 是将输入维度 d 投影到潜在维度 m的编码器矩阵，而 $W_d \in \mathbb{R}^{d \times m}$ 是投影回原始空间的解码器矩阵。

线性自编码器仅限于捕捉数据中的线性相关性。因此，模型在数学上更简单、计算开销也更小，但表达能力不如其非线性对应模型。尽管存在这一局限，线性自编码器仍提供了一个基线，用以评估

表 3. 超参数搜索空间

超参数	探索的取值
学习率	{0.0001、0.001、0.01}
编码器层数	{1, 2, 3}
编码器单元	{64, 96, 128, 160, 192, 224, 256}
编码器激活函数	{ReLU、SeLU、GeLU、Swish}
编码器正则化	{0.0, 0.001, 0.01}
编码器丢弃率	{0.0, 0.1, 0.2, 0.3, 0.4, 0.5}
编码器批归一化	{是}
潜在空间维度	{2, 3,....., 50}
潜在激活函数	{ReLU、SeLU、GeLU、Swish}
解码器层	{1, 2, 3}
解码器单元	{64, 96, 128, 160, 192, 224, 256}
解码器激活函数	{ReLU、SeLU、GeLU、Swish}
解码器正则化（L2）	{0.0, 0.001, 0.01}
解码器丢弃	{0.0, 0.1, 0.2, 0.3, 0.4, 0.5}
解码器批归一化	{True}

在我们的调优实验中，使用贝叶斯优化进行超参数搜索。无论编码器和解码器每次选择多少层，探索的单元数、激活函数、正则化、Dropout 和批归一化取值在各层中均相同。

更深层的非线性架构所带来的额外价值。线性自编码器可视为与主成分分析（PCA）相似，但省略了一些约束，例如额外的正交性约束。

在训练方面，我们保留了与非线性情形相同的基于重构的目标函数。模型使用上一节所述的基础损失函数进行优化，以确保与更复杂的自编码器架构具有可比性。

4.5 贝叶斯概率网络

我们使用通过 Chow–Liu 算法学习得到的树结构贝叶斯网络，对分类调查变量的联合分布进行建模（Chow和Liu 1968）。这一模型族将联合分布分解为局部边缘分布与成对条件分布的乘积，并将其组织成一棵树，从而生成一个完全无监督的受访者典型性评分器。不专心作答无法遵循题目之间占主导地位的条件依赖关系，因此在学习得到的模型下似然较低，进而会被标记为异常。

设 $\mathbf{X} = (X_1、\ldots\ldots、X_d)$ 为离散变量，其中 $X_j \in \{1、\ldots\ldots、K_j\}$ 是问题 j 的类别（我们将“未作答”视为一个明确的类别）。树结构贝叶斯网络 T 由一个根变量 X_r 以及从各父节点指向子节点的有向边构成，形成包含 $d-1$ 条边的有向无环图。其联合分布可分解为

$$p_T(\mathbf{x}) = p(x_r) \prod_{j \neq r} p\left(x_j \mid x_{\pi(j)}\right),$$

其中，$\pi(j)$ 表示树中节点 j 的父节点。

给定数据 D $= \{\mathbf{x}^{(n)}\}_{n=1}^{N}$，Chow–Liu 算法通过树结构近似经验分布，并使似然最大。该算法分两步进行：(i) 估计所有变量对之间的成对互信息（MI）；(ii) 根据这些互信息权重，选取最大权生成树。具体而言，对于变量 X_i 和 X_j，我们构造平滑后的经验联合

$$\widehat{p}_{ij}(a,b) = \frac{N_{ij}(a,b)+\alpha}{N+\alpha K_i K_j},$$

$$\widehat{p}_i(a) = \sum_b \widehat{p}_{ij}(a,b),$$

$$\widehat{p}_j(b) = \sum_a \widehat{p}_{ij}(a,b).$$

采用拉普拉斯平滑 $\alpha > 0$（我们使用 α=1）。 $N_{ij}(a,\ b)$ 是$X_i = a,\ X_j = b$共同出现的次数。互信息（自然对数）为

$$\widehat{I}(X_i;X_j) = \sum_{a=1}^{K_i}\sum_{b=1}^{K_j}\widehat{p}_{ij}(a,b)\ \log\frac{\widehat{p}_{ij}(a,b)}{\widehat{p}_i(a)\,\widehat{p}_j(b)}.$$

设 G 为由 {1、.....、d} 构成的完全图，边权重为 $I(X_i;\ X_j)$。 G 的任意最大权生成树 $E^\star$ 都定义了无向结构；从选定的根节点 r出发确定其边的方向，即可得到有向树。实践中，我们选择 r 作为“互信息中心”节点（与其他节点的互信息总和最大），并使用标准的最大生成树算法（Prim 算法）求得该树。

给定有向树后，我们采用拉普拉斯平滑来估计根节点边缘分布和各条件概率表（CPT）。对于根节点:

$$\widehat{p}(x_r{=}k) = \frac{N_r(k)+\alpha}{N+\alpha K_r} \qquad (k=1,\dots,K_r).$$

对于每个子节点—父节点对（$c,\ p$）：

$$\widehat{p}(x_c{=}k_c \mid x_p{=}k_p) = \frac{N_{pc}(k_p,k_c)+\alpha}{N_p(k_p)+\alpha K_c},$$

$$k_p = 1,\dots,K_p,\ \ k_c = 1,\dots,K_c.$$

平滑可防止出现零计数，从而在稀疏、高基数场景下稳定互信息估计和似然。

对于受访者$\mathbf{x}$，在学习得到的树下，其（平滑）对数似然为

$$\log\widehat{p}_T(\mathbf{x}) = \log\widehat{p}(x_r) + \sum_{j\neq r}\log\widehat{p}\big(x_j \mid x_{\pi(j)}\big).$$

对于排序，我们依据总对数似然计算基于排名的典型性百分位数。设 N为受访者人数， $r(\mathbf{x}) \in \{1,\dots,N\}$ 为按对数 $\widehat{p}_T(\mathbf{x})$ 降序排列时 $\mathbf{x}$的排名（因此 r=1 最具典型性，其对数似然最高）。我们定义

$$\mathrm{pct}(\mathbf{x}) = 1 - \frac{r(\mathbf{x})-1}{N-1} \in [0,1],$$

因此，pct=1表示最具典型性的受访者，而 pct=0 表示最不具典型性的受访者。注意力不集中检测直接使用此统计量，按 pct升序排列受访者。换言之，这一排序标准等同于自编码器情形下的重构损失排序。

Chow–Liu 树明确编码了题目间的主要成对依赖关系（例如题组内部或逻辑相关问题之间的依赖关系）。专心受访者往往会遵循这些依赖关系，因此具有较高的似然。相比之下，注意力不集中的受访者会破坏学到的条件结构（例如做出不连贯或随机的选择），从而降低其似然。该方法运用了与基于重构的方法相同的直觉——“结构化模式可以压缩，噪声则不能”，但它采用概率视角，并能对局部依赖关系进行解释。

附录A中提供了一个示例，说明该网络如何运作。

5 实验结果

5.1 评估所用指标

5.1.1 重构评估

我们首先评估模型重构原始数据的准确程度。对于每个变量，准确率（即正确重构值所占的比例）计算如下：

$$\text{Accuracy}(v_i) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}(x_{j,i} = \hat{x}_{j,i}),$$

其中，$\mathbb{I}(\cdot)$ 是指示函数，$x_{j,\,i}$和 $\hat{}x_{j,\,i}$分别表示样本 j中变量 v_i的真实值和重构值。所有 $|\mathcal{V}|$ 个变量的平均重构准确率为

$$\text{Mean Accuracy} = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\text{Accuracy}(v_i).$$

为评估性能，我们将其与每个变量的多数类基线进行比较：

$$\text{Baseline Accuracy}(v_i) = \max_{c\in C_i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}(x_{j,i} = c),$$

其中，C_i 是分类值的集合。模型相较于基线的提升幅度通过 Lift 指标量化：

$$\text{Lift} = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\frac{\text{Accuracy}(v_i)}{\text{Baseline Accuracy}(v_i)},$$

其中，Lift > 1表示模型性能优于多数类预测。

为补充准确率，我们计算一对其余 ROC AUC 指标 ORA，以衡量模型将正确回答排在错误回答之前的能力。对于具有类别 C_i的变量 v_i，

$$\text{ORA}(v_i) = \frac{1}{|C_i|}\sum_{c\in C_i}\text{ROC AUC}(c),$$

总体得分为

$$\text{Mean ORA} = \frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\text{ORA}(v_i).$$

5.1.2 随机性检测评估

我们使用注意力检查结果作为真实标签，评估模型检测注意力不集中的受访者的能力。由于某些数据集包含多个检查项，因此分别计算每项检查的指标。

计算每位受访者的重构误差（Chow-Liu 模型则计算负对数似然）后，我们按降序排列受访者。这将注意力不集中检测转化为信息检索任务：目标是将粗心的受访者排在列表前列。

排序指标： 我们报告 **Recall@h**，其中 h 是数据集中注意力不集中的受访者的真实数量。该指标（在异常检测文献中也称为“成功率”）衡量的是，当我们将拒绝阈值恰好设为真实污染率时，检出的真实异常样本所占的比例：

$$\text{Recall@h} = \frac{|\{\text{inattentive users in top-}h\}|}{h}.$$

我们还报告固定操作截点 ($k \in \{10,\ 50,\ 100\}$) 下的 **Precision@k**，以模拟“人在环路中”的审查场景：研究人员检查排名前列的 k 个标记案例：

$$\text{Precision@k} = \frac{|\{\text{inattentive users in top-}k\}|}{k}.$$

注意，精确率@h =召回率@h。

为了评估注意力不集中的受访者在排序靠前位置的表现，我们计算归一化折损累积增益：

$$\text{DCG@h} = \sum_{i=1}^{h}\frac{\mathbb{I}(\text{inattentive at rank } i)}{\log_2(i+1)}$$

$$\text{NDCG@h} = \frac{\text{DCG@h}}{\text{IDCG@h}},$$

其中，IDCG@h 表示理想（完美）的排序。

判别指标： 最后，我们使用 ROC 曲线评估不依赖特定阈值的整体区分性能，并以重构误差作为决策阈值：

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad \text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}.$$

在各阈值下将真正率对假正率作图，可得到ROC曲线及其曲线下面积

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}$$

可量化排序性能，其中更高的曲线下面积值表明专心受访者与注意力不集中的受访者之间的区分更强。为确保我们对曲线下面积的估计能够稳健应对异常检测中提出的“在测试集上训练”问题，我们依据交叉验证检查所生成的样本外分数对这些数值进行了确认。因此，所报告的曲线下面积反映的是模型真正学习注意力集中行为与注意力不集中行为之间结构差异的能力，而非对特定受访者噪声的过拟合。

5.2 结果

5.2.1 重建性能

表4 展示了各种模型上各类方法在重构任务中的结果。我们纳入了非线性自编码器的两种变体：(a) 一种带有 $p = 100$的变体，按标准自编码器进行训练；(b) 一种带有 $p = 85$ 的百分位损失变体，其结果最为 promising。总体而言，三种无监督模型在各数据集上均实现了强劲的重构性能，并持续优于多数类基线预测器。准确率通常超过 80%，几乎所有情况下提升度都远高于 1.3，这表明这些模型无需监督即可成功捕捉调查作答模式中的规律。其中，线性自编码器在各数据集上的表现最稳定且具有竞争力，准确率和提升度往往达到最高（例如，Uhalt 2020，Mastroianni and Dana 2022）。这表明线性映射能够捕捉调查数据中存在的大部分协方差结构。非线性自编码器的表现与之相当，在某些情况下略胜一筹（例如，O’Gradyet al. 2019、Alvarezet al. 2019）。

百分位损失变体在设计上，重构得分略低于其他方法。由于它将学习重点放在误差较低的样本子集上，因此以整体重构准确度换取稳健性和选择性敏感度；这是注意力不集中检测中预期且理想的表现（尽管会损害纯粹的重构指标）。总体而言，这些发现表明，简单的线性变换是捕捉分类调查数据中主要结构的强基线，而当需要稳健性或对细微依赖关系进行建模时，非线性模型和基于百分位数的模型则更具优势。最后，我们注意到，贝叶斯网络仅用于第二项任务，即检测不专心参与者，因为它们无法经过修改来执行重构任务。

5.2.2 随机性检测性能

表5、6和7 展示了第二项任务——注意力不集中的受访者检测——的结果。在各个数据集中，所有模型在检测注意力不集中的受访者方面都取得了有意义的表现，这证实了无监督学习无需监督即可可靠地区分连贯与不连贯的回答模式（曲线下面积指标有明确的 0.5 基线）。总体而言，Chow–Liu 贝叶斯网络表现最为稳定且出色，在各种数据集和注意力检查配置中经常获得最高的曲线下面积和精确率得分。当依赖关系更加局部化且呈分类形式时，Chow–Liu 网络表现出色；它利用成对条件概率来检测作答一致性的细微偏离。这些互补的结果表明，概率模型能够有效捕捉调查数据的典型性。非线性自编码器在某些情况下也取得了有竞争力的召回率

表 4. 不同数据集和方法的重构结果

数据集与方法	准确率	基线准确率	Lift	ORA
Robinson-Cimpian (2014)				
非线性自编码器（$p = 100$）	81.37	58.27	1.53	**0.73**
非线性自编码器（$p = 85$）	80.83	58.27	1.52	0.70
线性自编码器	**83.16**	58.27	**1.58**	0.70
Pennycook 等人(2020)				
非线性自编码器（$p = 100$）	86.84	66.22	1.38	0.83
非线性自编码器（$p = 85$）	86.86	66.22	1.38	0.83
线性自编码器	**89.18**	66.22	**1.44**	**0.86**
Alvarez 等人（2020）				
非线性自编码器（$p = 100$）	**90.31**	52.56	2.05	**0.88**
非线性自编码器（$p = 85$）	89.69	52.56	2.02	0.86
线性自编码器	90.07	52.56	**2.07**	0.85
Uhalt (2020)				
非线性自编码器 ($p = 100$)	85.94	35.66	2.49	**0.89**
非线性自编码器 ($p = 85$)	81.31	35.66	2.35	0.83
线性自编码器	**91.44**	35.66	**2.64**	**0.89**
O'Grady 等人（2019年）				
非线性自编码器 ($p = 100$)	**86.89**	58.78	**1.70**	**0.89**
非线性自编码器 ($p = 85$)	84.30	58.78	1.64	0.85
线性自编码器	79.80	58.78	1.68	0.86
Buchanan和Scofield （2018年）				
非线性自编码器（$p = 100$）	**94.62**	50.87	2.18	**0.90**
非线性自编码器（$p = 85$）	91.89	50.87	2.12	0.75
线性自编码器	92.80	50.87	**2.19**	0.84
Moss 等人（2023年）				
非线性自编码器 ($p = 100$)	83.46	59.78	1.51	0.76
非线性自编码器 ($p = 85$)	82.73	59.78	1.49	0.75
线性自编码器	**88.59**	59.78	**1.66**	**0.82**
Mastroianni 和 Dana(2022)				
非线性自编码器 ($p = 100$)	77.77	61.64	1.26	0.68
非线性自编码器 ($p = 85$)	77.57	61.64	1.25	0.68
线性自编码器	**83.11**	61.64	**1.34**	**0.72**
Ivanov等人 (2021)				
非线性自编码器（$p = 100$）	76.15	45.32	1.86	0.78
非线性自编码器（$p = 85$）	74.69	45.32	1.82	0.76
线性自编码器	**80.24**	45.32	**2.00**	**0.84**

在尝试"预测"样本外属性值时，对模型在各数据集上的性能进行评估。Acc 指模型重构原始数据的平均准确率。基线 Acc（基线准确率）表示在每个属性上都"猜测"多数类作为取值时的准确率。提升度的计算方式为准确率与基线准确率之比，用于表示我们的模型相较于朴素18预测器的提升。ORA 表示一对多 ROC AUC 指标，其中 0.5 为基线性能。

数据集（例如，Alvarez等人（2019年）和Ivanov等人（2021）。

PL 自编码器始终能提供均衡的结果，其精确率和曲线下面积通常优于基础自编码器，尤其是在注意力检查标签噪声较大或内容异质的数据集中。这印证了一种直觉：在训练过程中降低高误差样本的权重，会使模型较少受离群值影响，并在评估时对注意力不集中更加敏感。相比之下，线性自编码器的表现通常较差，尤其是在较小或高度离散的数据集中，这凸显了纯线性映射在注意力不集中检测方面的局限性，尽管它们在重构方面仍然表现出色。各种方法整体表现一致，表明无监督建模可作为一种强大且不依赖特定领域的方法，用于识别行为数据中注意力不集中的受访者。

6 分析

6.1 无监督方法何时表现良好?

所有无监督模型都展现出检测注意力不集中的受访者的显著能力，但其表现会因数据集和指标而异。了解这种差异的来源，对于评估模型的稳健性以及指导其在行为研究中的实际应用至关重要。因此，我们将探讨这些方法在何种情况下以及为何能或不能有效识别注意力不集中。

我们首先进行数据集层面分析，检验检测效果的变异是否与数据集特征相关：样本量（样本）、变量数量（变量）、独热特征总数（特征）以及每个变量的平均特征数（*AFV*）。对于随机性检测任务，我们用各方法的平均 AUC 汇总每个数据集。对于重构任务，我们使用平均提升度。我们计算皮尔逊 r 和斯皮尔曼 ρ 相关系数（双侧检验； $n = 9$ 个数据集），分别衡量线性关联和单调关联。

表8 显示，数据集规模和维度（样本、变量、特征、*AFV*）与检测性能之间不存在可靠关联（皮尔逊 $|r| \leq 0。54$；斯皮尔曼 $|\rho| \leq 0。48$；所有 $p>0。18$）。相比之下，平均重构提升度与平均 AUC 呈中等相关（$r\approx0。69$， $p\approx0。04$； $\rho\approx0。68$），这表明，当模型更好地捕捉结构规律时，注意力不集中状态会更容易区分。[1]

在数据集层面，这些发现表明，规模和稀疏性并非决定检测成效的主要因素。相反，依赖结构得到良好捕捉（即重构提升度较高）的调查问卷，能更清晰地区分注意力集中的受访者和注意力不集中的受访者。实际上，这意味着，决定无监督模型检测注意力不集中的受访者的效果如何的，是调查结构所包含的信息量，而非其规模。

无监督模型区分注意力集中与注意力不集中的受访者，主要取决于两个因素：调查问卷的内部结构，以及注意力检查标签的质量。像Uhalt (2020)和Alvarez等人(2019)这样的数据集代表了最佳情形；它们采用结构良好的多题项量表，其中多个问题测量同一潜在构念，而注意力检测题则直接检验指令遵从情况。这种组合能产生一致的回答模式，使模型易于学习。相比之下，像Mastroianni和Dana(2022)以及Buchanan和Scofield (2018)这样的数据集混合了态度、行为和基于元数据的变量，削弱了跨题项规律。非线性模型仍能识别结构，但是

[1]估计值 基于表格中的九个数据集1。方法特定相关性见附录B .

表 5. 各数据集与方法的随机性检测结果（第一部分）

数据集与方法	h	R@h	P@10	P@50	P@100	归一化折损累计增益@h	曲线下面积
Robinson-Cimpian（2014）							
非线性自编码器（$p = 100$）	230	0.07	**0.20**	0.08	**0.09**	0.08	0.71
非线性自编码器（$p = 85$）	230	**0.10**	**0.20**	**0.12**	**0.09**	**0.09**	0.74
线性自编码器	230	0.03	0.00	0.04	0.03	0.03	0.51
Chow-Liu	230	0.08	0.00	0.08	**0.09**	0.08	**0.75**
Pennycook 等（2020）							
注意力1							
非线性自编码器（$p = 100$）	636	0.73	**1.00**	0.88	0.84	0.75	0.52
非线性自编码器（$p = 85$）	636	0.74	**1.00**	0.86	0.87	0.76	0.52
线性自编码器	636	**0.75**	0.70	0.84	0.76	0.75	**0.53**
Chow-Liu	636	**0.75**	**1.00**	**0.94**	**0.91**	**0.77**	**0.53**
注意力 2							
非线性自编码器（$p = 100$）	236	0.30	0.50	0.30	0.34	0.32	0.52
非线性自编码器（$p = 85$）	236	0.31	**0.60**	0.44	0.37	0.35	**0.53**
线性自编码器	236	0.31	0.20	0.36	0.33	0.31	**0.53**
Chow-Liu	236	**0.33**	0.50	**0.50**	**0.39**	**0.36**	**0.53**
注意力 3							
非线性自编码器（$p = 100$）	68	0.25	0.30	0.26	0.21	0.24	0.59
非线性自编码器（$p = 85$）	68	0.25	**0.40**	0.28	0.22	0.30	0.62
线性自编码器	68	0.08	0.00	0.06	0.09	0.06	0.53
Chow-Liu	68	**0.31**	0.30	**0.32**	**0.26**	**0.33**	**0.67**
注意力 4							
非线性自编码器（$p = 100$）	172	0.24	0.40	0.36	0.29	0.27	0.50
非线性自编码器（$p = 85$）	172	0.26	**0.70**	0.40	0.36	0.32	0.51
线性自编码器	172	0.20	0.40	0.24	0.20	0.21	0.50
Chow-Liu	172	**0.28**	0.50	**0.48**	**0.38**	**0.33**	**0.52**
并集							
非线性自编码器（$p = 100$）	686	0.79	**1.00**	0.96	0.91	0.81	0.53
非线性自编码器（$p = 85$）	686	0.80	**1.00**	0.92	0.91	0.81	0.53
线性自编码器	686	**0.81**	0.80	0.94	0.86	**0.82**	**0.55**
Chow-Liu	686	0.80	**1.00**	**0.98**	**0.95**	**0.82**	0.53
交集							
非线性自编码器（$p = 100$）	28	0.14	0.10	**0.12**	0.07	0.12	0.56
非线性自编码器（$p = 85$）	28	0.14	**0.30**	**0.12**	0.11	**0.24**	0.61
线性自编码器	28	0.04	0.00	0.04	0.03	0.03	0.50
Chow-Liu	28	**0.18**	0.20	**0.12**	**0.13**	0.23	**0.64**
Alvarez 等（2019年）							
非线性自编码器（$p = 100$）	975	**0.69**	**1.00**	**1.00**	**1.00**	0.73	**0.80**
非线性自编码器（$p = 85$）	975	0.68	**1.00**	**1.00**	0.99	0.72	0.77
线性自编码器	975	0.43	0.30	0.30	0.39	0.43	0.58
Chow-Liu	975	**0.69**	**1.00**	**1.00**	**1.00**	**0.74**	**0.80**

跨数据集的注意力不集中检测评估。h 表示被识别为真实标签的注意力不集中用户数量。R@h 表示 Recall@h。在完美排序下，所有注意力不集中用户的误差都高于所有注意力集中的用户，R@h = 1。P@k 表示 Precision@k。我们对注意力检测题较多的数据集单独进行了评估。并集表示，只有当样本未通过某项注意力检测题时，我们才将其视为注意力不集中。交集表示，只有当样本未通过所有注意力检测题时，我们才将其视为注意力不集中。

表 6. 不同数据集和方法的随机性检测结果（第二部分）

数据集与方法	h	R@h	P@10	P@50	P@100	归一化折损累计增益@h	曲线下面积
Uhalt （2020）							
非线性自编码器 ($p = 100$)	6	0.17	0.10	0.04	**0.05**	0.15	0.67
非线性自编码器 ($p = 85$)	6	0.33	**0.30**	**0.10**	**0.05**	0.34	0.82
线性自编码器	6	0.00	0.00	0.02	0.03	0.00	0.60
Chow-Liu	6	**0.50**	**0.30**	**0.10**	**0.05**	**0.39**	**0.88**
O'Grady等人(2019年)							
非线性自编码器（$p = 100$）	20	0.10	0.00	0.04	0.04	0.07	0.55
非线性自编码器（$p = 85$）	20	**0.20**	**0.10**	0.08	0.06	**0.15**	0.55
线性自编码器	20	0.05	**0.10**	0.08	0.10	0.05	0.63
Chow-Liu	20	0.15	**0.10**	**0.16**	**0.13**	0.14	**0.76**
Buchanan和Scofield(2018年)							
非线性自编码器 ($p = 100$)	59	0.07	0.10	0.08	0.08	0.08	0.61
非线性自编码器 ($p = 85$)	59	**0.27**	**0.80**	**0.28**	**0.21**	**0.34**	**0.72**
线性自编码器	59	0.08	0.10	0.06	0.07	0.07	0.51
Chow-Liu	59	0.22	0.10	0.20	0.19	0.19	0.65
Moss 等人(2023年)							
注意力1							
非线性自编码器 ($p = 100$)	161	0.16	0.10	0.16	0.14	0.15	0.53
非线性自编码器 ($p = 85$)	161	**0.18**	**0.20**	0.20	**0.17**	**0.20**	**0.59**
线性自编码器	161	0.11	**0.20**	0.18	0.11	0.13	0.56
Chow-Liu	161	0.12	**0.20**	**0.22**	0.13	0.16	0.58
注意力 2							
非线性自编码器 ($p = 100$)	140	0.11	**0.30**	0.20	0.13	0.12	0.54
非线性自编码器 ($p = 85$)	140	**0.21**	**0.30**	**0.24**	**0.22**	**0.24**	**0.60**
线性自编码器	140	0.06	0.20	0.08	0.06	0.08	0.54
Chow-Liu	140	0.11	0.20	0.14	0.12	0.15	0.59
并集							
非线性自编码器（$p = 100$）	248	0.16	**0.30**	0.28	0.22	0.17	0.54
非线性自编码器（$p = 85$）	248	0.20	**0.30**	**0.30**	**0.25**	**0.23**	**0.59**
线性自编码器	248	0.16	**0.30**	0.20	0.13	0.17	0.53
Chow-Liu	248	**0.21**	**0.30**	0.28	0.20	**0.23**	0.57
交集							
非线性自编码器（$p = 100$）	53	0.08	0.00	0.08	0.06	0.07	0.56
非线性自编码器（$p = 85$）	53	**0.11**	**0.20**	**0.10**	**0.13**	**0.13**	**0.65**
线性自编码器	53	0.06	0.10	0.06	0.04	0.08	0.62
Chow-Liu	53	0.08	0.10	0.08	0.05	**0.13**	0.61

跨数据集的注意力不集中检测评估。h 表示被识别为真实标签的注意力不集中用户数量。R@h 表示 Recall@h。在排序完美的情况下，即所有注意力不集中用户的误差都高于所有注意力集中的用户，R@h = 1。P@k 表示 Precision@k。我们还单独评估了包含更多注意力检测题的数据集。并集表示只有当样本未通过任意一道注意力检测题时，才将其视为注意力不集中。交集表示只有当样本未通过所有注意力检测题时，才将其视为注意力不集中。

表 7. 不同数据集和方法上的随机性检测结果（第二部分 I)

数据集 & 方法	h	R@h	P@10	P@50	P@100	NDCG@h	曲线下面积
Mastroianni 和 Dana (2022)							
非线性自编码器 ($p = 100$)	60	**0.50**	**1.00**	**0.56**	0.30	**0.62**	0.66
非线性自编码器 ($p = 85$)	60	0.43	**1.00**	0.52	**0.31**	0.57	**0.68**
线性自编码器	60	0.05	0.00	0.04	0.04	0.04	0.52
Chow-Liu	60	0.45	**1.00**	0.54	0.28	0.58	0.61
Ivanov等人（2021）							
注意力1							
非线性自编码器（$p = 100$）	165	0.32	0.70	0.42	0.37	0.37	0.64
非线性自编码器（$p = 85$）	165	0.39	**0.90**	**0.60**	0.49	0.46	0.69
线性自编码器	165	0.22	0.40	0.28	0.26	0.23	0.51
Chow-Liu	165	**0.46**	**0.90**	0.58	**0.53**	**0.51**	**0.73**
注意力 2							
非线性自编码器 ($p = 100$)	55	0.09	0.20	0.10	0.13	0.15	0.61
非线性自编码器 ($p = 85$)	55	**0.22**	**0.30**	**0.20**	0.21	**0.20**	0.69
线性自编码器	55	0.09	**0.30**	0.10	0.07	0.13	0.51
Chow-Liu	55	0.18	0.10	0.16	**0.23**	0.16	**0.76**
并集							
非线性自编码器 ($p = 100$)	173	0.35	0.70	0.46	0.41	0.40	0.65
非线性自编码器 ($p = 85$)	173	0.42	**0.90**	**0.64**	0.53	0.48	0.70
线性自编码器	173	0.22	0.40	0.30	0.27	0.24	0.50
Chow-Liu	173	**0.48**	**0.90**	0.62	**0.56**	**0.53**	**0.74**
交集							
非线性自编码器 ($p = 100$)	47	0.06	0.20	0.06	0.09	0.13	0.59
非线性自编码器 ($p = 85$)	47	**0.15**	**0.30**	**0.16**	0.17	**0.15**	0.67
线性自编码器	47	0.09	**0.30**	0.08	0.06	0.12	0.52
Chow-Liu	47	0.09	0.10	0.12	**0.20**	0.08	**0.76**

跨数据集的注意力不集中检测评估。h 表示被识别为真实标签的注意力不集中用户数量。R@h 表示 Recall@h。在完美排序的情况下，即所有注意力不集中用户的误差都高于所有注意力集中用户时，R@h = 1。P@k 表示 Precision@k。对于注意力检测题更多的数据集，我们单独进行了评估。并集表示，只有当样本未通过任意一道注意力检测题时，我们才将其视为注意力不集中。交集表示，只有当样本未通过所有注意力检测题时，我们才将其视为随机样本。

表 8. 数据集特征与性能之间的相关性

	平均 AUC		平均提升度	
预测变量	皮尔逊 r	斯皮尔曼 ρ	皮尔逊 r	斯皮尔曼 ρ
样本	0.20	0.05	−0.24	−0。20
变量	−0.46	−0。16	−0。47	−0.46
特征	−0。37	−0。21	−0。54	−0。48
AFV	0.11	0.06	0.10	0.15
平均提升	0.69*	0。68*		

跨越 n=9 个数据集的皮尔逊和斯皮尔曼相关（双侧）。平均提升度 ↔ 平均 AUC 显示出中等程度的正相关（p≈0.04），而其他预测变量与重构和随机性表现均无可靠关系。

这些模式在概念上不够连贯，降低了注意力集中的受访者与注意力不集中的受访者之间的可区分性。

Pennycook 等人 的研究（2020）和 Moss 等人的研究（2023年）进一步说明，检测性能在很大程度上取决于标注精度。两项研究都包含重复的主题条目，理论上应能揭示注意力集中的受访者所呈现的清晰模式，但两者的注意力检测题在诊断性方面有所不同。自我报告和理解检测题产生的标签嘈杂且异质，而严格的指令遵循题则能提供清晰得多的真实标签。当注意力不集中被宽泛地定义（例如未通过任意一项检查）时，模型性能看起来较弱；当仅采用诊断性最强或“全部未通过”的标准时，区分效果会显著改善（例如，Pennycook 等人 （2020）中的“注意力 3”达到曲线下面积 = 0.67）。

总体而言，这些发现表明，连贯的调查设计为模型提供了有意义的结构供其学习，但只有精确的标注才能充分发挥这种潜力。如何界定注意力不集中仍是一项挑战，尤其是注意力检测题本身也并非完美；这既是我们评估的一项局限，也是更广泛的回答质量研究所面临的局限。

6.2 重建—检测权衡

标准自编码器通过最小化平均重构损失来学习重现所有输入。然而，对于注意力不集中检测，目标是对典型流形建模，同时允许不连贯回答保持较差的重构效果。PL 目标函数通过对每个批次中误差最低的 p% 样本取平均来实现这一点：较小的 p 值会在训练期间降低高损失样本的权重，使其在测试时仍表现为异常。这形成了一种权衡：较高的 p 能提升重建性能，而较低的 p 则能提高异常敏感度。

在各数据集上，我们在图p=100中展示相对于全批次目标函数（3–4）的变化。随着 p 增加， Δ提升度单调上升并趋近于零（重构效果更好），而 Δ曲线下面积呈倒 U 形，在 p=80–90 时有所提升，到 p=95时则减弱。数据集内配对检验证实了这一模式：在 p=85–90（Δ≈+0.04, p≈0.001）时，检测曲线下面积提升最为稳定；在 p=80 时提升较小，到 p=95时则没有提升。相比之下，随着 $p \to 100$ 增加，重建性能稳步提高（ Δ≈−0.07， p=80）。

总体而言，随着 p 增加，重构效果单调提升，而检测效果在

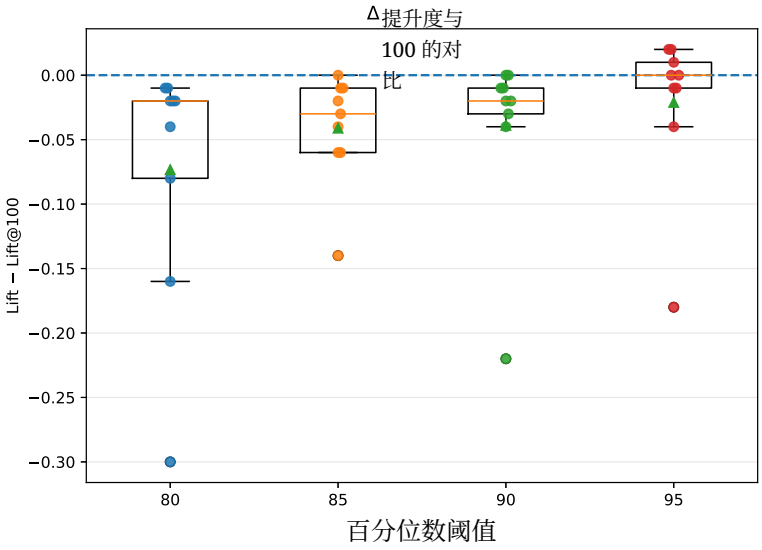


图3：百分位损失的权衡： Δ重构提升度与 p 相对于 p=100。在 p=80/85/90处，中位数为负；在 p=95处则趋近于0，这表明随着 $p \to 100$增加，重构效果单调改善。

p≈85–90 附近达到较高水平。在这些取值下，模型保留了足够多的典型案例来学习主要流形，同时让高损失（不连贯）的回答保持差异性。随着 $p \to 100$，对所有输入的过拟合会削弱异常对比度；如果 p 过低，模型就会过度关注琐碎模式，从而降低泛化能力。这可以概括为重构与随机性检测权衡：提升模型重构所有输入的能力，不可避免地会削弱其检测非典型输入的能力。

7 实际启示：自动化质量控制的经济学

我们的研究结果表明，无监督模型能够可靠地检测注意力不集中，尤其是在结构良好的调查问卷中。然而，对于调查平台和研究人员而言，是否采用这一框架，并非仅由检测准确率（曲线下面积）决定，还取决于数据质量、受访者负担与运营成本之间的经济权衡。本节将这一决策问题形式化，并推导无监督筛查优于传统注意力检测题的适用条件。

7.1 筛查的成本效益模型

我们将筛查流程建模为一个成本最小化问题。设 N 为受访者总数。传统的质量控制依赖于嵌入式注意力检测题（AC），而我们提出的方法则依赖于无监督建模（UM）。每种方法的总成本函数 L 可分解为负担成本 （施加于受访者的成本）和错误成本（保留噪声或丢弃有效数据所产生的成本）。

1. 注意力检测题的成本（L_{AC}）。注意力检测题会带来认知负担，并增加

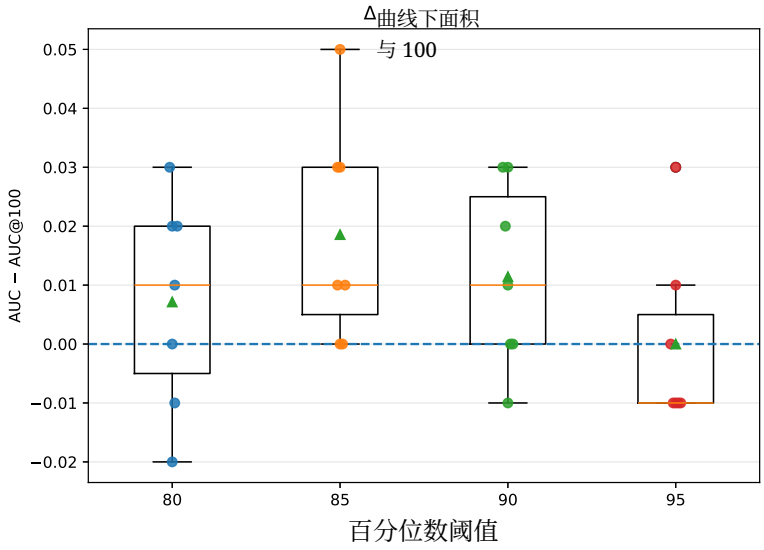


图4：百分位损失权衡： Δ曲线下面积（随机性检测）与相对于 p=100 的 p 。在 p=80/85/90 处有所提升，而在 p=95 处有所减弱，表明最优点大致位于 $p \approx$ 85–90 附近。

调查时长。设 c_{tax} 为每位受访者纳入检查项的边际成本（以时间折算的成本、增加的脱落风险或测量反应性）。如果将检查项视为“真实标签”，我们便假设其错误率可以忽略不计，但这项负担会施加于所有 N 受访者：

$$L_{AC} = N \cdot c_{tax} \tag{1}$$

2. 无监督建模的成本（L_{UM}）。我们的无标签方法不会增加受访者负担（$c_{tax} = 0$）。然而，由于它是一种概率代理指标，因此会引入分类错误。令 α 表示注意力不集中的受访者的患病率。令 FNR 表示假阴性率（漏掉的注意力不集中的受访者），令 FPR 表示假阳性率（被标记的注意力集中的受访者）。令 c_{noise} 表示将注意力不集中的受访者纳入最终数据集的成本（统计效力降低，可能产生伪相关）。令 $c_{discard}$ 表示错误排除有效受访者的成本（浪费招募费用、样本量减少）。无监督方法的总成本为：

$$L_{UM} = C_{compute} + N \cdot C_{model} \tag{2}$$

$$C_{model} = \alpha \cdot FNR \cdot c_{noise} + (1 - \alpha) \cdot FPR \cdot c_{discard} \tag{3}$$

其中， $C_{compute}$ 是训练模型的固定成本，对于现代硬件而言，这一成本可以忽略不计（每位受访者 ≈ 0 ）。

3. 决策规则。当 $L_{UM} < L_{AC}$时，平台应从注意力检测题切换到无监督建模。假设 $C_{compute} \approx 0$，则条件变为：

$$\alpha \cdot FNR \cdot c_{noise} + (1 - \alpha) \cdot FPR \cdot c_{discard} < c_{tax} \tag{4}$$

这一不等式揭示了我们框架的确切价值主张。在受访者时间成本高昂的场景中（例如 c_{tax} 较高的专家调查），或“测量反应性”威胁效度时，只要错误率（FNR， FPR）保持在有界范围内，无监督方法在经济性上更具优势，即使其准确性低于显式检查项。相反，在低风险的众包场景中，当 c_{tax} 较低但数据完整性（c_{noise}）至关重要时，混合方法可能是最优选择。

7.2 阈值的操作化

一个关键的管理决策是设定检测器的灵敏度。我们对百分位损失（PL）超参数 p （第 6.2 节）的分析，为这一权衡提供了直接的调节手段。

- **尽量降低噪声**（$c_{noise} \gg c_{discard}$）：如果不良数据的代价很高（例如医疗或政策研究），实践者应以我们所识别的“稳健区域”为目标（$p \in [85, 90]$）。如图4所示，这一区间能最大限度地拉开典型流形与异常之间的距离，从而有效降低 FNR，代价是重构误差略有增加。

- **尽量减少剔除**（$c_{discard} \gg c_{noise}$）：如果招募受访者困难（例如稀有人群），研究人员应采取保守的排除策略。我们不建议采用固定阈值，而是建议按照重构误差对受访者进行排序，并检查误差分布的“肘部”。

7.3 将设计作为治理

对于信息系统领域而言，我们最具实践指导意义的发现或许是，检测是设计的内生结果。我们表明，题项组连贯且彼此重叠的调查问卷（AFV 或平均提升度较高）有助于显著提升区分效果。这意味着调查设计是一种治理机制。寻求自动化质量控制的平台不必只依赖更优的算法；它们也可以强制执行设计指南，具体而言，要求进行有效的量表构建并包含冗余指标。这样可以提高响应数据的“信噪比”，在不添加显式检查题项的情况下，有效降低公式（3）中的 FNR 和 FPR。

7.4 合乎伦理的部署：人在回路

最后，我们提醒，不应仅凭无监督评分就完全自动化地排除受访者。如第4.3节所述，自编码器会标记“缺乏一致性”的情况，而这有时可能与合理但属于少数的观点重叠。为降低公平性风险（$c_{discard}$），我们建议采用两阶段治理流程：

1. **自动标记：** 使用重构误差得分对受访者进行排序，并将排名前 q%的受访者标记为“高风险”。 2. **人工审核：** 将这些高风险档案呈现给人工审查员。可视化的重构误差（如图1所示）可作为可解释界面，准确突出显示哪些项目偏离了预期模式。

这种“增强智能”方法兼顾了无监督学习的可扩展性与人类判断的细致入微，确保数据质量工作不会无意中强化同质性。

8 结论

本文提出一种无需标签且不依赖特定调查问卷的注意力不集中检测框架，结合了互补的建模视角：几何重构（自编码器）和概率依赖建模（Chow–Liu）。在九个异质的真实世界数据集上，我们表明

对跨题项规律进行建模，无需监督即可发现回答不一致的受访者，从而形成实用的两阶段工作流程：先进行无监督建模，再通过重构误差或（负）似然对受访者排序。在方法上，我们将自编码器应用于分类调查数据，并采用变量级损失加权，还提出了百分位损失，揭示了重构与检测之间明确的权衡，并指出在 $p \in [85, 90]$ 处存在适用于注意力不集中检测的稳健工作区间。在实质层面，我们表明，题项组连贯且相互重叠的调查工具能够产生稳定的协方差和依赖模式，供我们的模型利用，从而更有效地区分专心受访者与注意力不集中的受访者，并将心理测量设计选择与无监督异常检测性能直接关联起来。总体而言，这些结果将注意力不集中筛查从临时规则和被动检查转变为一种可扩展、透明的诊断工具，能够推广至不同的调查问卷、语言和人群。

9 局限性与未来研究

评估标签。 我们的检测性能是根据注意力检查项的结果进行评估的。这些结果是已知的注意力不集中代理指标，且可能存在噪声。尽管我们的研究结果在多个数据集和检查项中均表现稳健，未来研究仍应纳入其他真实标签（例如眼动追踪不专注标签），以便更精确地量化对标签噪声的敏感性。

文本限制与模态覆盖范围。 我们仅关注结构化类别变量和离散化数值回答，并有意排除开放式文本字段，以确保不同数据集之间具有可比性。这一设计选择限制了研究结果向包含大量自由文本内容的调查工具的泛化。一个自然的扩展方向是将文本嵌入（例如句子级表示）整合到我们的重构和似然流程管线中，从而实现对混合类型调查数据的统一多模态筛查。

建模范围。 我们评估了线性／非线性自编码器（采用百分位损失）以及 Chow–Liu 贝叶斯网络。其他概率模型族（例如结合作答时间的混合 IRT 模型、因子混合模型）和现代表格生成模型（例如流模型、类别空间上的扩散模型）或许能进一步提升典型性评分。

对科学结论的下游影响。归根结底，我们关心的是，剔除注意力不集中的用户是否会影响使用调查结果的研究结论。为了分析检测注意力不集中的用户是否有用，我们需要复现全部九项研究的结果，并考察各论文的结果如何随注意力不集中者剔除程度的不同而变化。这一过程类似于（Cimpian 等人，2018）的研究，该研究探讨了剔除恶作剧型受访者后，人口普查数据的估计值会如何变化。不过我们注意到，这类分析需要复现九篇论文的研究工作，是一项非常艰巨的任务，因此将其留待未来研究。

参考文献

Herv´e Abdi 和 Lynne J Williams。主成分分析。《Wiley 跨学科评论：计算统计》，2(4):433–459，2010。Andreas Alfons 和 Max Welz。关于运用机器学习识别粗心作答的开放科学视角：新的希望还是虚幻威胁？Social and Personality Psychology Compass，18(2):e12941，2024。R Michael Alvarez、Lonna Rae Atkeson、Ines Levin 和 Yimeng Li。关注注意力不集中的调查受访者。Political Analysis，27(2):145–162，2019。Jinwon An 和 Sungzoon Cho。基于变分自编码器的异常检测：利用重构概率。Special Lecture on IE，2(1):1–18，2015。Adam J Berinsky、Michele F Margolis 和 Michael W Sances。区分敷衍者与认真作答者？确保受访者在自填式调查中认真作答。American Journal of Political Science，58(3):739–753，2014。Erin M Buchanan 和 John E Scofield。检测低质量数据的方法及其对心理学研究的影响。《行为研究方法》，50(6):2586–2596，2018。CKCN Chow 和 Cong Liu。用依赖树近似离散概率分布。IEEE Transactions on Information Theory，14(3):462–467，1968。Joseph R Cimpian、Jennifer D Timmer、Michelle A Birkett、Rachel L Marro、Blair C Turner 和 Gregory L Phillips。大规模估计中，潜在恶作剧型受访者造成的偏差：女同性恋、男同性恋、双性恋或疑问期（LGBQ）青少年与异性恋青少年之间的健康差异。American Journal of Public Health，108（S4）：S258–S265，2018。Lee Anna Clark 和 David Watson。构建效度：客观量表开发中的基本问题。Psychological Assessment，7(3):309–319，1995。doi：10.1037/1040-3590.7.3.309。Lee Anna Clark 和 David Watson。构建效度：创建客观测量工具的新进展。Psychological Assessment，31(12):1412，2019。Lee J Cronbach。α系数与测验内部结构。Psychometrika，16(3):297–334，1951。Paul G Curran。检测调查数据中粗心的无效回答的方法。Journal of Experimental Social Psychology，66:4–19，2016。Alexander Philip Dawid 和 Allan M Skene。使用期望最大化算法对观察者错误率进行最大似然估计。Journal of the Royal Statistical Society: Series C (Applied Statistics)，28(1):20–28，1979。Thomas F Donlon 和 Frederic E Fischer。衡量个体对群体确定的题目难度认同程度的指标。Educational and Psychological Measurement，28(1):105–113，1968。Fritz Drasgow、Michael V Levine 和 Esther A Williams。采用多分测验反应模型和标准化指标进行适切性测量。British Journal of Mathematical and Statistical Psychology，38(1):67–86，1985。Simao Eduardo、Alfredo Nazabal、Christopher K. I. Williams 和 Charles Sutton。用于混合类型数据离群值检测与修复的稳健变分自编码器。载于 Silvia Chiappa 和 Roberto Calandra 编，《第二十三届人工智能与统计国际会议论文集》，第108卷《机器学习研究论文集》，第4056–4066页。PMLR，2020年8月26–28日。URL https://proceedings.mlr.press/v108/eduardo20a.html。Philippe Goldammer、Hubert Annen、Peter Lucas St¨ockli 和 Klaus Jonas。问卷测量中的粗心作答：检测、影响与补救措施。The Leadership Quarterly，31(4):101384，2020。

Adam Goodge、Bryan Hooi、See Kiong Ng 和 Wee Siong Ng。对抗性影响下用于异常检测的自编码器的鲁棒性。载于第二十九届国际人工智能联合会议国际会议论文集，第 1244–1250 页，2021 年。Joseph K Goodman、Cynthia E Cryder 和 Amar Cheema。平坦世界中的数据收集：Mechanical Turk 样本的优势与不足。《行为决策》，26(3):213–224，2013 年。Simon Hawkins、Hongxing He、Graham Williams 和 Rohan Baxter。使用复制器神经网络进行离群值检测。载于国际数据仓储与知识发现会议，第 170–180 页。施普林格，2002 年。Maxwell Hong、Jeffrey T Steedle 和 Ying Cheng。检测投入不足作答的方法：比较与实践建议。《教育与心理测量》，80(2): 312–345，2020。Stefan Ivanov、Meghan A Novisky 和 Matt Vogel。种族怨恨、共情与对新冠肺炎疫情期间释放囚犯的支持：一项调查实验的结果。*Socius*，7:23780231211005222，2021 年。Yujin Kim、Jennifer Dykema、John Stevenson、Penny Black 和 D Paul Moberg。直线作答：测量概述、指标比较及其对邮件-网络混合模式调查的影响。《社会科学计算机评论》，37(2):214–233，2019年。Diederik P Kingma 和 Max Welling。自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013 年。Anders Krogh 和 John Hertz。简单的权重衰减可以提升泛化能力。神经信息处理系统进展，4，1991 年。Mads Lindskou、Torben Tvedebrink、Poul Svante Eriksen 和 Niels Morling。使用条件高斯回归模型检测混合变量类型的高维数据中的离群值。*arXiv* 预印本 *arXiv:2103.02366*，2021 年。Adam M Mastroianni 和 Jason Dana。对长期态度变化的普遍误解。《美国国家科学院院刊》，119(11):e2107260119，2022 年。Adam W Meade 和 S Bartholomew Craig。识别调查数据中的粗心作答。《心理学方法》，17(3):437，2012 年。Nicholas Merrill 和 Colin C Olson。一种用于无监督高光谱异常检测的新型自编码器训练范式。载于*IGARSS 2020-2020 IEEE*国际地球科学与遥感研讨会，第 3967–3970 页。IEEE，2020 年。Aaron J Moss、Cheskie Rosenzweig、Jonathan Robinson、Shalom N Jaffe 和 Leib Litman。将 Mechanical Turk 用于行为研究是否合乎伦理？来自 Mechanical Turk 参与者代表性调查及其工资的相关数据。《行为研究方法》，55(8):4048–4067，2023年。David S Nichols、Roger L Greene 和 Peter Schmolck。评估明尼苏达多相人格调查表（MMPI）条目认可不一致模式的标准：理论依据、制定过程与实证检验。《临床心理学杂志》，45(2):239–250，1989 年。Trevor O'Grady、Donald Vandegrift、Michael Wolek 和 Gregory Burr。关于他人取向行为的决定因素：道德基础问卷的现场测试。《人格研究期刊》，81:224–237，2019年。Koken Ozaki。利用机器学习检测注意力不集中的受访者：一种可替代定向问题量表并弥补其不足的通用技术。《行为研究方法》，56 (7):7059–7078，2024年。Eyal Peer、David Rothschild、Andrew Gordon、Zak Evernden 和 Ekaterina Damer。在线行为研究平台和样本库的数据质量。《行为研究方法》，54(4):1643–1662，2022年。

Gordon Pennycook、Jonathon McPhetres、Yunhao Zhang、Jackson G Lu 和 David G Rand。在社交媒体上打击新冠肺炎错误信息：可扩展准确性助推干预的实验证据。《心理科学》，31(7):770–780，2020。Joseph P Robinson-Cimpian。恶作剧型受访者导致的差异估计不准确：评估结论的若干建议。《教育研究者》，43(4):171–185，2014。Zachary Joseph Roman、Holger Brandt 和 Jason Michael Miller。使用贝叶斯潜在类别模型自动检测在线调查中的机器人。《心理学前沿》，13:789223，2022。Jeniffer Rubio、Paolo Barucca、Gerardo Gage、John Arroyo 和 Ra´ul Morales-Resendiz。利用人工神经网络对支付模式进行分类：一种自编码器方法。《拉丁美洲中央银行学刊》，1(1-4):100013，2020。Ulrich Schroeders、Christoph Schmidt 和 Timo Gnambs。使用随机梯度提升检测调查数据中的粗心作答。《教育与心理测量》，82(1):29–56，2022。Tom AB Snijders。估计被试参数时被试拟合统计量的渐近零分布。Psychometrika，66(3):331–342，2001。Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。Dropout：一种防止神经网络过拟合的简单方法。《机器学习研究期刊》，15 (1):1929–1958，2014。J. Uhalt。调查数据中的注意力检查与作答质量。OSF，2020。URL https://osf.io/uz872/.[数据集]。Esther Ulitzsch、Steffen Nestler、Oliver L¨udtke 和 Gabriel Nagy。一种基于屏幕使用时间的混合模型，用于在生态瞬时评估数据中识别和监测粗心作答与投入不足作答。《心理学方法》，2024年。Pascal Vincent、Hugo Larochelle、Isabelle Lajoie、Yoshua Bengio、Pierre-Antoine Manzagol 和 L´eon Bottou。堆叠去噪自编码器：在深度网络中利用局部去噪准则学习有用的表示。《机器学习研究期刊》，11(12)，2010年。Ray Weaver 和 Drazen Prelec。利用贝叶斯真理血清创建如实报告激励。《营销研究期刊》，50(3):289–302，2013年。Max Welz 和 Andreas Alfons。受访者何时不再在意：识别粗心作答的起始点。arXiv 预印本 arXiv:2303.07167，2023年。Haowen Xu、Wenxiao Chen、Nengwen Zhao、Zeyan Li、Jiahao Bu、Zhihan Li、Ying Liu、Youjian Zhao、Dan Pei、Yang Feng 等。通过变分自编码器对 Web 应用程序中的季节性关键绩效指标进行无监督异常检测。载于《2018年万维网大会论文集》，第 187–196 页，2018年。Xiaoxin Yin、Jiawei Han 和 Philip S Yu。利用多个相互冲突的信息提供者在 Web 上进行真相发现。载于第 13 届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第 1048–1052 页，2007年。Chong Zhou 和 Randy C Paffenroth。利用鲁棒深度自编码器进行异常检测。载于第 23 届 ACM SIGKDD知识发现与数据挖掘国际会议论文集，第 665–674 页，2017年。

贝叶斯网络：一个示例

假设我们有一个包含 5 个变量的数据集：年龄、性别、年级、身高和体重。我们将每个变量离散化为几个直观的类别：

- 年龄 $\in \{A_1 = 12\text{–}14,\ A_2 = 15\text{–}17\}$
- 性别 $\in \{M, F\}$
- 年级 $\in \{G_1 = \text{初中},\ \ G_2 = \text{高中}\}$
- 身高 $\in \{H_1 = \text{矮},\ \ H_2 = \text{平均},\ \ H_3 = \text{高}\}$
- 体重 $\in \{W_1 = \text{轻},\ \ W_2 = \text{平均},\ \ W_3 = \text{重}\}$

我们希望捕捉的直观现实世界依赖关系：身高 $\leftrightarrow$ 体重（强）；年龄 $\rightarrow$ 年级（强；年纪较大的学生更可能在高中就读）；身高 $\leftrightarrow$ 年龄（中等；年纪较大的通常更高）；身高 $\leftrightarrow$ 性别（中等；男性的平均身高略高）。

我们假设有一个极小的数据集（N=10），其充分统计量反映了上述情况：

$$
\begin{aligned}
\text{Root (height):} \quad & \#(H_1, H_2, H_3) = (2, 5, 3). \\
\text{Weight} \mid \text{height}: \quad & \#(W_1, W_2, W_3 \mid H_1) = (2, 0, 0), \\
& \#(W_1, W_2, W_3 \mid H_2) = (1, 3, 1), \\
& \#(W_1, W_2, W_3 \mid H_3) = (0, 1, 2). \\
\text{Sex} \mid \text{height}: \quad & \#(M, F \mid H_1) = (0, 2), \\
& \#(M, F \mid H_2) = (3, 2), \\
& \#(M, F \mid H_3) = (2, 1). \\
\text{Age} \mid \text{height}: \quad & \#(A_1, A_2 \mid H_1) = (2, 0), \\
& \#(A_1, A_2 \mid H_2) = (2, 3), \\
& \#(A_1, A_2 \mid H_3) = (0, 3). \\
\text{Grade} \mid \text{age}: \quad & \#(G_1, G_2 \mid A_1) = (4, 0), \\
& \#(G_1, G_2 \mid A_2) = (1, 5).
\end{aligned}
$$

我们在所有估计中使用拉普拉斯平滑 $\alpha = 1$。根据经验联合分布（经过平滑）计算两两互信息后，最大的权重出现在（身高，体重）和（年龄，年级）这两组变量之间。由于最大生成树必须连接所有变量，一个自然的最大似然树会选择身高作为根节点（互信息中心性高），并添加边：身高 $\rightarrow$ 体重、身高 $\rightarrow$ 性别、身高 $\rightarrow$ 年龄，最后是年龄 $\rightarrow$ 年级。

其中 α=1，类别数为 K_H=3、K_W=3、K_S=2、K_A=2、K_G=2：

$$
\widehat{p}(H) = \frac{\#(H) + \alpha}{N + \alpha K_H} = \left(\tfrac{3}{13}, \tfrac{6}{13}, \tfrac{4}{13}\right) = (H_1, H_2, H_3).
$$

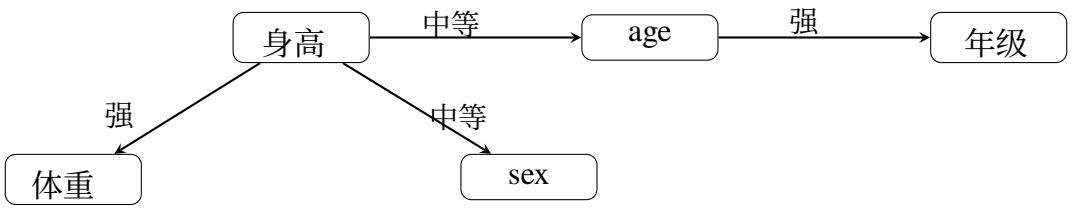


图 5: 玩具示例 Chow–Liu 树：以身高为根节点；边表示最强的互信息 n.

条件概率（经拉普拉斯平滑并按行归一化）：

$$\widehat{p}(W \mid H_1) = \left(\tfrac{3}{5},\ \tfrac{1}{5},\ \tfrac{1}{5}\right) \quad \widehat{p}(W \mid H_2) = \left(\tfrac{2}{8},\ \tfrac{4}{8},\ \tfrac{2}{8}\right)$$

$$\widehat{p}(W \mid H_3) = \left(\tfrac{1}{6},\ \tfrac{2}{6},\ \tfrac{3}{6}\right)$$

$$\widehat{p}(S \mid H_1) = \left(\tfrac{1}{4},\ \tfrac{3}{4}\right) \quad \widehat{p}(S \mid H_2) = \left(\tfrac{4}{7},\ \tfrac{3}{7}\right)$$

$$\widehat{p}(S \mid H_3) = \left(\tfrac{3}{5},\ \tfrac{2}{5}\right)$$

$$\widehat{p}(A \mid H_1) = \left(\tfrac{3}{4},\ \tfrac{1}{4}\right) \quad \widehat{p}(A \mid H_2) = \left(\tfrac{3}{7},\ \tfrac{4}{7}\right)$$

$$\widehat{p}(A \mid H_3) = \left(\tfrac{1}{5},\ \tfrac{4}{5}\right)$$

$$\widehat{p}(G \mid A_1) = \left(\tfrac{5}{6},\ \tfrac{1}{6}\right) \quad \widehat{p}(G \mid A_2) = \left(\tfrac{1}{4},\ \tfrac{3}{4}\right)$$

树的构建过程如下：

1. 计算平滑成对联合分布 $\widehat{p}_{ij}$和互信息 $I(X_i;\ X_j)$。

2. 构建边权为 I 的完全图，并提取最大生成树（我们使用 Prim 算法）。

3. 选择一个根节点（我们选择身高，它具有较高的互信息中心性），将边定向为从根节点向外，并为根节点边缘分布和每个条件概率表估计平滑参数。

现在，我们可以计算学习得到的模型下的完整概率：

$$\widehat{p}_T(x) = \widehat{p}(h)\ \widehat{p}(w \mid h)\ \widehat{p}(s \mid h)\ \widehat{p}(a \mid h)\ \widehat{p}(g \mid a).$$

我们来看看一个非常典型的回答：

$$x_{\text{good}} = \{\ H_3 = \text{tall},\ W_3 = \text{heavy},\ S = M,\ A_2 = 15\text{–}17,\ G_2 = \text{high}\ \}.$$

根据上述表格：

$$\widehat{p}_T(x_{\text{good}}) = \frac{4}{13} \cdot \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{4}{5} \cdot \frac{3}{4} \ \approx\ 0.055.$$

因此 $\log\ \widehat{p}_T(x_{\text{good}})\ \approx -2。90。$

但这是一个非常不典型的案例：

$$x_{\text{bad}} = \{\ H_1 = \text{short},\ W_3 = \text{heavy},\ S = F,$$
$$A_1 = 12\text{–}14,\ G_2 = \text{high}\ \}.$$

则会得到：

$$\widehat{p}_T(x_{\text{bad}}) = \frac{3}{13} \cdot \frac{1}{5} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{6}\ \approx\ 0.0043,$$

因此，对数 $\widehat{p}_T(x_{\text{bad}})\ \approx -5.44$，明显低于典型案例。这具体说明了违反已学习的依赖关系如何降低似然。

B 统计分析

在表 9中，我们分别针对每种方法报告详细的相关性结果。每个单元格列出性能指标（曲线下面积或提升度）与数据集特征（样本数量、变量、特征总数，以及每个变量的平均特征数，AFV）之间的皮尔逊相关系数（r）和斯皮尔曼秩相关（ρ）。曲线下面积部分的最后一列还报告了各方法的曲线下面积与其在不同数据集上的相应重构提升度之间的相关性，以表明重构质量是否与注意力不集中检测能力相一致。如表所示，大多数相关性较小且在统计上并不显著。尽管如此，百分位损失变体呈现出显著的正向曲线下面积–提升度关联，这表明使用百分位损失训练的模型能更好地使重构保真度与异常敏感性相一致——这与我们在第 6.2 节关于主要权衡的讨论一致。作为概率模型，Chow–Liu 不包含重构部分，因此不纳入提升度分析。

在图 6中，我们展示了各数据集上不同方法的平均曲线下面积表现。综合相关性和散点图可进一步印证我们的主要结论：（1）数据集规模、维度和特征粒度对注意力不集中检测性能的影响微乎其微；（2）重构模型相对提升度更高（提升度更大）的数据集，往往能更清晰地区分注意力不集中的受访者。这些发现支持这样一种观点：无监督检测成功与否，主要取决于潜在的调查结构，而非样本数或特征密度。

表9。 方法特定相关性与数据集特征

	AUC					Lift			
方法	样本	变量	特征	AFV	提升度（曲线下面积–提升度）	样本	变量	特征	AFV
非线性自编码器（p = 100）	0.38；0.38	−0。39；−0。27	−0。28；−0。23	0.18；0.20	0.36；0.40	−0.24；−0.17	−0.47；−0.39	−0.54；−0.45	0.10; 0.13
非线性自编码器 (p = 80)	0。25；0。25	−0.53；−0.41	−0。39；−0。25	0.42；0.38	**0.69***；**0.72***	−0.23；−0.17	−0.47；−0.39	−0.55；−0.45	0.10; 0.13
线性自编码器	−0.34；−0.48	0.04; 0.10	−0.11; 0.22	−0.45；−0.44	0。27；0。18	−0.26；−0.20	−0。46；−0。46	−0.52；−0.49	0。11；0。15
Chow–Liu 树	0。13；−0。13	−0.38；−0.11	−0。32；−0。22	−0.06；−0.10	—	（无重构指标）			

每个单元格显示在 n=9 个数据集上计算得出的皮尔逊 r；斯皮尔曼 ρ（双侧检验）。星号表示统计显著性达到 $p < .05$。Chow–Liu提供检测曲线下面积，但不提供重构提升度。

C 权衡图

本附录通过可视化百分位损失阈值 p 在重构保真度和注意力不集中检测之间的权衡，补充说明第 6.2 节的内容。每幅图都绘制了相对于全批次目标（p=100）的指标变化，即 $\Delta M(p)\ \equiv\ M(p)\ -\ M$(100)，并按以下维度汇总：

数据集/条件。与正文一致，随着 $p \to 100$，重构指标（准确率、提升度、ORA）单调提升；而检测指标（曲线下面积、归一化折损累计增益@h、精确率@10/50/100、召回率@h）在 $p \in [85, 90]$ 附近达到最高值，并在接近 p=95–100 时逐渐下降。

图 7a 展示了 Δ准确率与 p；中位数在 p=80/85/90 时为负值，到 p=95 时趋近于 0。图 3 展示了 Δ提升度呈现相同的单调变化趋势，趋向于 p=100。图 7c 展示了 ΔORA，也同样随着 p 增加而上升。综合来看，这些结果证实，PL 在较低的 p 时会牺牲重构效果，并在 p 接近 100 时恢复这一效果。另见第 6.2 节，其中的配对检验量化了这些变化。

图 4 展示了 Δ曲线下面积。增益在 p=85 时最大（在 p=90时也很稳健），在 p=95 时缩小，并在 p=100时消失或反转。图 7e–7i展示了 NDCG@h、P@10/50/100 和 R@h 的类似模式：对于 $p \in [85, 90]$，检测效果通常最强，这与正文报告的符号检验结果一致。

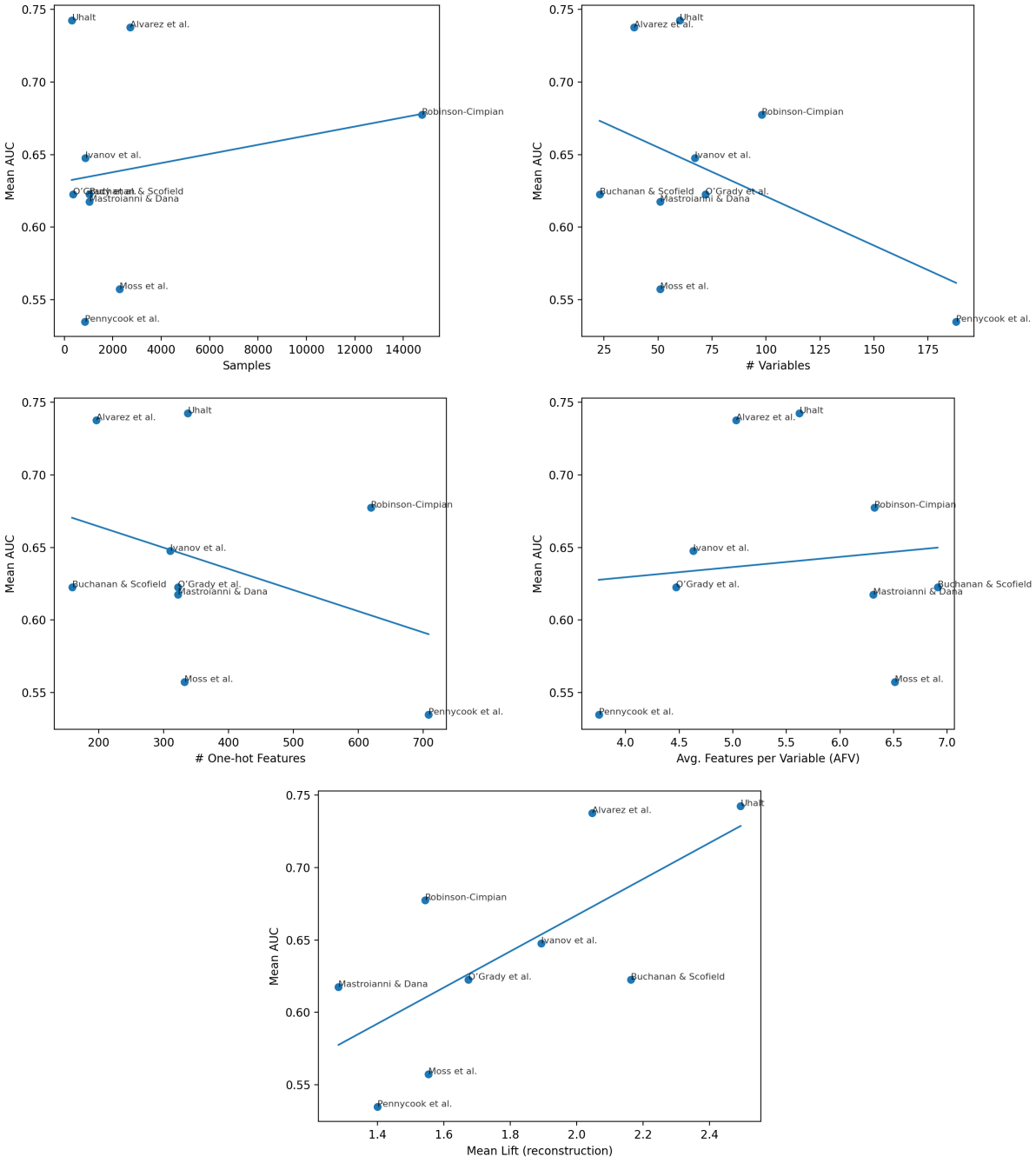
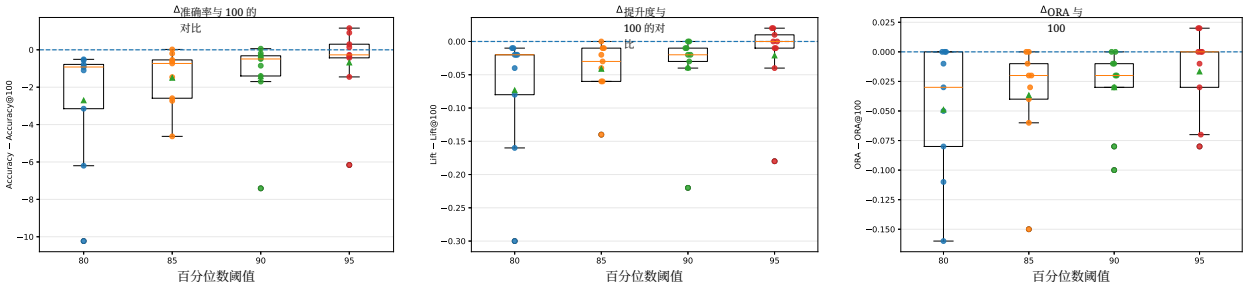


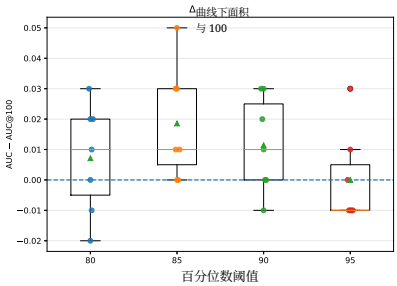
图 6：平均 AUC 与数据集特征的散点图。每个点代表一个数据集。实线表示最小二乘回归拟合。从上到下、从左到右依次为：AUC 与样本数、变量数、特征数、AFV 和平均提升度的关系。未发现明显的强单调趋势，这证实检测性能在很大程度上不受数据集规模或稀疏性的影响，而更多地取决于潜在结构一致性。



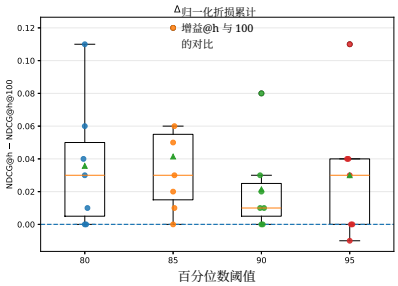
(a) Δ准确率

(b) Δ提升度

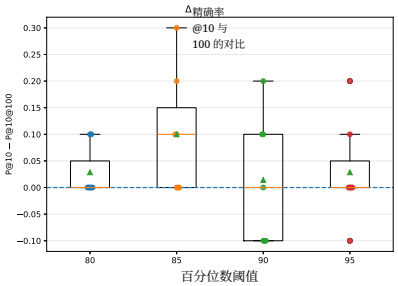
(c) ΔORA



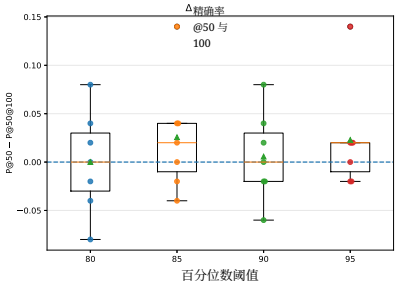
(d) Δ曲线下面积



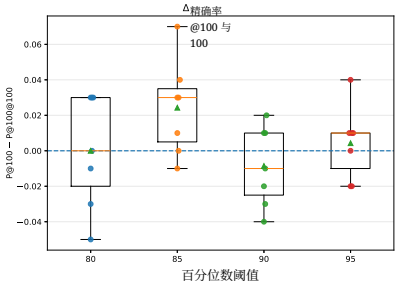
(e) Δ归一化折损累计增益@h



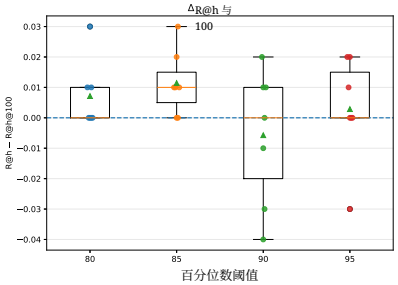
(f) Δ精确率@10



(g) Δ精确率@50



(h) Δ精确率@100



(i) ΔR@h

图7：相对于 p=100 的重构与检测权衡。(a) 在 p=80/85/90 时，中位数为负值，并在 p=95 时趋近于 0。(b) 重构效果随 p 增大而改善。(c) 较高的 p 可提升重构的排序校准效果。(d) 最优点大致位于 p≈85–90。(e) 排序增益与曲线下面积的变化相似，在 p=85–90 达到峰值。(f) 短列表中的精确率从 p=85–90 获益最多。(g) 改进仍然存在，但幅度小于 P@10。(h) 增益有限，与更宽的截断范围相符。(i) 召回率的提升在 p=85–90 左右达到峰值。