从匿名化和不完整表格数据中学习

Lucas Lange 莱比锡大学 & ScaDS.AI Leipzig，德国
lange@informatik.uni-leipzig.de

Adrian Böttinger 莱比锡大学 & ScaDS.AI Leipzig，德国 a.boettinger@t-online.de

Victor Christen 莱比锡大学 & ScaDS.AI Leipzig，德国
christen@informatik.uni-leipzig.de

Anushka Vidanage 澳大利亚国立大学 堪培拉，澳大利亚
anushka.vidanage@anu.edu.au

Peter Christen 澳大利亚国立大学 堪培拉，澳大利亚
peter.christen@anu.edu.au

Erhard Rahm 莱比锡大学 & ScaDS.AI Leipzig，德国 rahm@uni-leipzig.de

arXiv:2602.01217v1 [cs.LG] 1 Feb 2026

摘要

用户驱动的隐私保护使个人能够控制是否共享自己的数据以及共享数据的粒度，因此，同一记录和属性中可能混合原始值、泛化值和缺失值。尽管这种数据表示直观地体现了隐私保护，但也给机器学习带来了挑战：机器学习通常会将非原始值视为新类别或缺失值，从而丢弃其泛化语义。针对从此类表格数据中进行学习，我们提出了考虑异构匿名化的新型数据转换策略，并将其与标准插补及基于大语言模型的方法一并评估。我们采用多个数据集、隐私配置和部署场景进行实验，结果表明，我们的方法能够可靠地恢复效用。研究结果显示，泛化值优于纯抑制；最佳数据准备策略取决于具体场景；一致的数据表示对于维持下游效用至关重要。总体而言，我们的研究发现，能否有效学习取决于对匿名化值的恰当处理。

可用性： 本文的源代码和参考数据可在 https://github.com/luckyos-code/user-driven-privacy 获取。

1 引言

人们对数据隐私日益增强的意识，正在改变个人数据的收集、共享和分析方式 [8, 9, 24, 50]。用户不再被动地放弃控制权，而是越来越要求对数据的使用方式享有透明度和自主权。这一转变促成了个人在线数据存储库（POD） [32, 47]和联邦学习 [34], 等去中心化范式，使数据隐私决策由个人用户而非中央机构主导。

作为一个实际案例，我们正与健康研究人员开展合作项目，利用 POD 重新构想临床试验研究的开展流程。我们正在开发一种以参与者为中心的数据所有权模式，并筹备开展一项试验（于2026年2月开始），在该试验中，敏感医疗和个人数据将在每位参与者自己的 POD 中收集和管理。这种开展临床研究的新方法使试验能够远程且大规模进行，从而支持快速测试、验证以及加速收集研究结果，以获得监管审批。通过采用 POD，并让参与者参与向研究人员共享数据，临床试验如今能够尊重并保护每位参与者的数据隐私。

用户驱动的隐私保护会改变所收集数据集的结构，因为用户可能会在不同的粒度和敏感度层级上有选择地披露信息，而不是在全局范围内采用单一的匿名化策略 [32, 47]。如表 1 所示，此类数据集可能包含原始、泛化（例如，[30–39], 岁、欧洲）以及完全抑制或缺失值（⊥）等多种数据的混合形式。尽管这类表示提供了直观的隐私控制，并能增强用户信任 [8], ，但它们会降低特征特异性并引入歧义，从而给下游效用带来挑战 [37]。

仅靠传统机器学习（ML）流程难以应对这种设置，而缺失数据的标准处理方法会将缺失值视为随机噪声或错误 [46],，而泛化值则编码了正确但经过刻意粗化的信息，

表 1：细粒度数据。

Age	Sex	国家
26	m	法国
[40–49]	f	⊥
old	⊥	欧洲

并带有明确的语义约束。将泛化值视为缺失值会丢弃这些信息，并可能不必要地降低性能。与此同时，针对机器学习数据的隐私保护技术，例如 k匿名化 [53], 差分隐私 [7],以及合成数据生成 [25]，并未解决如何从经过记录级混合隐私变换的数据集中进行学习的问题。在真实的部署场景中，这些问题尤为突出，因为匿名化可能影响训练、推理或二者兼有期间的数据。训练与推理之间的数据表示不一致可能导致严重的分布偏移 [20, 37]。

在本研究中，我们采用以数据为中心的视角，将用户驱动的隐私保护下的学习视为数据准备问题，并着重改进此类数据表示。我们评估了不同数据准备策略对下游机器学习性能的影响，并深入探讨如何处理匿名化表格数据以提升效用。

我们的贡献： (1) 我们对匿名化表格数据上的学习问题进行了形式化，其中原始值、泛化值和缺失值并存。(2) 我们通过实证评估不同场景下的数据准备方法，以解决这一问题。(3) 我们提出的基于粒度的插补方法通过专业化利用泛化语义，实现了较高的效用和鲁棒性。(4) 我们发现：(i) 泛化优于纯抑制；(ii) 数据准备的效果取决于具体场景；(iii) 一致的数据准备对于下游效用至关重要。

2 相关研究

本节通过总结相关研究，为我们的问题和贡献提供背景。我们的工作探讨了数据质量、隐私和机器学习交叉领域中的挑战。

个人数据与用户驱动的隐私保护： 传统隐私机制通常对整个数据集统一应用转换，例如 k-匿名等句法隐私模型 [53]及其扩展 [27, 31], , 它们通过全局抑制或泛化属性来实现不可区分性，直到满足基于分组的约束 [26]。差分隐私 [7] 采用不同范式：向查询或模型更新中注入校准噪声，以限制单条记录的影响。虽然这两种方法都提供形式化隐私保证，但非专业用户难以理解其参数，而且隐私决策通常由集中式方式作出，而非针对个人分别作出 [8]。这些机制的个性化版本基于敏感度或用户偏好，但要么应用于机器学习模型，要么通过将用户排除在数据集之外来实现 [2, 18, 23, 48, 60]。

POD [32, 47]和联邦学习 [34] 等去中心化方法将隐私控制权交给用户，使用户能够决定以多高的精度共享数据。实证研究表明，不同个体和情境的隐私偏好差异显著 [50], , 因此数据集中不同记录的各项属性可能具有不同粒度。尽管这种用户驱动的隐私保护能够提升透明度，并且在某些情况下，由于用户的隐私要求较宽松而提高效用 [1], , 但它也会产生现有隐私感知机器学习流水线尚未解决的异构表示。此前关于 POD 的研究主要关注系统架构和访问控制 [38], , 而对其下游机器学习影响的探讨仍十分有限。值得注意的是，与传统隐私机制不同，用户驱动的方法通常缺乏形式化隐私保证。

机器学习的数据质量： 以数据为中心的机器学习强调，提高数据质量通常比改进模型带来更大的收益 [58, 62]。在此背景下，研究 [28, 37] 证实，缺失值和特征不准确性会对机器学习性能产生显著影响，而且这些影响取决于数据质量问题影响的是训练数据、推理数据，还是两者兼有。我们同样关注这三个问题，其中部署场景、特征完整性和特征准确性是本问题的关键因素。然而，我们的设置在一个重要方面有所不同：异构性是有意为之的，而非错误造成的。泛化值和缺失值源于用户驱动的隐私保护选择，并编码了部分但有效的信息。因此，挑战不在于将错误数据纠正为假定的真实值，而在于以一种既保留效用、又利用隐私约束所固有的剩余价值的方式准备匿名化数据。

缺失值填补： 缺失数据通常使用 Rubin [46],提出的框架进行形式化，该框架将缺失性建模为一个随机过程。经典插补技术会使用点估计替换缺失值，例如均值/众数替换 [17]和多元填补 [59]。基准测试中出现的较新方法 [20, 35] 使用训练得到的模型，包括自动编码器 [11, 33, 41] 和生成模型 [16, 29, 61] , 从观测数据中推断合理的值。在这些方法中，缺失值通常被视为噪声或不完整的观测。

我们的设置有一个根本区别：有意抽象并不等同于缺失数据。泛化值编码了区间或集合值约束（例如年龄 $\in [30–39]$），能够保留部分但具有语义意义的信息。若将此类值视为不受约束的缺失数据，就会丢弃这些约束，并可能导致不合理的插补。关于删失数据、区间值数据或层次数据 [21, 51, 56]的相关研究，将粗观测建模为对潜在精确值的不确定性，假设采用统一的观测机制，并以针对固定分析模型的有效统计推断为目标。相比之下，我们关注列和记录内部的混合粒度，重点研究能够恢复效用而非进行后验恢复的准备策略。

在最近的一项大规模评估中，Khan等人 [20] 进一步凸显了以填补为中心的方法的局限性。他们表明，在不同设置下，没有任何一种插补方法始终优于其他方法；插补的效果取决于缺失是否影响训练数据、推理数据或两者；而常见的填补质量指标并不能很好地预测下游机器学习性能。我们的工作基于这些发现，直接探讨其对机器学习效用的影响，并研究训练与推理条件各异时的准备策略。

最后，数据清洗相关研究利用机器学习检测并修复错误或不完整的记录。Lopster [44],HoloClean [45], 和 CPClean [19] 等系统假设存在一个真实值，以此为目标纠正数据，通常还需要干净的训练数据或外部约束。相比之下，在我们的设置中，泛化值并非错误，而是有意为之的抽象，仍编码了部分语义信息，且没有可用的干净参考数据。我们的目标不是修复数值，而是以能够利用这些信息进行下游学习的方式准备匿名化数据表示。

大语言模型和基础模型： 大语言模型（LLM）已被应用于表格数据的插补、错误检测和预测 [10, 12, 15, 40, 55, 63]。近年来，表格基础模型（TFM）被专门设计用于结构化数据 [54],, 而专用 TFM 与通用大语言模型究竟哪一种更适合数据管理仍存在争议 [43]。TabPFN [14] 是一种 TFM，通过上下文学习且无需针对特定任务进行训练，便可实现出色的预测性能。然而，这些模型通常假设特征表示同质且缺失模式符合传统规律。因此，TabPFN 可能会受到数据质量问题的影响，错误会扭曲受损样本和干净样本的结果 [42]。

在大多数情况下，大语言模型和TFM会将泛化值视为缺失值或类别标记，从而丢弃具有语义意义的信息。如果明确对这类上下文进行编码，模型可能会加以利用，并依赖其通用预训练所获得的先验知识。然而，这些模型会产生高昂的计算成本。

为在不同方法之间进行公平比较，我们考察零样本大语言模型方法，即除了给定的提示上下文之外不进行微调，因为我们直接处理匿名化数据。

3 问题定义

我们对如何从细粒度隐私化数据中学习这一问题进行定义。随后，我们将在第4节介绍解决方案，并在第5节说明用于评估这些方案的实验环境。

3.1 数据模型与符号

令 $D = \{(x_i, y_i)\}_{i=1}^{n}$ 表示一个包含 n 条记录的数据集，其中 $x_i \in \mathrm{X}$ 是特征向量，$y_i \in \mathrm{Y}$ 是记录 i 的标签。每个特征向量 $x_i = (x_{(1)i}, \ldots, x_{(d)i})$ 由 d 个属性组成，这些属性可以是数值型（例如年龄、收入），也可以是类别型（例如性别、职业）。

隐私转换：在用户驱动的隐私系统中，每位用户 u_i（负责控制记录 i）都会在将其记录添加到数据集之前，对该记录应用隐私转换 $\pi_{u_i}: \mathrm{X} \to \mathrm{X}^A$。我们考虑每个属性的三种转换类型：

(1) 原始 (π^{orig})：保持不变，$\pi^{\text{orig}}(x_{(j)i}) = x_i^{(j)}$ (2) 泛化 (π^{gen})：替换为不那么具体的值：• π^{gen} 35 = [30 39] • 数值型：替换为区间，例如（）–类别型：替换为分组，例如（“博士”）“研究生”（3）缺失 (π^{miss})：值被移除，$\pi^{\text{miss}}(x_i^{(j)}) = \perp$

用户可以对任意属性应用任意变换。令 $\tau_i \in \{\text{orig}、\text{gen}、\text{miss}\}^d$ 表示记录 i 的变换类型向量，其中 $\tau_{(j)i}$ 表示对记录 i 的属性 j 所应用的变换，由此得到匿名化数据集 $D^A = \{(\pi_{u_i}(x_i), y_i)\}_{i=1}^{n}$。

隐私配置：我们根据各条记录上的变换分布来刻画数据集。隐私配置元组 (p_{orig}、p_{gen}、p_{miss}) 指定了处于各变换状态的属性-记录对所占的比例，其中 $p_{\text{orig}} + p_{\text{gen}} + p_{\text{miss}} = 1$。

为便于阅读，我们使用 orig-gen-miss记法的简写形式表示各项配置。例如，66-17-17 表示 66% 为原始值、17% 为泛化值、17% 为缺失值。

3.2 学习目标与数据准备

给定匿名化数据集 $D^A = D^A_{\text{train}} \cup D^A_{\text{test}}$，目标是学习一个模型 $f: \mathrm{X}^A \to \mathrm{Y}$，使其期望预测误差最小：$\mathrm{L}(f) = \mathbb{E}_{(x,y)\sim D^A_{\text{test}}}[\ell(f(x), y)]$，其中 ℓ 是损失函数（例如用于分类的 0-1损失）。模型 f 仅使用匿名化 D^A_{train} 进行训练，不接触原始数据；唯一可获取的信息是从 D^A 本身的取值中直接观察到的信息。

标准机器学习面临的挑战：传统机器学习假设 D_{train}和 D_{test} 来自相同的分布，且特征表示一致。我们的异构匿名化数据模型违背了这些假设，因此带来了若干挑战。我们聚焦于三个关键方面：(1) 特征歧义，泛化值（如年龄区间 [30–39] ）代表多个可能的真实值，导致特征空间存在不确定性；(2)信息损失，与原始数据相比，泛化值和缺失值会降低特征与标签之间的互信息；(3) 语义稀释，泛化值扩大了特征域，但在语义上仍包含有关原始值的信息，而标准机器学习算法无法利用这些信息。

数据准备：为应对这些挑战，我们引入一个数据准备函数 $\phi: \mathrm{X}^A \to \mathrm{X}^P$，将记录的匿名化特征映射为一种可能更适合机器学习的表示。此类函数 ϕ可应用于训练数据、测试数据或两者，从而得到：$D^P = \{(\phi(x_i), y_i) \mid (x_i, y_i) \in D^A\}$。数据准备的目标是找到一个 ϕ，使使用准备后的匿名化数据与原始数据之间的效用差距最小：$\phi^* = \arg\min_{\phi\in\Phi} \left|\mathrm{U}(f^*) - \mathrm{U}(f_\phi)\right|$，其中：

- f^* 是在原始数据上训练和评估的基线分类器，
- f_ϕ是在 $\phi(D^P$训练集)上训练、在 $\phi(D^P$测试集)上评估的分类器，• U(·) 衡量模型效用（例如准确率、F1 分数等），• Φ 是可能的数据准备方法构成的空间。

理想的预处理方法可实现 $\mathrm{U}(f_\phi) \approx \mathrm{U}(f^*)$，从而在匿名化后有效维持预测性能。

3.3 部署场景：

随着匿名化可能影响训练、推理或两者的数据，预处理要求也会随之变化。通常应一致地应用数据转换，以避免分布偏移，因为预处理不一致（$\phi_{\text{train}} \neq \phi_{\text{test}}$）会导致特征分布不匹配。然而，当只有一组数据经过匿名化时，预处理可能会使另一组数据发生信息损失。我们确定了三种代表不同现实情境的部署场景：

(1) 匿名化训练（*AnTr*）：训练数据具有敏感性并经过匿名化处理，但推理在可访问原始值的可信环境中进行（例如，使用匿名化的患者记录进行训练，并部署于安全的医院系统中）。(2) 匿名化推理（*AnTe*）：训练使用公开数据或经同意使用的数据，但推理面向注重隐私、会定期匿名化输入的用户（例如，模型使用人口普查数据训练，并由对敏感属性进行泛化处理的用户查询）。(3) 双方均匿名化（*AnBo*）：隐私约束贯穿整个流程，因此两个阶段都必须一致地处理匿名化数据（例如，采用用户可控隐私的联邦学习）。

在 (1) 和 (2) 中，数据准备仅应用于匿名化部分，而在 (3) 中，两者都需要采用一致的数据准备方法。

4 数据准备方法

现在，我们提出一组用于处理异质匿名化表格数据集的数据准备方法。如图 1 所示，我们根据这些技术的核心运行原理将其分为若干类别。每个类别都体现了一种在数据粒度和效用之间权衡的特定策略。这些方法对匿名化数据的影响如图 2 所示。

4.1 不进行预处理（匿名化基线）

该基线将匿名化数据直接输入机器学习模型，不做任何修改。这种方案的缺点是，列域的粒度各不相同，使模型在训练过程中更难将表示与标签正确对齐。如果只有训练数据或测试数据经过匿名化，问题尤其突出，因为模型学到的统计特征与预测时使用的统计特征会不匹配。当两组数据采用相似的匿名化模式时，模型或许能够学到合适的映射。

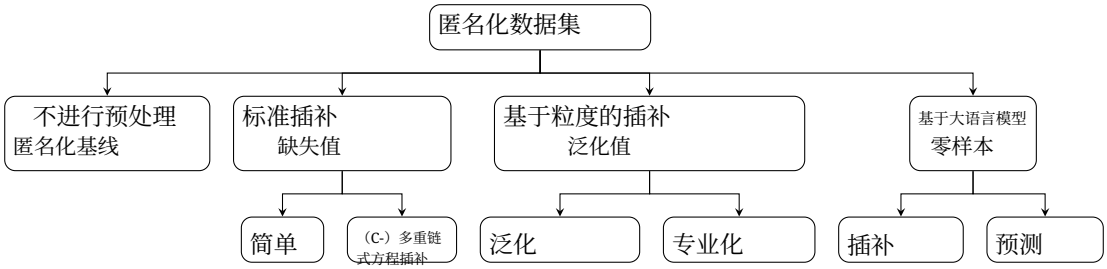

图 1：评估的　　匿名化数据集的数据准备方法　　包含泛化值和缺失值。

（a）匿名化数据集

	ID	Age	Sex	y
原始值:	R1	52	w	1
泛化值:	R2	[50–59]	m	0
缺失值:	R3	42	⊥	1

(b) 标准插补

ID	Age	Sex	y
R1	52	w	1
R2	42	m	0
R3	42	m	1

(c) 泛化

ID	Age	Sex	y
R1	[50–59]	w	1
R2	[50–59]	m	0
R3	[40–49]	⊥	1

(d) 专业化（$\tilde{v} \geq 2$）

ID	Age	Sex	y	w
R1	52	w	1	*1.0*
R2	55	m	0	*1.0*
R3	42	w	1	*0.5*
R3	42	m	1	*0.5*

图 2：匿名化数据样本上可能的准备结果对比，其中 ⊥ 表示缺失值。

4.2 标准插补方法

作为一种常见方案，也是与我们其他方法进行比较的基准，我们采用标准的统计填补策略 [30]。这些统计技术会从可用数据中的模式学习如何填补合理的替代值，但其关注的是数据集中的缺失值。因此，直接将其应用于我们的数据，会把泛化变换视为缺失数据。

简单插补：用根据剩余原始值 [17]计算得出的属性众数（类别型）或均值（数值型）替换缺失值。这种方法快速且简单，但忽略属性间相关性，并可能扭曲特征分布。

MICE：（链式方程多重插补） [59]使用迭代回归，将每个属性建模为其他所有属性的函数。对于存在缺失值的每个属性，它会：(1) 使用具有观测值的记录训练预测模型；(2) 根据其他属性填补缺失值；(3) 迭代直至收敛。因此，MICE 能捕捉属性间依赖关系，并生成统计上合理的值。然而，由于必须为泛化条目假定一个缺失值，它可能违反泛化约束。例如，对于用户指定年龄为 $\in$ [30–39], 的记录，可能会填补年龄= 25 ，从而忽略泛化值中编码的语义约束。

约束型 MICE（C-MICE）：为纳入泛化值中的可用信息，我们在 MICE 基础上增加后处理，将插补值裁剪到其泛化域内。例如，若 MICE 为一条年龄为 $\in$ [30–39][1], 的记录插补出年龄 = 25 ，我们将其替换为 30。该方法利用了语义边界，但可能在泛化集合边界处产生分布伪影。

4.3 基于粒度的填补方法

与基线填补不同，我们基于粒度的插补直接处理泛化值的层次结构。这些方法旨在统一列内的粒度级别，并通过泛化所有数据或将泛化值扩展为其可能的原始值，统一训练数据与测试数据之间的粒度。

强制泛化：该方法通过将所有原始值或较低粒度的值提升到统一的粒度级别，在列域内建立一致性，而缺失值保持不变。这一方法在算法 1 中形式化，能够高效地形成同质化表示，但会丢弃原始值及泛化程度较低的值中更详细的信息。

带筛选的加权特化：与强制泛化相反，专业化通过将每个泛化值扩展为匿名化列数据中所观察到的可能原始值，从而逆转泛化。由于缺失值无非是“极度泛化”的值，我们也对其应用专业化。这与常见的填补方法相似，但我们旨在将数据集的粒度特性纳入方法设计。为此，我们必须应对这种数据准备方法带来的两个关键挑战：（1）每个泛化或缺失条目的可能取值经笛卡尔积造成的数据集指数级增长；（2）对合理性各不相同的候选项给予同等对待。该算法见算法 2，我们将按其组成部分——专业化、加权和筛选——来说明其工作原理。

专门化：对于每条匿名化记录，我们都会生成与其泛化属性相一致的所有具体取值组合。可能的取值仅限于相应匿名化数据集列中观察到的取值。数值属性使用区间中点进行实例化（例如，年龄 $\in$ [30–39] $\rightarrow$ 35），而类别属性则展开为泛化范围内的所有取值（例如，“高等教育” $\rightarrow$ {“学士”、“硕士”、“博士”}，或性别列中缺失值“⊥” $\rightarrow$ {“男”、“女”}）。虽然可以直接应用这种方法，但它有两个关键缺点：(1) 如果 c 个类别属性被泛化为平均大小为 g 的分组，单条记录就会扩展为 g^c 个专门化变体，导致数据集组合式增长，最多可达 $O(n \cdot g^c)$；(2) 由于泛化记录会扩展为许多变体

[1]在准备过程中，所有方法都将“老年”等区间类别作为区间处理。

算法 1强制泛化

要求： 数据集 D^A，每个属性 j 的泛化函数 $\{g_\ell^j\}$ 和层级 ℓ
确保： 泛化数据集 D^P

$D^P \leftarrow D^A$
对于每个属性 $j \in \{1, \ldots, d\}$ **执行**
  $\ell \leftarrow$属性 j 在层次结构中的最高层级
  对于每个记录特征向量 $x_i \in D^P$ **执行**
    如果 $x_i[j]$ 不缺失 **则**
      $x_i[j] \leftarrow g_\ell^j(x_i[j])$ // 泛化到层级 ℓ
    结束条件判断
  结束循环
结束循环
返回 D^P

而未专门化的记录仍保持为单条记录，因此生成的数据集会明显偏向专门化条目，呈现出严重的分布偏斜。

筛选：为缩小第一个问题的范围并控制数据集大小，我们在展开得到可能的变体后，立即为每条记录仅保留 $\tilde{v}$ 个最相关的变体。这限制了输出规模，使其从 $O(n \cdot g^c)$ 缩减为 $O(n \cdot \tilde{v})$。为筛选出最相关的变体，我们采用一种排序策略：根据变体与列概况 P 的相似度为其评分；列概况 P 基于可观测到的原始数据集取值构建，并选取排名前$\tilde{v}$ 的变体。每个变体 s' 的评分方式如下：$\text{score}(s') = \frac{1}{d}\sum_{j=1}^{d} \text{sim}_j(s[j], P_j)$，其中 d是属性数量，而 sim_j 对于类别属性返回归一化频率，对于数值属性则返回 z 分数相似度 $\max(0, 1 - |z_j|/3)$；z_j 是相对于轮廓均值和标准差的 z 分数。这样，原始值中具有常见取值组合的变体将获得更高的评分。

加权：第二个问题——分布偏斜——通过引入记录级权重来解决。原始记录的权重保持为 1，即 $w = 1$；每个派生变体的权重则设为$w = 1/|S_i|$，其中 $|S_i| \le \tilde{v}$表示记录 i在筛选后保留的变体数量。通过这种归一化，无论匿名化输入数据集中的每条记录只出现一次（即未经过专业化），还是扩展为多个专门化变体，其总权重都相同。

在机器学习训练期间，这些权重通过样本加权的方式纳入训练，从而降低各个变体在 $w < 1$时的影响。因此，专业化导致的记录数量膨胀不会对训练得到的模型造成不成比例的影响。在测试数据上进行专业化时，我们可能会从一条匿名化记录生成多条候选记录。在推理期间，这会为同一个体产生多个预测结果，因此需要将其汇总为针对原始条目的最终决策。具体做法是计算专门化条目中所有预测结果的加权平均值：$p_{\text{avg}} = \sum_{i=1}^{\tilde{v}} w_i p_i / \sum_{i=1}^{\tilde{v}} w_i$，其中 $\{p_1、\ldots\ldots、p_{\tilde{v}}\}$ 是来自 $\tilde{v}$个专门化条目的预测结果，其权重为 $\{w_1、\ldots\ldots、w_{\tilde{v}}\}$。这样，我们便能获得标准的推理输出，而不受记录拆分为多个变体的影响。权重在此发挥的作用与训练期间相同，使预测结果能够平等地考虑所有候选项。

不含原始值：对于完全匿名化且不再保留任何原始值的数据集，我们既无法构建概况，也没有可供抽样的观测原始值来处理类别特征。因此，我们无法创建类别变体，必须保留其

算法 2带筛选的加权特化

要求： 数据集 D^A，泛化映射 $\{G_j\}$，筛选后保留的变体数量 变体数量 $\tilde{v}$
确保： 特化数据集 D^P及其权重W

$D^P \leftarrow \emptyset W \leftarrow \emptyset$
如果存在，则根据 D^A 中的原始值构建列概况 P
对于每个记录特征向量 $x_i \in D^A$ **执行**
  $S_i \leftarrow \{x_i\}$ // 该记录的候选项
  对于每个属性 $j \in \{1, \ldots, d\}$ **执行**
    $S_i' \leftarrow \emptyset$
    对于每个条目 $s \in S_i$ **执行**
      如果 $s[j]$ 是原始的**则**
        将 s 添加到 S_i'
      否则如果 $s[j]$ 是数值型且已泛化**则**
        $s[j] \leftarrow$ s[j]中泛化范围的中点
        将 s 添加到 S_i'
      否则如果 $s[j]$ 是数值型且缺失**则**
        $s[j] \leftarrow D^A$中总 $s[j]$ 区间的中点
        将 s 添加到 S_i'
      否则如果 $s[j]$ 是类别型且已泛化 **那么**
        $V \leftarrow G_j(s[j])$ // 基于 D^A中的原始值
        对 $v \in V$ 中的每个 **执行**
          $s' \leftarrow$ 复制(s); $s'[j] \leftarrow v$
          将 s' 添加到 S_i'
        结束循环
      否则如果 $s[j]$ 是类别型且缺失**则**
        $V \leftarrow D^A$的所有可能原始值，位于 $s[j]$中
        对于每个 $v \in V$ **执行**
          $s' \leftarrow$ 复制(s); $s'[j] \leftarrow v$
          将 s' 添加到 S_i'
        结束循环
      结束条件判断
    结束循环
    $S_i \leftarrow S_i'$
  结束循环
  $S_i \leftarrow$ 筛选$(S_i, \tilde{v}, \text{P})$ //筛选为 $\tilde{v}$ 变体
  $w \leftarrow 1/|S_i|$ // 每个保留变体的权重相等
  对每个 $s \in S_i$ **执行**
    将 s添加到 D^P; 将 w添加到 W
  结束循环
结束循环
返回 D^P, W

泛化版本。对于数值特征，我们仍可使用泛化区间值的中点。对于缺失数值，我们可以利用该特征所有泛化区间的边界来推定整体界限，并据此得出中点。

4.4 大语言模型方法

在海量文本语料库上训练的大语言模型，已内化了有关现实世界实体和属性的统计模式 [40]。我们利用这些知识对匿名化数据集进行插补和预测；然而，由于这些方法严重依赖先验知识，因此在处理新任务或专业领域时可能会遇到困难。

大语言模型插补：我们向具备推理能力的预训练通用大语言模型提供提示，要求其根据有关匿名化的上下文信息插补匿名化值。提示配置如下：

```
'系统_提示': "你是一名数据分析师，负责填补缺失值或泛化值。
",↪
```

```
'提示_模板': """数据集：{dataset_name}
任务：为每条记录中指定的列填补具体明确的值。↪

上下文：
- "⊥" 表示值完全缺失 - 像 [30–39] 这样的值表示泛化范围 - 像 "年轻" 这样的语义值属于泛化值

说明：
- 对每记录,h为“目标”中列出的每一列提供一个具体明确的值。↪ - 预测具体值（例如，使用“35”而不是“[30–39]”，使用“Private”而不是“private_sector”）。↪

要处理的记录数：{records_block}

输出格式说明：-每条记录返回一行，使用以下确切格式（不使用 JSON，不使用 Markdown）：REQ_ID<TAB>列1=值1|列2=值2|...↪- 使用管道字符 `|` 分隔列预测值。↪- 值可以包含空格，但绝对不能包含 `|`或制表符。↪- 如果某个值未知，则为该列返回字符串“UNK”。↪

示例行：REQ_0 年龄=35|工作类别=Private REQ_1 职业=Sales

仅返回这些行，不要添加其他内容。"""
```

大语言模型非常适合此用途，因为即使在无需额外训练的零样本设置下，它们也能捕捉复杂的非线性模式。这些模型可以进行微调，但对于缺少真实标注的敏感数据而言，这并不现实。然而，对于数据集中的大量插补而言，这种插补方法受限于 API 访问所带来的高昂计算开销。在金钱和时间限制方面，这可能相当显著，同时也会造成有问题的电力消耗，对环境不利。

大语言模型预测： 除了在插补步骤中使用大语言模型外，我们还可以直接向大语言模型请求预测。在这种方法中，大语言模型会接收测试集记录；根据具体情况，这些记录可能包含泛化值和缺失值，然后直接预测目标标签，除了内部推理所进行的处理外，不执行任何中间插补。在其提示中，大语言模型会接收匿名化记录，并尝试推断其目标标签：

```
'系统_提示': "你正在预测目标变量。",'提示_模板': """数据集：{dataset_name}任务：为以下记录预测目标变量 '{target_name}' （0 或 1）。{target_info} ↪

说明：- 每条记录返回一行，且必须严格采用以下格式（不要使用 JSON 或 Markdown）：REQ_ID<TAB>值 ↪- 值必须为 0 或 1

要处理的记录：{records_block}

示例：REQ_1 0REQ_2 1

仅返回这些行，不要包含其他内容。"""
```

表 2：数据集；Cf. 表示当前相关研究中的使用情况。

Name	领域	#行数	#属性	#类别	#数值	Cf.
信用	金融	1K	20	13	7	[20, 37]
收入	金融	45K	13	8	5	[44]
糖尿病	健康	70K	21	14	7	([20])
就业	招聘	381K	16	15	1	[20]

target_info 变量简要说明了目标变量，例如，对于收入：“目标值：0（收入为 <=50K）或 1（收入为 >50K）”。和之前一样，大语言模型不会在敏感数据上进行预训练，而是用于零样本设置。由于我们直接预测目标值，而这项任务比在较大的训练集上插补许多列的值成本更低，因此与大语言模型插补相比，这种方法的开销更小。然而，与训练经典机器学习模型进行推理相比，我们需要发送许多可能成本高昂的 API 请求，而且推理时间可能更长。直接预测的另一个缺点是，大语言模型几乎无法获取有关数据集特定模式的信息，尤其是在专业领域中，这可能会降低其性能。

5 实验设置

本节将介绍我们实验设计的详细内容。

数据集： 我们在隐私与公平性的背景下，评估了表 2 中来自常见领域的公开数据集。这些数据集的规模各异，类别型属性与数值型属性的比例也不尽相同。德国信用 数据集 [13] 是一个规模较小、包含混合特征的金融数据集，主要用于预测信用状况。成人普查收入 数据集 [22] 包含美国人口普查记录及人口统计属性，用于预测个人年收入是否超过 5 万美元。我们还使用了糖尿病 再入院数据集 [52],的均衡划分二元版本，该数据集基于医院记录，主要包含（二元）类别型医疗特征，旨在推断患者的糖尿病状态。最后，Folk就业 数据集 [6], 同样源自美国人口普查数据，包含大量以类别型为主的申请者描述信息，用于预测就业状态。对于所有数据集，我们都关注二元分类，并采用 80/20 的训练-测试划分。

隐私分布： 为了模拟本问题中不同的数据分布（如第 3 节所述），我们采用多种配置，以代表用户对隐私和数据质量的不同偏好。除了原始数据基线外，我们还考察了原始列中三分之一、三分之二或全部被泛化或设为缺失的情况。为了更好地研究泛化数据的影响，我们还纳入了仅包含缺失数据的设置。这些分布以原始值-泛化值-缺失值的百分比（O-G-M）表示，具体如下：1-0-0（基线）、66-17-17、66-0-34、33-33-34、33-0-67 和 0-66-34。我们纳入不含原始数据的设置，以考察在这一最坏情况下哪些方法仍然有效。我们没有完全抑制所有值，而是将剩余比例分配给泛化值，以保留部分语义信息。不过，由于这种极端情况会使总体结果严重偏向大语言模型等预训练方法，因此我们对该设置进行了部分单独评估。

隐私模拟：为了评估隐私分布，我们需要通过拟合变换来创建匿名化数据集模拟。为此，我们首先必须为评估数据集中的每个属性定义粒度级别。粒度尺度的两个极端分别是原始和缺失，即保留值不变或将其完全移除。介于两者之间的是泛化级别，它们会在语义上将一组原始值归为一类。这些级别及其数量因特征的具体情况而异；例如，对于年龄，我们可以定义两个级别：第一个是范围 [50–59]，对应原始值 52；第二个则是更大的范围，例如“中年”。随后，根据所选分布，对每个属性实施按列匿名化，以应用这些级别的变换。

根据数据集所有属性的目标隐私分布（例如 33-33-34），我们随机选取某个属性中 33% 的值保留原样，33% 的值进行泛化，34% 的值设为缺失。如果存在多个可能的泛化级别，则以相同概率选取其中一个。应用这些转换后，会形成原始值、泛化值和缺失值的异质混合，从而反映目标隐私分布。然而，由于二元属性（例如性别）无法进行有意义的泛化，因此即使它们本应进行泛化，也会被设为缺失。这可能会导致实际配置偏离目标配置，具体取决于可用特征。最后，根据第 3 节所述的场景，这些转换可以应用于训练数据、测试数据或两者。

性能和隐私指标：我们使用宏平均 F1 分数 [4],评估每种方法的性能。该指标通过各类别精确率（$P = \frac{TP}{TP+FP}$）和召回率（$R = \frac{TP}{TP+FN}$）的调和平均值，再对类别 C 取平均，提供均衡的衡量方式：$F1 = \frac{1}{C}\sum_{c=1}^{C} 2 \cdot \frac{P_c \cdot R_c}{P_c + R_c}$。选择该指标而非准确率，是因为它不会受到数据集类别不平衡的影响，并且平等对待所有类别。不过，我们也会分别考察宏平均精确率和召回率 [4]。我们重点比较相对于原始基线的性能保持率，因为我们的数据准备过程应尽量减少匿名化造成的效用损失。因此，我们不采用基于误差的指标，因为我们的方法并非专门旨在精确重构原始数据，而且在某些情况下，甚至不会生成可比的表示。

除了性能之外，我们还需要评估隐私属性，以了解数据准备方法是否会无意中降低匿名化所提供的隐私保护。我们还需要检验这些由用户驱动的转换是否能带来额外的隐私保护，因为它们并不提供形式化隐私保证。为此，我们采用成熟的隐私指标，首先是营销者风险 (R_m) [5]，其定义为攻击者依据准标识符唯一重新识别一条记录的平均概率：$R_m(D) = \frac{1}{|D|}\sum_{x_i \in D} \frac{1}{|\mathrm{EC}(x_i)|}$，其中 $\mathrm{EC}(x_i)$ 是记录 x_i 在准标识符属性 Q 下所属的等价类，即所有这些特征的取值均相同的记录集合。等价地，$R_m = |\mathrm{EC}|/|D|$，其中 $|\mathrm{EC}|$ 是不同等价类的数量，值越低越好。作为另一项基线隐私指标，我们使用k-匿名 [53],；如果数据集中每条记录至少与 $k-1$ 条其他记录不可区分，则该数据集满足此指标：$\forall x_i \in D: |\mathrm{EC}(x_i)| \geq k$。在此，我们报告最小等价类大小，以此表示所有记录达到的最大 k，数值越大越好。对于这两项指标，我们保守地假设 Q 包含所有属性，因为敏感度由用户定义，并非预先确定。

实现：我们使用 Python 实现了实验，并采用 XGBoost [3] 模型作为机器学习流水线。该模型适用于表格数据，且凭借其树结构 [49]，能够鲁棒地处理缺失值。不过，大语言模型预测会直接在测试集上评估，无需进行模型训练。为确保可复现性，我们在整个流程中将所有随机种子设为 42，涵盖从确定性数据准备到机器学习评估的各个环节。可复现性还包括将使用开源推理模型 Mistral-Small- 24B-Instruct-2501 [36] 的插补和预测方法中的大语言模型温度设为 0。所有代码均可在 https://github.com/luckyos-code/user-driven-privacy 获取，该代码库还支持重新创建我们的匿名化数据集，以供进一步研究。

6 结果

我们在涵盖金融、医疗和就业领域的四个数据集上评估了数据准备方法。在原始数据上，这些数据集的基线 F1 分数分别为 0.68（信用）、0.81（收入）、0.75（糖尿病）和 0.82（就业）。图3展示了所有场景、数据集和隐私配置下的完整结果。

6.1 子方法评估

标准插补和专业化均包含第 4 节所述的多种子方法。因此，我们首先确定它们表现最佳的配置，以用于后续评估。我们使用汇总 F1 性能比较各方法，排除 0-66-34，因为该数据集没有可用于插补的原始值，否则会使结果产生偏差。对于标准插补，我们评估了三种方法：简单填补、MICE 和 C-MICE。简单填补取得了最高的 F1 分数（0.61），优于 MICE（0.59）和 C-MICE（0.59），这主要归功于它在 Income 数据集上的表现更佳（+0.13 F1）。它在不同设置下也更加稳定，比基于 MICE 的方法在多出 35% 的情况下表现最佳。这表明，MICE 可能会对数据集特定模式过拟合，进而产生人为的标签相关性，损害模型训练；而众数和均值替换只会强化现有的特征统计量。对于专业化，我们评估了每条记录保留的变体候选项数量（$\tilde{v} \in \{1, 2, 3, 5\}$）。配置文件引导的筛选采用 $\tilde{v} = 2$ 时取得了最佳权衡，汇总 F1 分数达到 0.69。它优于选择 1 个（0.68）或 3 个（0.68）候选项；而增加到 $\tilde{v} = 5$ 仅带来微小提升（+0.003 F1），却显著增大了数据集规模。基于这些结果，我们在后续实验中选择简单填补，以及采用 $\tilde{v} = 2$ 的配置文件引导的专门化。

6.2 匿名化数据与隐私分布

我们也希望考察为评估而创建的匿名化数据集。我们按照第5节所述，使用不同的隐私分布生成这些数据集。重要的是，隐私变换是根据指定的百分比按列应用的。因此，原始值的占比指的是各个特征值，而非完全原始的记录。因此，每条记录包含原始、泛化或抑制特征的概率相同。这种设置产生了更加多样且更具挑战性的数据集，优于那些强制采用固定比例的完全

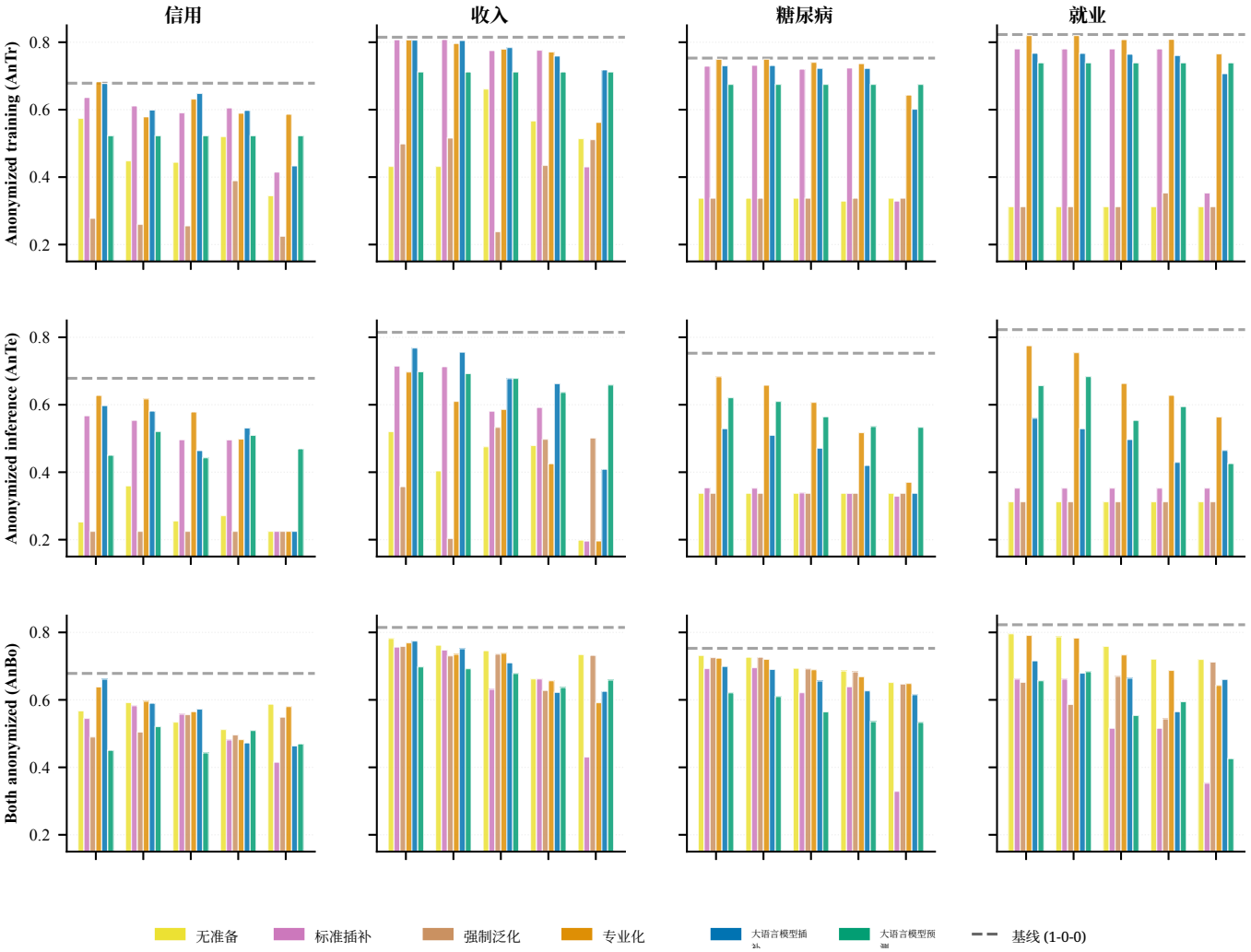

图3：各评估场景、数据集和隐私分布下预处理方法的F1分数。横轴以原始-泛化-缺失百分比表示分布。虚线表示原始数据上的基线性能。

原始记录与完全匿名化记录。通过避免关于用户级分布的假设，我们在整个数据集中获得了细粒度记录。相比之下，按记录处理的方法可能倾向于丢弃匿名化记录而仅保留完全原始的记录，从而将问题简化为更窄的一组情形。

如图3所示，随着匿名化程度提高（即原始值的比例降低），效用会持续下降。从较轻度匿名化（66% 原始值）到较重度匿名化（33%），所有方法的 F1 分数都逐渐下降，但它们的相对排序基本保持稳定。在 0-66-34 这一设置下，原始值已全部消失，达到一个关键阈值。与原始基线相比，即使是最强的方法，性能也相对下降了最多 31%。在其余分布中，我们发现，尽管没有完全原始的记录，只要原始值占有足够比例，性能仍然稳健。由于采用了按列匿名化，在原始值占 66% 时，完全原始记录的比例已经很低（约为 2.5%），而在 33% 及以下时则降为零。最后，每条记录中都存在粗化值，因此在专业化过程中，所有条目都会生成变体；这使数据集大小成为一个重要考量，也促使我们采用过滤策略。

6.3 场景依赖的方法选择

进一步审视各个场景可以发现，图3已经表明不同方法在各设置下的表现存在差异。图4中的热力图报告了相对于原始基线的F1保留率排名前三的方法，这些结果在各数据集和隐私分布上汇总，并排除了0-66-34。我们观察到，没有任何单一预处理方法在所有场景和数据集上都占主导地位，尽管专业化始终位列表现最佳的方法之中。总体而言，最优方法的选择在很大程度上取决于部署环境。

在 AnTr场景中，模型在匿名化数据上训练、在原始数据上部署时，其表现仍接近于

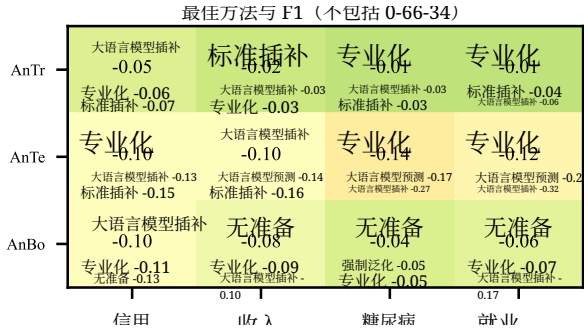

图4：除特殊情况 0-66-34 外，三个最佳数据准备方法在不同隐私分布下相对于原始基线的平均 F1 效用损失热力图。

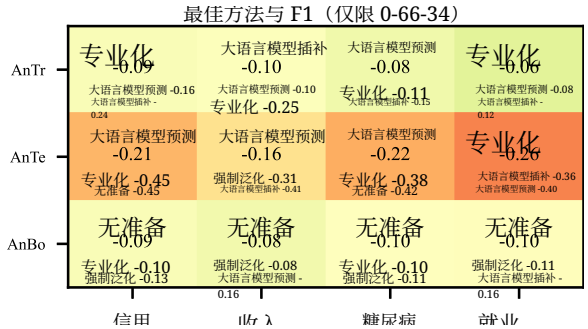

图5：仅针对 0-66-34 分布（不含原始值），三种最佳数据准备方法相对于原始基线的平均 F1 效用损失热力图。

原始基线，在大多数数据集上都是如此，信用数据集是主要例外。AnTe场景中，所有数据集的表现都明显下降，即使采用表现最佳的方法也是如此，因此这是最具挑战性的设置。这种困难源于训练时使用的主要是原始值，而推断时使用的是粗化输入，两者之间存在分布偏移。值得注意的是，直接进行基于大语言模型的预测（无需特定任务模型训练）在这一场景中成为一种有竞争力的替代方案，在除信用数据集外的所有数据集上均排名第二。这表明，预训练所获得的通用世界知识可以部分弥补分布不匹配。

场景 AnTr 和 AnTe 主要通过专业化受益于数据准备，而 AnBo 在直接使用匿名化数据、不进行任何预处理时表现最佳。在这种情况下，训练与推理阶段采用一致的匿名化方式，消除了分布偏移，使模型能够直接从粗化表示中学习有效的决策边界。因此，即使不使用复杂的方法，也能保留相当可观的性能，尽管这类方法仍能取得出色的结果。只有在 Credit 数据集上，也就是规模最小、整体上最具挑战性的数据集上，基于大语言模型的插补仍具有优势，这表明即使分布一致，预训练知识也能弥补训练数据有限的问题。更普遍而言，基于大语言模型的方法在 Credit 和 Income 上的表现优于在 Diabetes 和 Employment 上的表现，这说明前两个领域的预训练覆盖更充分。

数据准备的益处： 通过比较图4中表现最佳的方法与直接使用匿名化数据而不进行预处理的效果，可以量化场景 AnTr 和 AnTe 与 AnBo 之间的差异。AnTr 和 AnTe 的平均 F1 分数分别提升了 +89% 和 +93%，而 AnBo 仅从中获得微乎其微的益处，且仅限于 Credit 数据集。这进一步凸显了结果对部署环境的高度依赖，也清楚表明 AnBo 更适合不进行预处理。

无原始数据： 隐私分布 0-66-34 中不再保留任何原始值，代表最强的隐私保护——
隐私设置，并测试所有数据准备方法的极限。图5展示了在这一极端条件下，各场景与数据集组合中表现最佳的方法。由于大多数其他方法依赖可观察到的原始值，这一设置自然更有利于能够利用预训练知识的基于大语言模型的方法。

在 0-66-34 设置下，场景依赖性总体上依然存在：AnTr 和 AnTe 从数据准备中获益显著，而 AnBo 再次倾向于不进行预处理，且这一趋势如今适用于所有数据集。AnTe 仍是最具挑战性的场景，相较于基线，其 F1 分数大幅下降，因此在缺少原始数据时基本不可行。尽管 AnTr 和 AnBo 的表现相较于隐私保护程度较低的设置也有所下降，但在当前隐私层级下仍然可行。在这一极端情况下，方法选择转向基于大语言模型的预测，其总体表现最佳。基于大语言模型的插补似乎也受限于原始值的缺失，但在某些设置下，专业化仍然可行。这可能是因为缺少原始值时，专业化与不进行预处理较为相似：类别泛化值保持不变，而数值特征仍能从区间中点中获益。总体而言，即使没有原始数据，数据准备方法仍能恢复显著效用（AnTr 中为 +89%，AnTe 中为 +120%），尽管与基线相比，剩余差距在 AnTe 中尤为严重。相比之下，AnBo 在不进行预处理的情况下仍能取得出色结果。

强制泛化： 强制泛化方法表现出的行为与其预期设计不同。尽管该方法旨在通过强制采用统一的泛化层级来解决训练-测试不匹配问题，但由于丢弃了细节，它在各种场景中的表现始终不佳。值得注意的是，强制泛化通常需要额外的领域知识来定义层级结构，而其他方法仅依赖观测到的匿名化值。即使在 AnBo 中，其表现与不进行预处理相当，也没有明显优势，不如保持数据不变。

6.4 汇总性能结果

汇总所有场景、数据集和隐私设置的结果后，可得出表 3 所示的总体方法排名。我们的专门化方法整体表现最为出色，在这些汇总结果中的优势甚至比图4和图5中更加明显。尽管它主要在 AnTr 和 AnTe 场景中胜出，但在 AnBo 场景中仍能与不进行预处理相竞争。在所有设置下，与基线相比，专门化方法的平均 F1 分数损失最低；排除和纳入 0-66-34 分布时，平均损失分别降低 0.09 和 0.12。它的胜率也最高，在所有聚合组合中，取得最佳 F1 分数的比例为

表 3：按场景、数据集和分布进行聚合后排名的方法。左侧为排除 0-66-34 的结果，右侧为包含 0-66-34 的结果；在 0-66-34 中已无原始数据剩余。我们展示每个场景和数据集的 F1 分数，以及平均 F1（F1）、精确率（P）和召回率（R）。Best 表示该方法在聚合案例中取得最佳 F1 分数的比例。注意：宏平均 F1 源自各类别 F1 分数，而非宏平均 P 和 R。缩写：AnTr（Tr）、AnTe（Te）、AnBo（Bo）、信用（C）、收入（I）、糖尿病（D）、就业（E）。

	排除 0-66-34											包含 0-66-34										
	聚合				场景			数据集				聚合				场景			数据集			
方法	F1	P	R	最佳	训练	Te	Bo	C	I	D	E	F1	P	R	最佳	Tr	Te	Bo	C	I	D	E
原始基线。	0.77	0.78	0.77	–	0.77	0.77	0.77	0.68	0.81	0.75	0.82	0.77	0.78	0.77	–	0.77	0.77	0.77	0.68	0.81	0.75	0.82
#1 专业化	**0.68**	**0.70**	**0.69**	**42%**	**0.74**	**0.62**	**0.69**	**0.59**	0.70	**0.69**	**0.76**	**0.65**	**0.69**	**0.67**	**38%**	**0.72**	0.56	0.67	**0.57**	0.65	**0.66**	**0.74**
#2 大语言模型插补	0.65	**0.70**	0.66	17%	0.73	0.56	0.65	0.58	**0.74**	0.63	0.64	0.62	0.67	0.65	15%	0.71	0.52	0.64	0.54	**0.71**	0.60	0.64
#3 大语言模型预测	0.61	0.63	0.63	4%	0.66	0.59	0.59	0.49	0.69	0.61	0.66	0.61	0.62	0.63	10%	0.66	**0.58**	0.58	0.49	0.69	0.61	0.64
#4 标准插补	0.61	0.66	0.64	10%	0.73	0.47	0.62	0.56	0.71	0.58	0.57	0.55	0.59	0.61	8%	0.66	0.43	0.57	0.52	0.64	0.53	0.53
#5 无准备	0.49	0.53	0.57	27%	0.42	0.35	**0.69**	0.44	0.58	0.46	0.46	0.48	0.51	0.57	28%	0.41	0.33	**0.69**	0.43	0.56	0.46	0.46
#6 强制泛化	0.43	0.45	0.54	0%	0.34	0.32	0.64	0.41	0.47	0.48	0.37	0.44	0.44	0.55	0%	0.34	0.32	0.64	0.34	0.53	0.46	0.42

42% 和 38%。因此，专门化方法脱颖而出，成为适用性最广的方法，在不同场景、数据集和隐私分布中均展现出强劲且稳定的表现。即使纳入表现不利的 0-66-34 设置，专门化方法也不逊于其他竞争方法，尽管与能够在缺少原始值时利用通用预训练知识的基于大语言模型的方法相比，它处于劣势。基于大语言模型的插补总体排名第二，凸显了外部知识相较于标准插补技术的优势。相比之下，不进行预处理的方法虽然在 AnBo 场景中表现出色，但在 AnTr 和 AnTe 中表现不佳；在这两个场景中，数据准备显著提升了效用。因此，不进行预处理和强制泛化在总体排名中位居末尾

在所有场景中进行汇总时的排名。

精确率与召回率：除 F1 分数外，表 3 还报告了精确率（P）和召回率（R）。在大多数方法中，精确率与召回率之间的平衡保持相当稳定。显著的例外是强制泛化，它以牺牲精确率为代价，表现出不成比例的高召回率。0-66-34 分布也出现了类似模式：精确率的下降明显大于召回率，而其他分布则保持稳定。这种失衡可能带来问题，因为正确预测（高召回率）会越来越多地被假阳性（低精确率）所掩盖。

计算权衡：若干方法会引入显著的计算开销；在 AnBo 中，当不进行预处理时，该开销可以忽略不计，但在其他场景中则变得不容忽视。专业化会使有效数据集大小从 n 增加到最多 $O(n \cdot \tilde{v})$，具体取决于保留的变体数量 $\tilde{v}$ 以及类别匿名化的普遍程度。这会导致机器学习训练或推理工作负载最多增加至 $\tilde{v}$ 倍。相比之下，基于大语言模型的方法即使采用批处理，仍需要对匿名化记录发起大量大语言模型请求，从而导致运行时间增加，并可能因付费 API 令牌而产生货币成本。此外，必要的大语言模型预训练（或微调）也可能带来额外成本。总体而言，基于大语言模型的方法比专业化产生更高的开销：在 66-17-17 设置下，插补平均大约多耗时 40×，预测平均大约多耗时 8×。专业化所增加的计算成本，可由其在 AnTr 和 AnTe 场景中带来的效用收益加以证明。

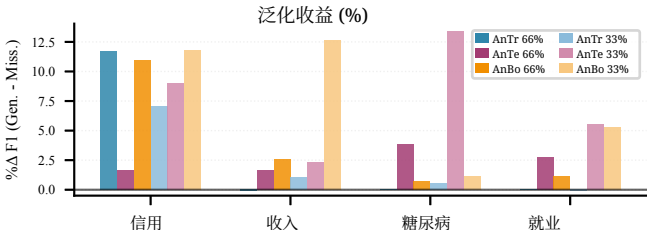

图 6：纳入泛化值（66-17-17 和 33-33-34）与仅将值设为缺失（66-0-34 和 33-0-67）相比，最佳 F1 分数结果的百分比变化。

6.5 泛化收益

我们数据模型中的一项关键评估是：允许使用泛化值是否确实比仅使用缺失值更能提高效用。在图 6 中，我们将由泛化值与缺失值混合而成的分布所取得的最佳结果，与缺失量相应的对照设置进行比较。纳入能够保留底层原始数据部分信息的泛化值，始终能取得优于或类似于完全抑制的结果。收益在我们的专门化方法中最为明显，该方法会显式利用层次结构；在基于大语言模型的插补中也同样明显，后者可将泛化值用作上下文信息。尽管泛化收益的幅度因场景和数据集而异，但它从未表现出可察觉的负面影响。相反，每个数据集都获得了一定的效用提升。

Credit 数据集在所有情况下响应最为明显；而在 AnTr 场景中，泛化对其余数据集的益处似乎较小，效用基本保持不变。相比之下，对于 AnTe 和 AnBo，所有数据集都能持续从纳入泛化值中获益，尤其是在较低的原始设置（33%）下。

总体而言，这些结果表明，与抑制相比，泛化造成的损害较小，并且能够主动提升下游任务的效用。这种效果在 Credit 这类较小的数据集上似乎更为显著，因为每个数据点都至关重要。不过，要实现这些益处，数据准备方法必须能够利用泛化值所提供的额外信息。

表4：按设置和数据集划分的隐私指标，包括营销者风险（R_m↓）、k-匿名性（k↑）及其聚合均值。

分布	信用		收入		糖尿病		就业		mean	
	R_m	k	R_m	k	R_m	k	R_m	k	R_m	k
1-0-0	1.00	1.0	1.00	1.0	0.98	1.0	0.51	1.0	0.87	1.0
66-17-17	0.96	1.0	0.98	1.0	0.95	1.0	0.58	80.4	0.87	20.9
66-0-34	0.96	1.0	0.98	1.0	0.94	1.0	0.55	80.1	0.86	20.8
33-33-34	0.96	1.0	0.90	1.0	0.81	1.0	0.58	38.2	0.81	10.3
33-0-67	0.95	1.0	0.79	1.0	0.74	1.0	0.54	35.4	0.76	9.6
0-66-34	0.73	119.3	0.41	5.4K	0.26	8.4K	0.12	46K	0.38	15K
mean	0.93	20.7	0.84	0.9K	0.78	1.4K	0.48	7.7K	0.76	2.5K

6.6 隐私结果

如第5节所述，我们在表4中报告了营销人员再识别风险（R_m）和所达到的 k-匿名性。在所有数据集中，隐私结果主要取决于所选的隐私分布，而非采用的数据准备方法；表5所示的按方法汇总结果也体现了这一规律。

对于所有保留原始值的配置，R_m 仍然相对较高，只有当原始值的比例降至 33% 时才会出现明显下降。较少匿名化的设置（66-17-17 和 66-0-34）的均值 R_m 值与原始基线（0.87）相当。相应地，对于 Income、Diabetes 和 Credit，实现的 k-匿名性均为 1，这表明在我们将所有属性视为准标识符的保守假设下，许多记录仍然是唯一的。Employment 则是一个部分例外：即使在原始值占 66% 的中等隐私分布下，其 k-值仍然较大，说明数据集特定的特征也发挥着关键作用。对于极端的 0-66-34 配置，可以观察到性质不同的行为，其中不再保留任何原始值。在这种情况下，R_m 在所有数据集上都急剧下降（均值 $R_m = 0.38$），而 k-匿名性则提升了数个数量级，Employment 最高达到 $k = 46$K。这凸显了完全移除原始属性值所产生的强烈聚合效应，因为该效应有效消除了记录层面的唯一性。

表5：按方法划分的隐私（不包括0-66-34）。

方法	R_m	k
原始基线。	0.87	1.0
无预处理。	0.92	1.0
标准插补。	0.73	33.1
强制泛化	0.81	1.0
专业化	0.91	1.0
大语言模型插补	0.79	31.7
大语言模型预测	0.92	1.0

表5总结了我们各数据准备方法的隐私指标（不包括 0-66-34，因为它会主导聚合结果），并表明不同方法之间的差异较为适中。令人意外的是，仅使用匿名化数据而不进行任何数据准备，得到的结果最差（$R_m = 0.92$、$k = 1.0$）；对基于该数据进行预测的大语言模型而言，结果也完全相同。这表明，传统隐私指标可能无法充分反映异质化、用户驱动的匿名化所具有的隐私属性，因为将泛化值和缺失性视为精确标识符时，可能会造成记录层面的唯一性。尽管专门化会逆转泛化，但这一特性仍然存在，因为将泛化值扩展为多个变体，会形成不同的记录概况，其中每个变体的重构值都略有不同，从而提高了基于语法的重新识别指标所衡量的唯一性。相比之下，标准插补和基于大语言模型的插补所带来的隐私保护略好，这可能是因为它们插补出的属性值更为同质，从而增大了等价类大小。

总体而言，所评估的数据准备方法均未对抗或逆转用户驱动的匿名化，而是维持了相当的隐私层级。相反，观察到的隐私指标差异主要由匿名化数据集的用户驱动隐私分布以及剩余原始值的比例决定。值得注意的是，在 R_m 和 k-匿名性下，我们所提出的能够提升效用的方法（例如专业化）并不会损害匿名化数据经测量的隐私性。

7 讨论

我们讨论了针对具有异构隐私转换的表格数据集的数据准备研究结果。所有方法均直接作用于匿名化数据，通过可能改变其表示形式来恢复下游效用，同时不修改机器学习模型。

效用： 数据准备的有效性高度取决于具体场景。AnTr 场景是最宽容的设置，F1 分数仅出现轻微下降，尤其是在我们提出的专业化方法下。AnTe 也能从数据准备中获益匪浅，但它是最具挑战性的场景。特别是当不再保留任何原始值时，学习几乎变得不可行。

相比之下，AnBo 无需进行数据准备即可保持较高效用。在这种设置下，数据准备仍然有效，但不进行预处理也能取得相同结果，并可避免计算开销。在此，训练与推理阶段保持一致的匿名化，使模型能够直接从粗化表示中学习，而基于重构的方法则可能引入不必要的噪声。此外，匿名化通常不会引入错误，泛化值仍然传达正确的信息，只是信息经过了粗化。因此，AnBo 设置挑战了异构值必然存在问题这一假设：保留异构性的任何预处理都表现良好，而强制泛化——即明确地强制实现同质性——则表现不佳。

当没有原始数据可用时，直接大语言模型预测在 AnTe 中表现更优，它利用预训练来弥补测试特征域不可观测的问题。专业化仍具竞争力，但在缺少原始值时存在固有限制，其他基于插补的方法也有类似局限。与基线填补相比，专业化使平均 F1 相对提升了 12% 至 18%。

总体而言，专业化是最稳健、最通用的预处理策略，在各种场景下表现始终出色，且所需假设最少。它对泛化值和缺失值都有效；与完全抑制相比，保留泛化值显然更有优势。最后，如果已知部署场景，我们建议 AnBo 不进行预处理，而对 AnTr 和 AnTe 采用专业化。不过，如果场景未知或可能发生变化，专业化是最稳妥的默认选择。

隐私： 针对匿名化数据进行数据准备时，我们关注的是这是否会通过逆转匿名化而损害隐私。结果表明，我们的方法不会显著增加重新识别风险。相反，隐私在很大程度上取决于数据分布，尤其是保留的原始值占比。值得注意的是，未经数据准备的匿名化数据似乎比

原始基线更不隐私，尤其是在营销者风险和 k-匿名性指标下，尽管从概念上说，它的隐私保护程度更高。这些指标将泛化值视为精确标识符，未能捕捉集合值语义。因此，异构性增加可能会推高记录唯一性，却并未真正削弱预期的隐私保护；尤其是在我们保守地假设所有属性均为准标识符的情况下。经典指标仍然有用，但在评估异构的用户驱动匿名化时应谨慎解读。需要指出的是，这类隐私转换无法提供形式化保证，因此未来研究可引入更适合为用户驱动的隐私决策提供依据的指标。

可扩展性： 我们在第6.4节比较了大语言模型与专业化的计算开销，结果表明，基于大语言模型的方法运行时间显著更长。我们的实验代表了对专业化而言近乎最坏情况的设置，因为所有特征都被视为敏感特征，并按列匿名化，从而产生混合记录，因此每条记录都有大量变体。尽管如此，通过筛选出较少的候选项（例如 $\tilde{v} = 2$），我们仍能将数据集增长控制在合理范围内，最多仅使机器学习计算量增至原始水平的 $\tilde{v}$ 倍；相比之下，其他数据集级或生成式增强方法已知会产生高得多的成本 [39, 57]。然而，对于具有大量类别特征且值域非常大的数据集，即使是筛选步骤也可能带来不可忽视的开销。在实际应用中，通常只有一部分属性会被视为敏感属性，这能显著提升可扩展性 [20]。还可以通过依据重要性评分优先对特征进行专业化，或在数据收集期间合并较大的类别型值域，来缩减特征空间。探索此类预筛选策略是未来值得研究的方向。

假设： 在无法获取完整特征层级结构或外部知识的情况下，我们对领域知识采取保守假设。我们仅依赖匿名化后观察到的值，不使用真实值、辅助数据或预训练。我们的专业化只需将泛化值与观察到的原始值关联起来，而缺失值则可直接观察到。在实践中，领域层级结构通常是已知的或可以推断出来的；未来的工作可以利用大语言模型对其进行近似，尤其是在原始值已全部不可用时。我们进一步假设隐私分布是按列而非用户特定的，因为后者会引入关于用户群体的额外假设，并且由于各条记录的相对重要性不同，可能产生不均衡的影响。我们将这些隐私转换应用于所有属性，从而消除因选择哪些属性可能对某个用户敏感而产生的潜在偏差。由于缺乏带有隐私真实值的真实世界数据，我们的评估依赖合成生成的数据集；这是一个常见的局限。最后，我们的目标并非精确重构原始数据，而是在隐私约束下通过数据准备提升下游效用。

可迁移性： 我们的研究结果适用于分类之外涉及匿名化数据统计分析的任务。除下游机器学习外，这还包括对聚合结果或近似输出进行操作的问题，例如聚合、筛选和分组查询。这种可迁移性源于我们的方法针对的是数据层面的特征表示，而非下游应用本身。由于数据集成、模式演变，或不一致的数据收集也会导致特征域异构，因此我们的方法在隐私驱动场景之外仍然适用。此外，如前所示，这些方法能够直接处理仅由缺失数据导致的不完整数据集。

与先前工作的关系： 我们的发现补充了第 2 节和表 2 中关于数据质量与机器学习性能的先前研究，将研究拓展至混合隐私转换，并提出了相应的数据准备策略。与早期研究类似 [20, 37]，我们观察到，在训练与推理条件不匹配时，匿名化会导致性能大幅下降。不过，我们证明，恰当的数据准备能够挽回大部分损失的效用。与敏感属性集较小（性别、年龄和种族）的设置相比 [20]，我们采取的保守隐私假设代表了更严格的情形。将属性集从所有属性缩减为仅包含一个子集，可能会提升效用、可扩展性和隐私指标。不同于假设能够访问一些基本干净的训练数据的数据清洗系统 [44]，我们仅使用匿名化输入，并将泛化值视为有效约束，而非错误。保留并利用这些约束，能够实现标准填补方法往往无法达到的效用提升。

8 结论

在用户驱动的隐私保护下进行数据收集，会产生这样的表格数据集：原始值、泛化值和缺失值可能在属性和记录中并存。这给标准机器学习流水线带来了挑战，因而导致特征域不一致并引入分布偏移。在本研究中，我们提出并评估了数据准备方法，以支持在现实部署场景中针对这类异构数据表示进行学习。结果表明，与纯抑制相比，泛化值能够始终保留更多效用；数据准备的效果高度依赖于具体场景；而训练与推理阶段对数据表示的处理不一致，是性能损失的主要来源。我们还发现，在训练与推理阶段采用一致的匿名化方式，可以完全免去数据准备。在所有设置中，我们提出的基于专业化的基于粒度的插补，是最稳健的数据准备策略；在假设有限且不损害所测隐私的前提下，它能够挽回大量分类效用。这些发现表明，在用户驱动的隐私保护下实现有效学习，并不需要修改机器学习模型或逆转匿名化。相反，通过相应地准备匿名化数据，便能恢复显著的效用，因此恰当的数据准备对于实用的隐私感知机器学习至关重要。

未来的工作可以致力于收集真实世界数据集，或拓展至基于可比数据表示的其他下游机器学习和分析任务。若能为用户驱动的隐私保护开发更合适、信息量更丰富的隐私指标，也将大有裨益。

致谢

作者感谢德国联邦教育与研究部以及萨克森州科学、文化与旅游部为 ScaDS.AI 提供的资金支持。本研究的部分计算工作使用了莱比锡大学计算中心的资源。本研究还获得了澳大利亚大学联盟和德国学术交流中心（DAAD）57701258号项目的部分资助。

参考文献

[1] Alessandro Acquisti、Curtis R. Taylor 和 Liad Wagman。2016年。《隐私经济学》。载于《信息系统经济学手册》，Shane Greenstein 和 Avi Goldfarb 编。第2卷。爱思唯尔，442–492。[2] Franziska Boenisch、Christopher Mühl、Adam Dziedzic、Roy Rinberg 和 Nicolas Papernot。2023年。按您的方式定制：DP-SGD 的个性化隐私分配。《神经信息处理系统进展》36（2023），19073–19103。[3] Tianqi Chen 和 Carlos Guestrin。2016年。XGBoost：一种可扩展的树提升系统。载于第22届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集。785–794。[4] Peter Christen、David J Hand 和 Nishadi Kirielle。2023年。F值综述：其历史、性质、批评及替代方案。《计算综述》56卷3期（2023），1–24。[5] Fida Kamal Dankar 和 Khaled El Emam。2010年。一种评估营销人员再识别风险的方法。载于2010年 EDBT/ICDT 研讨会论文集。1–10。[6] Frances Ding、Moritz Hardt、John Miller 和 Ludwig Schmidt。2021年。告别 Adult：面向公平机器学习的新数据集。《神经信息处理系统进展》34（2021）。[7] Cynthia Dwork。2006年。差分隐私。载于自动机、语言与编程国际研讨会。施普林格，1–12。[8] Claus-Peter Hermann Ernst 和 Alexander W. Ernst。2016年。隐私风险对智能手表使用的影响。载于美洲信息系统会议。https://api.semanticscholar.org/CorpusID:33031778。[9] 欧洲议会和欧盟理事会。2016年。欧洲议会和理事会2016年4月27日关于在处理个人数据过程中保护自然人以及此类数据自由流动的条例（EU）2016/679（通用数据保护条例）。https://eur-lex.europa.eu/eli/reg/2016/679/oj 第20条：数据可携权。[10] Xi Fang、Weijie Xu、Fiona Anting Tan、Ziqing Hu、Jiani Zhang、Yanjun Qi、Srinivasan H. Sengamedu 和 Christos Faloutsos。2024年。表格数据上的大语言模型（LLM）：预测、生成与理解——综述。《机器学习研究汇刊》（2024）。https://openreview.net/forum?id=IZnrCGF9WI。[11] Lovedeep Gondara 和 Ke Wang。2018年。Mida：使用去噪自编码器进行多重插补。载于亚太知识发现与数据挖掘会议。施普林格，260–272。[12] Stefan Hegselmann、Alejandro Buendia、Hunter Lang、Monica Agrawal、Xiaoyi Jiang 和 David Sontag。2023年。TabLLM：使用大语言模型对表格数据进行少样本分类。载于人工智能与统计学国际会议。PMLR，5549–5581。[13] Hans Hofmann。1994年。Statlog（德国信用数据）UCI机器学习仓库。https://doi.org/10.24432/C5NC77。[14] Noah Hollmann、Samuel Müller、Lennart Purucker、Arjun Krishnakumar、Max Körfer、Shi Bin Hoo、Robin Tibor Schirrmeister 和 Frank Hutter。2025年。使用表格基础模型对小规模数据进行准确预测。《自然》637（2025），319–326。[15] Xiang Huang 和 Shuang Hao。2024年。通过带有可训练提示和检索增强的预训练语言模型填补缺失值。。载于 VLDB 研讨会：DATAI。[16] Niels Bruun Ipsen、Pierre-Alexandre Mattei 和 Jes Frellsen。2021年。not-MIWAE：基于非随机缺失数据的深度生成建模。载于国际学习表征会议。[17] Mortaza Jamshidian 和 Matthew Mata。2007年。数据不完整时均值与协方差结构分析的进展。载于《潜变量及相关模型手册》。爱思唯尔，21–44。[18] Zach Jorgensen、Ting Yu 和 Graham Cormode。2015年。保守还是宽松？个性化差分隐私。载于2015年 IEEE 第31届国际数据工程会议。IEEE，1023–1034。[19] Bojan Karlaš、Peng Li、Renzhi Wu、Nezihe Merve Gürel、Xu Chu、Wentao Wu 和 Ce Zhang。2020年。不完整信息上的最近邻分类器：从确定答案到确定预测。《VLDB 基金会汇刊》14卷3期（2020）255–267。10.14778/3430915.3430917），–。https://doi.org/。[20] Falaah Arif Khan、Denys Herasymuk、Nazar Protsiv 和 Julia Stoyanovich。2025年。空值的更多面貌：负责任缺失值插补评估套件。《VLDB 汇刊》18卷9期（2025），2899–2913。https://doi.org/10.14778/3746405.3746416。[21] John P Klein 和 Melvin L Moeschberger。2003年。生存分析：删失与截断数据的分析技术。施普林格科学与商业媒体。[22] Ron Kohavi。1996年。提高朴素贝叶斯分类器的准确率：一种决策树混合方法。载于 Kdd，第96卷。202–207。[23] Lucas Lange、Ole Borchardt 和 Erhard Rahm。2025年。通过客户端采样实现个性化隐私的联邦学习。载于2025年第10届机器学习技术国际会议（ICMLT）。213–217。https://doi.org/10.1109/ICMLT65785.2025.11193238。[24] Lucas Lange、Tobias Schreieder、Victor Christen 和 Erhard Rahm。2025年。切分数据：揭示智能手表健康数据中的用户身份。载于第20届 ACM 亚洲计算机与通信安全会议论文集。710–726。[25] Lucas Lange、Nils Wenzlitschke 和 Erhard Rahm。2024年。用于隐私保护型可穿戴压力检测的合成健康传感器数据生成。《传感器》24，10（2024年），3052。[26] Kristen LeFevre、David J DeWitt 和 Raghu Ramakrishnan。2006年。Mondrian 多维 k-匿名。（2006年），25–25。[27] Ninghui Li、Tiancheng Li 和 Suresh Venkatasubramanian。2007年。t-接近性：超越 k-匿名与 l-多样性的隐私保护。载于《2007年 IEEE 第 23 届国际数据工程会议》。IEEE，106–115。[28] Peng Li、Xi Rao、Jennifer Blase、Yue Zhang、Xu Chu 和 Ce Zhang。2021年。CleanML：评估数据清洗对机器学习分类任务影响的研究。载于《第 37 届 IEEE 国际数据工程会议（ICDE）论文集》。13 24 10.1109/51399.2021.00009 –。https://doi.org/。/ICDE。[29]。Steven Cheng-Xian Li、Bo Jiang 和 Benjamin Marlin。2019年。MisGAN：利用生成对抗网络从不完整数据中学习。载于《国际学习表征会议》。[30] Roderick JA Little 和 Donald B Rubin。2019年。缺失数据的统计分析。第 793 卷。John Wiley & Sons。[31] Ashwin Machanavajjhala、Daniel Kifer、Johannes Gehrke 和 Muthuramakrishnan Venkitasubramaniam。2007年。l-多样性：超越 k-匿名的隐私保护。载于 ACM《数据知识发现汇刊》（TKDD），第 1 卷。ACM，3。[32] Essam Mansour、Andrei Vlad Sambra、Sandro Hawke、Maged Zereba、Sarven Capadisli、Abdurrahman Ghanem、Ashraf Aboulnaga 和 Tim Berners-Lee。2016年。Solid 平台在社交网络应用中的演示。载于《第 25 届万维网国际会议伴随论文集》。223–226。[33] Pierre-Alexandre Mattei 和 Jes Frellsen。2019年。MIWAE：不完整数据集的深度生成建模与插补。载于《国际机器学习会议》。PMLR，4413–4423。[34] Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。2017年。去中心化数据上的深度网络通信高效学习。载于《人工智能与统计学》。PMLR，1273–1282。[35] Xiaoye Miao、Yangyang Wu、Lu Chen、Yunjun Gao 和 Jianwei Yin。2022年。缺失数据插补算法的实验性综述。IEEE《知识与数据工程汇刊》35，7（2022年），6630–6650。[36] MistralAI。2025年。Mistral Small 3。https://mistral.ai/fr/news/mistral-small-3。[37] Sedir Mohammed、Lukas Budach、Moritz Feuerpfeil、Nina Ihde、Andrea Nathansen、Nele Noack、Hendrik Patzlaff、Felix Naumann 和 Hazar Harmouch。2025年。表格数据的数据质量对机器学习性能的影响。《信息系统》132（2025年），102549。https://doi.org/10.1016/j.is.2025.102549。[38] Richard Mortier、Hamed Haddadi、Tristan Henderson、Derek McAuley 和 Jon Crowcroft。2016年。使用数据盒进行个人数据管理：盒子里有什么？《2016年 ACM 云辅助网络研讨会论文集》（2016年），49–54。[39] Alhassan Mumuni 和 Fuseini Mumuni。2022年。数据增强：现代方法的全面综述。《Array》16（2022年），100258。[40] Avanika Narayan、Ines Chami、Laurel Orr 和 Christopher Ré。2022年。基础模型能否处理你的数据？载于《VLDB 基金会论文集》，第 16 卷。VLDB 基金会，738–746。[41] Alfredo Nazabal、Pablo M Olmos、Zoubin Ghahramani 和 Isabel Valera。2020年。使用变分自编码器处理不完整异构数据。《模式识别》107（2020年），107501。[42] Vasileios Papastergios 和 Anastasios Gounaris。2025年。走向真实世界：探究不完美数据对表格基础模型的影响。载于 VLDB 研讨会：QDB。[43] Paolo Papotti 和 Carsten Binnig。2025年。神经关系数据专题讨论：表格基础模型、大语言模型……还是二者兼有？载于《VLDB 基金会论文集》，第 18 卷。5513–5515。[44] Eduardo Reis、Mohamed Abdelaal 和 Carsten Binnig。2024年。潜在空间中的可泛化表格数据清洗。《VLDB 汇刊》17，13（2024年），4786–4798。10.14778/3704965.3704983。https://doi.org/。/。[45] Theodoros Rekatsinas、Xu Chu、Ihab F. Ilyas 和 Christopher Ré。2017年。HoloClean：利用概率推断进行整体数据修复。《VLDB 汇刊》10，11（2017年），1190–1201。10.14778/3137628.3137631 –。https://doi.org/。/。[46] Donald B Rubin。1976。推理与缺失数据。《Biometrika》63，3（1976年），581–592。[47] Andrei Vlad Sambra、Essam Mansour、Sandro Hawke、Maged Zereba、Nicola Greco、Abdurrahman Ghanem、Dmitri Zagidulin、Ashraf Aboulnaga 和 Tim Berners-Lee。2016年。Solid：一种基于关联数据的去中心化社交应用平台。MIT 计算机科学与人工智能实验室与卡塔尔计算研究所，技术报告 2016（2016年）。[48] Maja Schneider、Erik Buchmann 和 Erhard Rahm。2024年。分布式隐私感知位置数据聚合。载于《2024年 IEEE 第 6 届智能系统与应用中的信任、隐私与安全国际会议（TPS-ISA）》。IEEE，31–40。[49] R. Shwartz-Ziv 和 A. Armon。2021年。表格数据：深度学习并非你所需要的一切。arXiv 预印本 arXiv:2106.03253（2021年）。[50] Chan S. Song 和 Yong-Ki Kim。2021年。消费者愿意向时尚销售机器人分享个人信息的预测因素。《零售与消费者服务杂志》63（2021年），102727。[51] Matthias Speidel、Jörg Drechsler 和 Shahab Jolani。2020年。R 软件包 hmi：用于分层多重插补及其他功能的便捷工具。《统计软件》

统计软件 95, 9 2020 1 48 10 18637 095 09 ( ), – (2025年). [58] Steven Euijong Whang、Yuji Roh、Hwanjun Song和Jae-Gil Lee. 2023. 深度学习中的数据收集与质量挑战：以数据为中心的人工智能视角。*VLDB*期刊 32, 4 (2023), 791–813. [59] Ian R White、Patrick Royston和Angela M Wood. 2011. 使用链式方程的多重插补：实践中的问题与指导。医学中的统计学 30, 4 2011 377 399 ( ), – .[60] Jiale Xu、Benjamin S Glicksberg、Chang Su、Peter Walker、Jiang Bian和Fei Wang. 2020. 具有异构隐私的联邦学习。*IEEEAccess* 8 (2020), 125011–125020.[61] Jinsung Yoon、James Jordon和Mihaela Schaar. 2018. Gain：使用生成对抗网络进行缺失数据插补。收录于国际机器学习会议。PMLR, 5689–5698. [62] Daochen Zha、Zaid Pervaiz Bhat、Kwei-Herng Lai、Fan Yang、Zhimeng Jiang、Shaochen Zhong和Xia Hu. 2025年. 以数据为中心的人工智能：综述。计算综述 57, 5 (2025年), 1–42. [63] Haochen Zhang、Yuyang Dong、Chuan Xiao和Masafumi Oyamada. 2024年. 大语言模型作为数据预处理器。*VLDB* 研讨会：*TaDA* (2024 年). [52] 。https://doi.org/ 。/jss.v .i Beata Strack 、 Jonathan P DeShazo 、 Chris Gennings 、 Juan L Olmo、Sebastian Ventura、Krzysztof J Cios 和 John N Clore。2014年。糖化血红蛋白测定对医院再入院率的影响：对 70,000 条临床数据库患者记录的分析。国际生物医学研究 2014（2014年）。[53] Latanya Sweeney。2002年。k-匿名：一种保护隐私的模型。《不确定性、模糊性与基于知识的系统》 10，05（2002年），557–570。[54] Boris van Breugel 和 Mihaela van der Schaar。2024年。立场：为什么表格基础模型应成为研究重点。载于《第 *41* 届国际机器学习会议论文集》。48976–48993。[55] Jian wei Wang、Kai Wang、Ying Zhang、Wenjie Zhang、Xiwei Xu 和 Xuemin Lin。2025年。关于大语言模型增强的混合类型数据填补与高阶消息传递。《*VLDB* 汇刊》 18，10（2025年），3421–3434。[56] Wei Wang 和 Zhi-Hua Zhou。2015年。从不完整且不准确的监督中学习。知识发现与数据挖掘 （2015年），1–10。[57] Zaitian Wang、Pengfei Wang、Kunpeng Liu、Pengyang Wang、Yanjie Fu、Chang-Tien Lu、Charu C Aggarwal、裴健和 Yuanchun Zhou。2025年。数据增强综合综述。*IEEE*《知识与数据工程汇刊》