
利用大型语言模型生成具有多样写作风格的隐私保护合成评论

Tevin Atwal, Chan Nam Tieu, Yefeng Yuan, Zhan Shi, Yuhong Liu

Department of Computer Science and Engineering

Santa Clara University

{ tatwal, ctieu, yyuan4, ashi2, yhliu } @scu.edu

Liang Cheng

eBay

liacheng@ebay.com

ABSTRACT

由大型语言模型 (LLM) 生成的合成数据的日益增多，在数据驱动的应用中既带来了机会也提出了挑战。尽管合成数据为模型训练提供了一个经济高效、可扩展的替代方案，以替代真实世界的真实数据，但其多样性和隐私风险仍未得到充分探索。聚焦于基于文本的合成数据，我们提出了一整套指标来定量评估几种最先进的 LLM 所生成的合成数据集的多样性（即语言表达、情感和用户视角）和隐私（即重新识别风险和风格异常）。实验结果揭示了 LLM 在生成多样性和隐私保护的合成数据方面的显著限制。在评估结果的指导下，我们提出了一种基于提示的方法，以提高合成评论的多样性，同时保护评论者的隐私。

Keywords Synthetic Data Generation · Writing Style Diversity · Privacy · Prompt Optimization a · LLM

1 介绍

最近在人工智能技术方面的进展显著增加了对海量数据训练的需求。在获取高质量、多样化且注重隐私保护的数据集代价高昂或受到法律限制的情况下，合成数据通过减少对现实世界数据的依赖同时解决隐私问题，提供了一种有希望的替代方案 [1]。尽管拥有这些优势，合成数据往往缺乏真实数据所提供的语言多样性和变异性，可能导致偏见结果及在人工智能系统中的失真表现 [2]。此外，最近的研究表明，由大型语言模型生成的合成数据可能无意中包含来自训练阶段的原始数据记录，引发了对大型语言模型记忆和重复敏感训练数据能力的广泛担忧 [3, 4]。鉴于这些问题，提高合成数据的多样性和隐私性已成为确保道德人工智能发展的关键。

作为解决这些问题的第一步，本研究旨在定量评估 LLM 生成的合成数据的多样性和隐私性，探索权衡，并提出潜在的解决方案来增强多样性，同时保护用户隐私。特别是，我们聚焦于在线产品评论数据，因为由于其情感、语言表达和用户视角的混合，它们在用户的定制信息和变异性方面特别丰富，使其成为本研究的理想背景。

我们提出了一整套指标，用于评估：(1) 从词汇、语义和情感方面的多样性，以及 (2) 从个人可识别信息的存在和用户层面的风格独特性可能造成的隐私风险。通过将所提出的指标应用于真实用户数据和由不同大型语言模型 (LLMs) 生成的几个合成数据集，我们详细讨论了结果和主要发现。根据评估反馈，我们引入了一个自动提示优化流程，该流程根据评估不合格的指标自适应更新生成指令。具体来说，此流程为每个多样性或隐私标准（例如词汇变化或异常值控制）叠加定向提示优化，从而产生高质量、更加平衡的合成评论数据集。与静态提示相比，这种动态策略在极少人工调整情况下提高了对所需属性的整体覆盖率。

本研究对围绕合成数据的伦理和实践考虑的更广泛讨论做出了贡献，特别是在评论数据应用中，旨在促进更具包容性和安全性的 AI 系统。该工作的主要贡献总结如下：

- 我们提出了一套全面的指标来定量评估用户评论数据的多样性和隐私性。这些指标的有效性通过包含超过 250 万条真实用户评论的亚马逊评论数据集得到了验证。
- 我们通过多种最先进的 LLM，包括 GPT-4o 和 Claude 3.7 Sonnet，生成一组多样化的合成评论数据，并根据提出的指标评估这些合成数据的多样性和隐私性。我们分析和比较评估结果，以调查不同模型在合成数据生成方面的能力，以及写作风格多样性与作者身份识别隐私风险之间的权衡。
- 提出了一种基于迭代提示的增强方法，旨在在保护评审者隐私的同时增强写作风格的多样性。该方法的特点是具有自动提示优化组件，可根据指标失效动态调整指令，从而生成语义更加丰富、风格更多样化且符合隐私要求的评论。

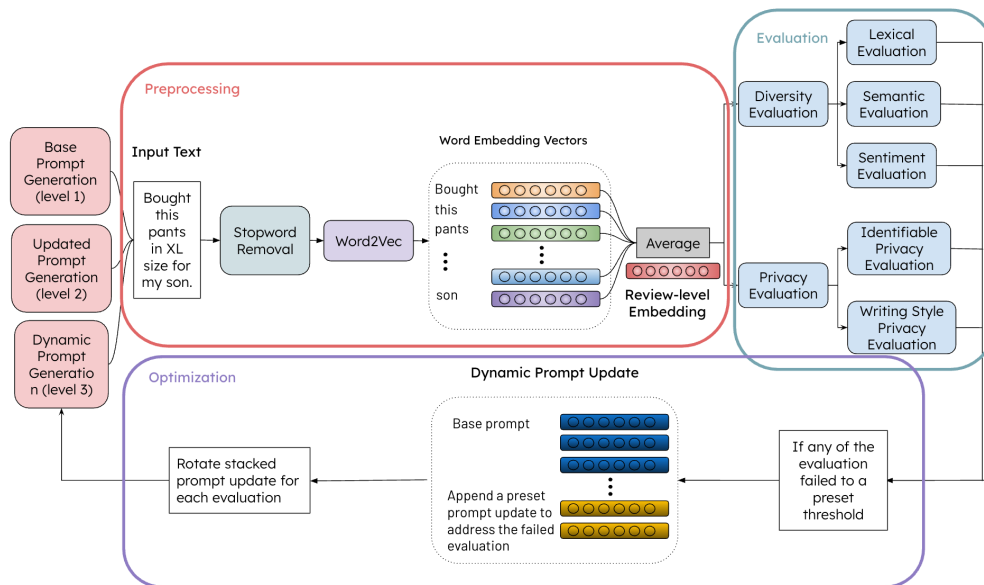


Figure 1: 评估框架概述

2 相关工作

大型语言模型（LLM）对于生成合成文本，在隐私限制下增强或替代真实数据至关重要。然而，仍然存在大量挑战，尤其是涉及到真实文本数据固有的复杂性、真实性和表示多样性。例如，Hämäläinen 等人 [5] 揭示合成数据表现出显著较低的多样性，并包含特定的、可辨识的异常。此外，Carlini 等人 [3] 强调了在某些情况下，LLM 会记忆并再现训练数据中的精确短语或个人可识别信息（PII），暴露出明显的脆弱性，易于提取敏感信息。这些问题促使研究人员建立严格的指标和评估协议来评估合成数据的质量。研究人员经常使用如 Distinct- n 、Self-BLEU、熵、句长和通过基于嵌入的聚类方法的语义多样性等指标来量化词汇和结构多样性。Shaib 等人 [6] 通过比较基于单词对平均相似性和各种标记/类型比率的指标，建立了一个标准化的多样性测量框架。在我们的工作中，我们采用了这种结构化的方法来评估词汇多样性。此外，我们基于最小生成树（MST）提出了多种语义多样性指标，以捕捉语义多样性方面。Yuan 等人 [7] 通过关联评论情感分布与相应评分值来评估合成评论文本的有效性，激发了本研究情感多样性指标的设计。Montahaei 等人 [8] 专注于文本生成模型质量和多样性的联合评估。[9] 联合评估了合成数据的保真度和隐私保护完整性。在这一研究线索的基础上，我们同时评估了模型多样性和隐私，以进一步阐明这两者之间的权衡，从而加深对其相互作用的理解。为了提高合成数据生成的质量，一些研究结合迭代细化技术、细致的提示工程和明确的偏差控制机制 [10]。Barbierato 等人 [10] 介绍了一种基于概率网络的方法，该方法明确管理合成数据集内的偏差和公平性，允许对特征相关性和相互依赖性进行精确操作。Whitney 和 Norman [11] 警告潜在风险，如“多样性漂白”，即合成数据集可能表面上看起来多样化，但未能解决潜在的代表性偏见和同意规避问题。为了减轻这些问题，我们提出了一种基于良好定义的综合指标评估的迭代提示方法，涵盖多样性和隐私的考虑。该方法在性能上体现出良好的表现，同时提供了更高的灵活性。

3 方法

在本节中，我们讨论了如图 1 所示的所提出的评估框架，它包括一些针对评论嵌入的预处理步骤，从多样性和隐私角度提出的一套全面的定量指标，以及一个动态提示生成模块，根据需要调整提示内容。

为了定量评估合成生成的评论数据的多样性和隐私性，本研究实施了一个结构化的预处理和向量化工作流程，如图 1 所示。该工作流程首先去除诸如“the”、“is”、“at”等停用词，这些词频繁出现在文本中但提供的有意义的语义信息很少。然后使用 Word2Vec 嵌入模型将评论转换为数值表示。与使用可能引入不相关外部语义偏见的通用预训练嵌入模型不同，专门针对所研究的数据集训练 Word2Vec 确保嵌入精确捕捉数据集特有的语言特征。为了定量表示每条评论，通过对每个评论中的所有词向量进行平均，计算出一个句子级别的嵌入。最后，将嵌入输入到多样性和隐私评估模块中进行进一步分析。更多详情请参见附录 A。

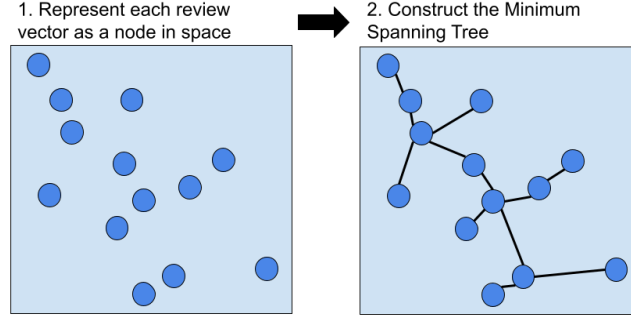


Figure 2: 最小生成树的视觉表示，其中每个节点代表一个单独的评论向量，边是评论向量之间计算的余弦距离。

3.1 评估指标

为了系统地评估数据集的多样性和隐私风险，我们提出了一套全面的指标。具体而言，多样性从三个方面进行评估：(1) 包括词汇层面的词汇丰富度，(2) 评论层面的语义多样性，以及 (3) 情感多样性。提出了两种互补的隐私指标：(1) 个人可识别内容的存在，以及 (2) 用户层面的风格独特性，以衡量两种文本隐私风险：非故意记忆性和作者归属 [12]。两个指标均量化了可以在生成模型的训练集中包含数据时暴露个人用户隐私漏洞的信号。

我们提出通过利用基于 N -gram 的方法来评估词汇层级的丰富性，该方法分析从一元组（单词）到五元组（五个单词的序列）的连续单词序列。这种多层次的方法为分析局部词汇多样性和更广泛的语言模式提供了关键的洞察。具体来说，我们为每个 N -gram 级别采用了两个指标：词汇独特性比例（即，用 L_r 标记），这是相对于识别出的 N -gram 总数的独特 N -gram 的比例（即，表示词汇丰富性），以及规范化词汇熵（即，用 H_n 标记），这反映了这些词汇元素的分布复杂性（即，表示结构变异性）。较高的词汇独特性比例结合较低的规范化词汇熵可能表明词汇多样性，但短语结构的分布不均；而在两个指标中都取得高分则表示词汇和结构的多样性。

在本研究中，我们建议在嵌入空间中评估评论之间的语义多样性。具体来说，通过图 1 中所示的预处理，我们提取每个评论的嵌入，这作为下面两个评估指标的基础。

语义比率：不同嵌入向量与评论总数的比率。该指标评估数据集中重复的程度。较高的比率表示较少的重复和评论之间语义独特性较高，而较低的比率表示显著的重复，暗示多样性有限。

平均最小生成树边长。受 [13] 的启发，我们采用基于最小生成树（MST）的方法来评估评论层面的语义多样性。MST 的建立基于评论嵌入之间的两两语义距离（即，计算为余弦距离），如图 2 所示。采用 MST 是因为它在一个语义网络中用最少的边连接所有评论，同时保持连通性，确保多样的评论得以链接而无冗余连接。它还以一种突显结构可变性的方式捕捉评论之间的结构关系。平均 MST 边长提供了连接评论之间平均语义距离的洞察。较高的平均边长表明评论不太相似，从而反映出增加的语义丰富性和减少的冗余。

3.1.1 情感多样性

在这项研究中，我们将情感多样性视为一个重要的指标，用于衡量合成评论是否能够现实地表现出不同评分级别下各种类似人类的满意或不满意的模式。其基本原理是，合成评论应该以多样化的情感生成，这样一星级的评论中负面情感的比例会更高，而五星级的评论中负面情感的比例会更低。

计算过程如下。首先，我们根据相应的评分值将所有合成评论分为五个部分。在每个部分中，对每条评论进行情感分析。具体来说，我们选择了 Flair 情感分类器 [14]，该分类器利用了 distilBERT 嵌入，因为它在将细微的文本情感分类为二元值（即正面或负面）时具有稳健性和精确性。接下来，我们计算每个部分内的情感分布，并评估各部分间的情感分布是否能够明显反映用户满意度。具体来说，在每个部分内，我们计算情感分数为具有正面情感的评论的百分比。

为了评估合成评论的情感多样性，我们建立了一个理想的线性情感分布作为基准。这个基准反映了一种直观的期望，即情感评分应该随评分单调增加（例如，评分 1 主要为负面，评分 5 主要为正面）。对于一个合成评论数据集，将其各段的情感分布与这个理想线性分布进行比较，可以直接揭示这些合成评论与理想人类情感模式之间的契合程度。较大的偏差突出了合成评论情感上的偏差。因此，我们建议用不同评分值下平均绝对误差（MAE）来衡量情感多样性，如等式 (1) 所示。

Table 1: 涉及隐私的评论在高实体/名词数量和密度情况下的代表性例子。命名实体用下划线表示；名词提及被突出显示。

Review Text	Entity Density	Nominal Density
“ <u>I</u> love this baseball <u>cap</u> . <u>I</u> graduated from the <u>University of Hawaii</u> with <u>my Bachelor’s degree</u> ...and <u>I</u> love advertising <u>Hawaii</u> on the <u>top</u> of <u>my</u> head! The many years <u>I</u> lived in <u>Hawaii</u> it was/is absolutely gorgeous, calm, safe, friendly and multi-ethnic. Great memories...thus, a happy cap to bring back happy memories.”	0.086	0.293
“ <u>I</u> have plantar fasciitis and have been trying and using various compression socks and sleeves... <u>I</u> ordered the large/extra-large because <u>I</u> take a <u>9</u> to <u>9.5</u> shoe... <u>I</u> ’m passing <u>them</u> off to my boyfriend... <u>These</u> are not only ‘fun’ but <u>they</u> are medically helpful for <u>my</u> plantar fasciitis...”	0.025	0.284
“Bought <u>this</u> in <u>XL</u> for <u>my</u> 11yo who is <u>5’8</u> and <u>110</u> .”	0.250	0.417
“ <u>My</u> granddaughter loves <u>these</u> !”	0.0	0.750

$$D_{sen} = \frac{1}{5} \sum_{i=1}^5 (1 - |y_i - \bar{y}_i|) \quad (1)$$

其中 y_i 表示片段 i 的正面评价百分比（即评分值等于 i ），而 $\bar{y}_i$ 表示线性基准的情感得分。

3.1.2 文本中的上下文标识符：命名实体和名词性提及

此指标旨在量化用户生成评论中与个人身份识别和情境揭示有关的信息的存在。我们关注两种对隐私风险有贡献的语言特征：命名实体，通常反映具体的现实世界参考（例如，“Jeff Bezos”，“西雅图”，“圣诞节”），以及名词性提及，表示个人叙述或角色的存在（例如，家庭成员、接收者、句子的主体）。这些元素作为评估 LLM 中意外记忆性风险的代理。

对于命名实体和普通提及，我们评估了两种互补信号：总计数和密度。总计数捕捉长且上下文丰富的评论——其中可能积累多个敏感术语并可用于重新识别，而密度则突出较短、更零散的评论，这些评论包含高浓度的揭示性术语，即使是单个字词也可能仍能进行关联推断。基于命名实体和普通提及，我们可以在内容层面评估数据集的潜在隐私泄露。高频率和密度的这些元素可能表明存在个人可识别或披露的信息，生成模型在训练期间可能会记住这些信息。因此，这些指标作为潜在非故意记忆性的指示器，特别是在训练于用户生成内容的 LLM 中。它们还为原始数据集与合成数据集之间隐私属性的比较提供了基础。

3.1.3 通过基于嵌入的用户画像进行风格异常检测

该度量指标侧重于在用户级别捕捉风格的独特性，以评估作者身份识别的风险。具有高度独特写作风格的用户可能更容易受到身份再识别攻击的影响，特别是在生成模型在合成输出中保留这些风格线索时。

具体来说，我们提出通过构建用户级别的嵌入向量来捕捉用户的写作风格，该向量计算为用户所有评论嵌入的平均值。如果一个用户的嵌入在全球和局部范围内都远离其他人，则被视为风格上的离群者。如果一个用户与其他用户的平均相似度显著低于大多数用户，则他/她是全局远离的；如果他/她与最近的邻居的距离很远，则是局部远离的。对于给定的数据集，我们建议使用识别出的风格离群者的数量（即 $|\mathcal{U}|$ ）来定量表示隐私风险，其中具有截然不同写作风格的用户越多，表明隐私风险越高。

为了评估用户与他人的全球风格距离，我们计算用户与数据集中所有其他用户（即，标记为 S_i ）的平均余弦相似度。然后，作为与所有用户的平均相似度分布相关的每个用户的相似度分数，计算其 z 分数（即， Z_i ）：

$$Z_i = \frac{S_i - \mu_S}{\sigma_S} \quad (2)$$

其中 μ_S 和 σ_S 分别是所有用户的 S_i 值的均值和标准差。较低的 z -得分表明该用户与大多数人整体不相似，因此是进一步本地检查的潜在异常值。具体来说，我们定义一个全局阈值 θ_g ，并提取所有 z -得分低于此阈值的用户作为全局独特用户（即， $\mathcal{U}_g$ ）以进行附加检查。

$$\mathcal{U}_g = \{u_i | Z_i \leq \theta_g\} \quad (3)$$

我们的局部基于距离的过滤步骤是受到 k -匿名性原则的启发。只有当一个用户的写作风格与数据集中其他用户的风格不相似时，他才能被认为是真正的风格上的异常值，从而增加作者身份归属的风险，这实际上相当于 $k = 1$ 。仅依赖全局相似性评分可能会错误地将用户的边缘群（数量达到数百甚至数千）归类为异常值，仅仅因为它们与总体人群有一定距离，尽管它们形成了一个相互一致的小组。由于我们的目标是检测真正的异常值，而不是小而内部一致的子群体，我们通过检查每个用户到其最近邻的距离来优化我们的候选集。

要实现这一步，我们对全局低 z 值分数（即 $\mathcal{U}_g$ ）的用户集应用本地基于距离的过滤器。对于从 $\mathcal{U}_g$ 中选择的每个用户 i ，我们计算其在嵌入空间中到其最近邻的余弦距离。最近邻距离 d_i^{nn} 的计算方法为：

$$d_i^{nn} = \min_{j \neq i} d_{cos}(u_i, u_j) \quad (4)$$

请注意，用户 j ，作为与用户 i 最相似的用户，可以是整个数据集中任何一个用户，而不一定来自 $\mathcal{U}_g$ 。在识别出所有全局远距离用户的最近邻后，我们按他们的 d_i^{nn} 值降序排列。如果用户 i 被确认为风格异常值，那么：

$$\mathcal{U}_o = \{u_i \in \mathcal{U}_g | d_i^{nn} \geq \theta_l\} \quad (5)$$

其中 θ_l 是局部阈值。

通过这种异常值检测方法识别出的用户，其写作风格与数据集的其他部分相比表现出与众不同或特有的特征。通过将这些个体隔离出来，我们可以评估数据的哪些部分最有可能对作者归属风险产生不成比例的影响。这项分析为刻画数据集的风格隐私特征提供了一个有用的视角，并能为数据过滤措施、用户匿名化或有针对性的隐私保护策略提供信息。

3.2 通过评估引导反馈的自适应提示优化

为了提高 LLMs 生成的合成产品评论的质量和多样性，我们引入了一种动态提示优化流程，该流程根据多指标反馈调整其指令。我们的方法并不依赖于在所有数据生成周期中应用的固定提示，而是根据在预定义评估指标中的失败逐步调整提示。此方法可以在避免过度拟合到静态指令模板的同时，迭代地优化输出结果。

系统首先提供一个基础提示，指导 LLM 生成真实多样的产品评论。在每个生成周期之后，生成的批次会在六个核心维度上进行评估：

- 词汇层面的多样性：在评论中确保多样化的单词使用和表达。
- 评论级别语义多样性：鼓励独特的背景、用例和用户体验。
- 情感多样性：将情感基调与评论评分相匹配。
- 异常值检测：识别语义或风格上异常的评论。
- 独特性：避免生成重复的内容。
- 长度分布：强制执行一个固定的评论长度分布，例如 25 % 非常短（1-10 字），40 % 中等（11-40 字），25 % 长（41-80 字），以及 10 % 超长（80+ 字）

每个评估指标独立运行，并通知提示调整机制。如果生成的一批数据未通过任何指标，则会从与该指标相关的预定义提示池中提取相应的指令。这些针对特定指标的提示在下一个提示的专用部分中堆叠，按评估类型组织（例如，“LENGTH DIVERSITY GUIDELINES”）。每个部分最多积累三个活跃指令用于每个指标。当某个指标需要超过三个提示指令时，旧的指令会以循环方式替换，以维持一个指导的滑动窗口。这样的设计避免了提示无限增长，因为这可能导致指令膨胀，并增加大模型提示饱和的风险——即模型开始忽略或低估关键指令，因为上下文过多或相互竞争。限制指令的数量有助于确保每个指标获得集中且可解释的指导，同时保持总体提示的有效性和连贯性。

这一过程创建了一个反馈回路，随着时间的推移对模型的行为进行微调：模型反复暴露于强调失败维度的动态指令下，从而改善了评审的多样性、质量和现实性。整个流程保留所有生成的输出，从不直接舍弃批次。相反，它优先考虑迭代改进，同时积累高质量的样本。

4 结果与讨论

为了验证所提出的指标是否能有效反映用户评论的多样性和隐私性，我们首先将这些指标应用于一个大型真实用户数据集。该真实用户数据集包括 250 万条亚马逊产品评论。我们主要关注三列数据，包括用户 ID、评级值和评论文本。结果验证了所提出的指标能够有效反映真实用户数据的多样性和隐私风险。由于篇幅限制，详细结果包含在附录 C 中。

然后，我们将相同的指标应用于由最先进的 LLM（包括 GPT-4o 和 Claude 3.7 Sonnet）生成的合成数据集，以评估这些模型在评论多样性和隐私方面的生成能力。设计了三个不同级别的提示词，以生成多个合成数据集进行评估和比较。

- 提示级别 1：具有最小约束的基本提示。
- 提示等级 2：手动改进的提示，具有特定的约束条件。
- 提示级别 3：根据评估反馈迭代优化的自适应提示。

基本提示是生成一批约束最小的合成产品评论。提供了示例评论来指导语调和情感分布，同时禁止直接复制以确保原创性和隐私。这个初始提示作为评估生成数据中文本多样性和隐私敏感性的基线。然后，我们引入了第二级提示，具有特定领域重点（时尚和服装）和更严格的限制，以增强现实性。它强调词汇和语义多样性、真实感情模式以及评论之间的独特内容。来自此提示的合成数据使用之前相同的标准进行评估。第三级提示是如第 3.2 节中讨论的建议自适应提示方案，它根据多指标反馈调整其指令。结果如图 3 所示。

隐私分析 使用基础提示生成的合成数据往往较短且更为零散，词汇使用最少，表达细节有限。因此，这些评论很少包含敏感词或可识别的关注模式。相比之下，强调词汇和语义多样性、真实情感分布和独特内容的二级提示生成了更长且更具表现力的评论。这导致代词使用的显著增加以及敏感词的出现，尤其是那些涉及家庭关系（例如：“我的女儿”）、身体测量（例如，身高、体重）和服装尺寸的术语。例如，原始评论“为我的女儿购买了这款领结”被转化为“给我的女儿买了这顶豆豆帽，她从没脱下来过”——这是一种更丰富、更个人化的表述，引入了亲属关系和情感联系。这些发现表明，尽管更严格的生成约束提高了真实性，但它们也可能增加在合成评论中出现与隐私相关内容的可能性。

到目前为止，自适应提示生成了最逼真的合成评论，具有自然混合的短小简洁的条目和较长的、内容丰富的叙述性评论。然而，这种逼真性的提高伴随着隐私风险的增加。我们注意到敏感内容显著增加，特别是涉及家庭和亲属的术语，这表明随着合成数据变得更加逼真，意外生成与隐私相关的语言的可能性也在增加。此外，嵌入级模式进一步突显了潜在的隐私问题。当应用全局 z -score 的 -1.0 阈值时，低于此阈值的用户与最近邻的余弦距离明显更高。这表明这些用户的写作风格比一般人群更为独特，从而增加了重新识别或暴露记忆化语言模式的风险。

多样性分析 在分析多样性评估时，我们首先查看了 Claude 1 级和 2 级的生成。对于 2 级提示，在高 N -gram 中词汇独特性比率有了显著的提高。这些结果表明，Claude 的 2 级提示在更高的 N -gram 级别上有效地产生了更多独特且多样的短语，且提示更为详细。语义多样性指标揭示了 2 级生成中的权衡。平均 MST 边长显著减少，这意味着评论之间的语义关系更紧密。然而，Claude 在 2 级达到完美的语义比率 1.0，这意味着没有重复的评论。虽然多样性指标没有评估每条评论中的词长，但 2 级提示大大增加了 Claude 和 ChatGPT 评论的长度。这种差异可以在 Claude 一级和二级生成的这两个示例评论中看到，分别是：“这个产品完全是垃圾。一天之内就坏了。”和“这些牛仔裤超出了我的期望。腰部非常合身，长度对我的身高也刚刚好。材料感觉既耐用又舒适，适合全天穿着。”因此，虽然 2 级生成的评论之间的语义关系更紧密，但即兴提示增加了评论的长度和独特性。还值得注意的是，情感多样性的 D_{sen} 评分没有显著变化。

在分析 ChatGPT 生成的数据时，可以看到其在两个级别之间的词汇多样性指标上表现出显著的一致性。ChatGPT 的数据中也观察到了 Claude 数据中相同的语义趋势，即级别 2 的平均 MST 边长有所下降。然而，与 Claude 不同的是，语义比率保持非常相似。同样，可以观察到 ChatGPT 的级别 2 数据中，评论的长度明显长于级别 1 生成的数据。类似的权衡也出现了，即评论本身变得更长，但语义关系变得更接近而多样性减少。然而，词汇和情感多样性能够保持非常相似，表明一个更具描述性的提示可以维持这两种类型的多样性。

使用 Level 3 提示技术的最显著改善是平均 MST 边长 (0.000914) 和 N 元语法 L_r 得分。平均边长大大超过了其他合成生成级别的平均值。这个显著的改进以及 0.9980 的语义比率得分表明这种技术能够保持语义独特性，几乎每条评论都是独立的，同时其评论在语义上比以前的技术更远。此外，每个 N 元语法级别的 L_r 和 H_n 得分都比以前的生成技术高。这个增加的 L_r 得分显示了在词汇层次上的更高变异性，并且在这个生成的数据集中重复的词更少。这个分数与 H_n 值相结合，表明分布比之前的几代至少同样均匀。自适应提示在提高语义和词汇多样性方面都有显著增长，同时保持了各个评论的独特性。

5 结论与未来工作

本文提出了一种综合的合成客户评论数据生成系统，该系统能够生成无限的、高质量的、保护隐私的评论数据。这些数据可以用于微调第三方的大型语言模型 (LLMs)，从而增强我们对 eBay 客户的理解并提高客户满意度。此外，这种方法可以广泛应用于匿名化和平衡任何自由格式的文本数据，为任何基础 LLM 模型的预训练提供了宝贵的资源。

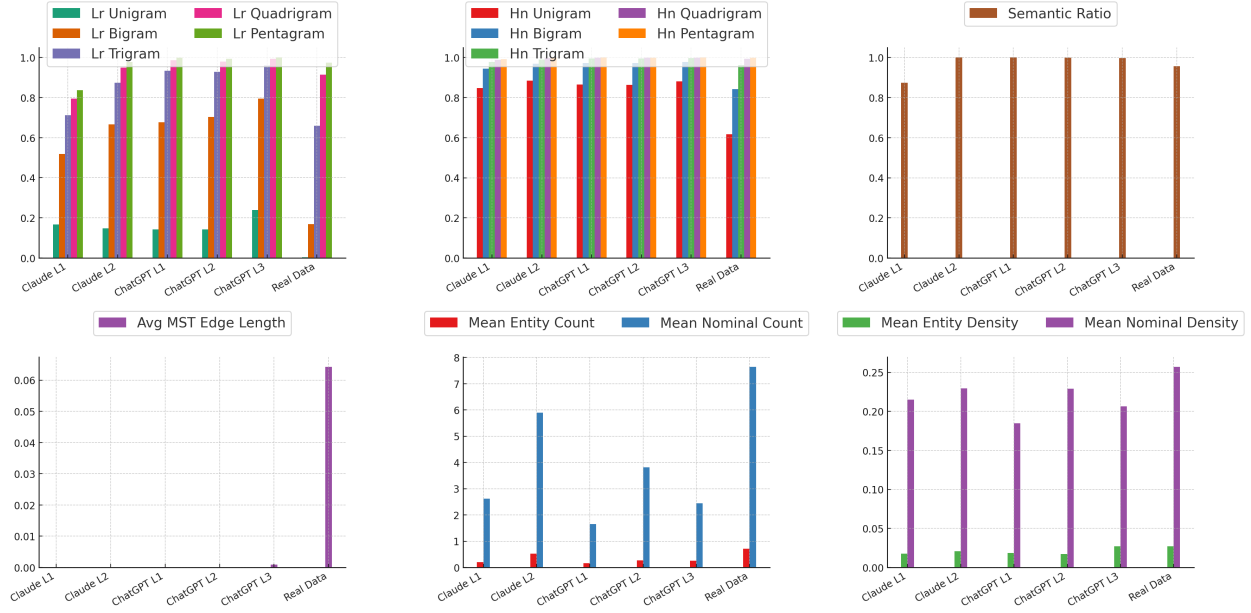


Figure 3: 每个评估指标结果的可视化。完整详细的数据表请参见附录 D。

虽然我们当前的方法使用的是人工设计的评估指标和一组静态的提示改进列表，但未来的工作可以探索基于学习的技术，以实现更高效和智能的提示适应。一个有前景的方向是将提示优化问题框架化为一个强化学习任务，其中 LLM 充当一个黑箱环境，提示更新是为了最大化奖励（例如，批次质量得分）而选择的动作。

另一个扩展涉及合并额外的评估指标，如命名实体的多样性、各评论子集中的情感一致性，或与现实世界数据分布的一致性。最后，用生成的或学习得到的机制替换人工选择的提示池，可以在不同行业中提供更大的可扩展性和适应性。

References

- [1] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*, 2024.
- [2] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*, 2024.
- [3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- [4] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025.
- [5] Perttu Hämmäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [6] Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenikova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores, 2025.
- [7] Zhigang Yuan, Fangzhao Wu, Junxin Liu, Chuhan Wu, Yongfeng Huang, and Xing Xie. Neural review rating prediction with user and product memory. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2341–2344, New York, NY, USA, 2019. Association for Computing Machinery.
- [8] Ehsan Montahaei, Danial Alihosseini, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models, 2019.

- [9] Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1), December 2024.
- [10] Enrico Barbierato, Marco L. Della Vedova, Daniele Tessera, Daniele Toti, and Nicola Vanoli. A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences*, 12(9), 2022.
- [11] Cedric Deslandes Whitney and Justin Norman. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1733–1744. ACM, June 2024.
- [12] Sakib Shahriar, Rozita Dara, and Rajen Akalu. A comprehensive review of current trends, challenges, and opportunities in text data privacy. *Computers & Security*, 151:104358, 2025.
- [13] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–35, 2021.
- [14] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59, 2019.
- [15] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. Software available from <https://spacy.io>.

Appendices

6

附录 A: 预处理工作流程

为了定量评估合成生成的评论数据的多样性和隐私性，本研究实施了一个结构化的预处理和向量化工作流程，如图 1 所示。该工作流程从文本预处理开始，然后进行词嵌入，最后对每个评论进行句子级别的嵌入。初始的预处理步骤包括去除诸如"the"、"is" 和"at" 之类的停用词，尽管这些词在文本中频繁出现，却几乎不提供有意义的语义信息。去除停用词显著减少了数据中的噪音，强调了具有明确语义表现的富含情感的有意义的词语的存在。通过消除这些常见但在语义上不重要的词，随后的嵌入过程能够有效地捕捉每个评论的真实内容和情感细微差别，提高了下游评价的准确性。

在预处理之后，评论将通过 Word2Vec 嵌入模型转换为数字表示。这种方法之所以被特别选择，是因为它能够基于评论语料库中的上下文共现关系捕捉到词语之间细微的语义关系。与可能引入无关外部语义偏见的通用预训练嵌入模型不同，专门在研究中数据集上训练的 Word2Vec 可确保嵌入精确反映数据集中独特的语言特征。句子参数明确提供来自评论数据集的预处理标记，这意味着 Word2Vec 模型仅根据这些评论中存在的词汇和词语模式学习语义关系。Word2Vec 模型的一个关键参数是窗口大小（记作 W_s ），它决定了在每个词的嵌入中，计算其贡献的相邻词语数量。在这项工作中，我们设置 $W_s = 5$ 来捕捉足够的上下文信息，以建立有意义的语义关系，同时平衡语义上下文的深度与计算资源及捕捉过于遥远、可能无关上下文的风险之间的关系。另一个关键参数是每个词的嵌入向量的维度。在本研究中，它被设置为 100，以降低计算复杂性，同时保持足够的表示复杂性以捕捉有意义的语义差异。

为了定量表示每个完整的评论，通过对每个评论中包含的词向量进行平均来计算句子级别的嵌入。这种平均方法之所以被选用，是因为它提供了一种平衡且具有代表性的嵌入，减少了来自个别异常词或影响过大的词造成的扭曲风险。这确保了句子嵌入能够捕捉评论的整体语义精髓，从而在后续步骤中能够准确且具有可解释性的比较评论之间的差异性。这些比较在评估多样性和检测合成数据集内部的冗余性方面至关重要，因为它们量化了评论之间的语义相似性或不相似性。在预处理之后，嵌入将被输入到多样性和隐私评估模块中进行进一步分析。

7

附录 B: 提议指标的方程

词汇唯一性比率：对于每个 N-gram 级别，该比率通过计算唯一 N-gram（不同的词组合）相对于识别出的 N-gram 总数的比例，来量化词汇多样性。更高的词汇唯一性比率表明更高的词汇丰富度和更低的重复性，这对于评估合成文本的变化性和原创性至关重要。

在数学上，这表示为：

$$L_r = \frac{U}{T} \quad (6)$$

，其中 U 是唯一 N 元组的数量， T 是该特定 N 元组级别的 N 元组总数。

归一化词汇熵：熵评估每个 N-gram 层次分布中的随机性或不可预测性。给定 N-gram 层次的熵 H 通过以下公式计算：

$$H = - \sum_{j=1}^U p_j \log_2(p_j) \quad (7)$$

$$p_j = \frac{f_j}{T} \quad (8)$$

其中 p_j 表示 j^{th} 唯一 N-gram 的相对频率（概率）。 p_j 由 f_j 即 j^{th} N-gram 的频率计算得出， T 则是该 N-gram 层次下所有 N-gram 的总数。对于熵 H ， U 表示该 N-gram 层次的唯一 N-gram 总数。

为了确保不同的 N-gram 水平之间的可比性，熵通过以下公式进行归一化：

$$H_n = \frac{H}{\log_2(U)} \quad (9)$$

。归一化的词汇熵值范围在 0 到 1 之间，较高的值表示分布更加均匀和语言模式更加多样化。

命名实体提取为了提取命名实体，我们利用来自 spaCy 自然语言处理 (NLP) 库的 `en_core_web_sm` 模型 [15]，该模型识别诸如人物、地点、组织和日期等实体。该模型处理每个评论并输出所有命名实体及其类型。对于每个评论，我们提取两部分信息：(1) 命名实体的总数（即， e_i 对于评论 i ），以及 (2) 实体密度（即， ρ_i 对于评论 i ），定义为命名实体的数量与标记总数的比率，不包括空格和标点符号（即， T_i 对于评论 i ），如方程式 (10) 所示。

$$\rho_i = \frac{e_i}{T_i} \quad (10)$$

为了评估整个数据集中的实体暴露，我们计算了以下统计数据：

- 平均实体数量： $\mu_e = \frac{1}{N} \sum_{i=1}^N e_i$
- 平均实体密度： $\mu_\rho = \frac{1}{N} \sum_{i=1}^N \rho_i$
- 最大实体计数： $\max_e = \max_{1 \leq i \leq N} e_i$
- 最大实体密度： $\max_\rho = \max_{1 \leq i \leq N} \rho_i$

这些统计数据提供了对现实世界引用整体水平的洞察，以及可能带来更高隐私风险的极极端案例。

名词指称提取除了命名实体之外，我们还提取名词指称，这些是指代句子中特定角色或参与者的词。其包括：

- 主语（例如，“我”，“我的女儿”）
- 对象（例如，“我儿子”，“送他的礼物”）
- 专有名词（例如，“Palo Alto”，“Stanford”）
- 代词（例如，“她”，“他们”）

。我们使用 spaCy 的依存解析器和词性标注器来识别这些词性和句法角色。对于每篇评论，我们提取两种信息：(1) 唯一名词词的总数（即， n_i 用于评论 i ），以及 (2) 名词密度（即， δ_i 用于评论 i ），其定义为唯一名词词数与排除空白的总词数之比（即， T_i 用于评论 i ），如等式 (11) 所示。这些特征可以提供关于文本在多大程度上依赖参与者指称的洞察，这可以间接揭示用户的意图、关系或人口统计信息。

$$\delta_i = \frac{n_i}{T_i} \quad (11)$$

同样，对于整体评估，我们计算以下四个统计数据：

- 平均名义计数： $\mu_n = \frac{1}{N} \sum_{i=1}^N n_i$
- 平均名义密度： $\mu_\delta = \frac{1}{N} \sum_{i=1}^N \delta_i$
- 最大标称计数： $\max_n = \max_{1 \leq i \leq N} n_i$
- 最大标称密度： $\max_\delta = \max_{1 \leq i \leq N} \delta_i$

每个用户的平均余弦相似度

我们计算每对嵌入向量在用户级别的平均余弦相似度。设 N 表示用户总数，对于用户 i ，其平均余弦相似度 S_i 计算如下：

$$S_i = \frac{1}{N-1} \sum_{j=1}^N \cos(u_i, u_j) \quad (12)$$

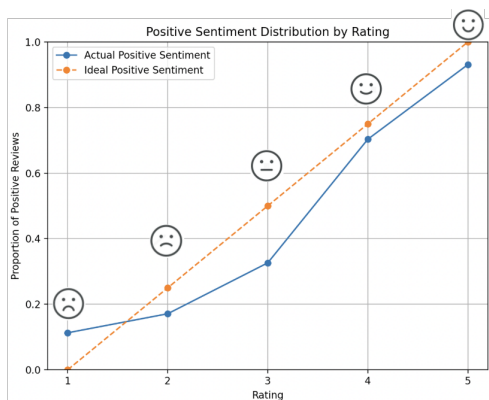
8

附录 C：真实用户数据的度量验证结果

8.1

1. 多样性评估

情感多样性分析 在分析真实用户评论的情感多样性时，我们观察到一种趋势，这种趋势与作为基准建立的理想情感分布线一致。图 4 中的图表和随附表格清楚地展示了这种一致性，显示出正面情感评论的比例随着星级评分的增加而持续上升。



Rating	Negative	Positive
1	88.72 %	11.28 %
2	82.91 %	17.09 %
3	67.38 %	32.62 %
4	29.68 %	70.32 %
5	6.91 %	93.09 %
D_{sen} score: 0.9036		

(b) 实际用户数据值的情感分布和计算出的 D_{sen} 分数

(a) 真实用户数据的正面情绪分布（蓝色）与理想线性分布（橙色）的比较。

Figure 4: 用于情感多样性评估的真实用户数据的图形和表格表示。

根据方程 1 计算得出的情感多样性得分 D_{sen} 约为 0.9036，这表明真实用户评论非常符合理想的情感分布。这个高分突显了一个事实，即实际人类生成的评论通常反映了自然趋势，即更高的星级评分与更积极的情感紧密相关。然而，情感对齐度并非完全为 1.0，显示出实际数据中的一些差异。这样的差异是由于人类评论的细微性质引起的，其中情感的主观表现并不总是严格与评分值一致。值得注意的是，最明显的偏差出现在 3 星级评分上，其积极情感为 32.62%，明显低于基准的中性值 50%。这种偏差是可以理解的，因为 3 星级评分通常代表不确定性，捕捉到混合甚至略带负面的情感，而不是纯粹的中立。例如，一条评为 3 星的真实人类评论写道“洗了几次就散了。”，表现出一种负面情感。逻辑上，如果评论者在批评产品并留下 3 而不是更高的 4 或 5 的评分，那么文本评论中通常可以发现负面情感。在其他评分水平上的轻微偏差，例如 1 星级类别，其记录了 11.28% 的积极率，进一步表明了人类情感的复杂性。即使是负面评分有时也包含积极的方面，如在这个例子中，评分为 1，“做得不错.. 订了名字 ‘Samantha’ 却得到了名字 ‘Ava’ ..”，评论者写了一句讽刺的话，却被解读为积极情感。

词汇层面的多样性分析 分析真实用户评论数据集的词汇多样性揭示了不同 N-gram 层次上的显著洞见。首先从一元组（单个词）开始，词汇独特率明显低至 0.0042。这表明重复率很高，许多单词在评论中频繁出现。出现这种结果是意料之中的，因为在大型语料库中，单个词自然会更频繁地重复出现，尤其是与被评论产品相关的常见形容词、名词和描述性术语。如表 1 所示，一些示例评论在同一评论中多次重复某些 N-gram，例如“夏威夷”或“足底筋膜炎”。相应地，一元组的归一化词汇熵较低，为 0.6176，表明词汇使用分布不均衡。这种不均衡是由常用词显著压制不常用词导致的，造成了分布倾斜。

随着我们转向高阶 N 元语法，词汇唯一性比率显著增加。例如，二元语法展示了 0.1696 的比率，明显高于单一词元，反映了重复的减少和独特性的增加。这一趋势持续进行，三元语法的词汇唯一性比率大幅跃升至 0.6588，紧接着四元语法增长到 0.9133，五元语法则进一步提高到 0.9743。增加的比率表明，高阶 N 元语法本质上更具独特性，并且在整个数据集中重复出现的频率大大降低。

类似地，归一化词汇熵值随着 N 元组大小的增加而增加，并随着 N 元组阶数的增长接近完美分数。例如，归一化词汇熵从二元组的 0.8419 上升到五元组的高达 0.9987。这些较高的值表明在高阶 N 元组中存在更均匀的分佈，突出了短语和表达的多样化使用。这个趋势的出现是因为较长的短语相较于单词或较短的组合，自然具有更大的独特性和较少的重复性。

N-gram Level	Lexical Uniqueness Ratio	Normalized Lexical Entropy
Unigram	0.0042	0.6176
Bigram	0.1696	0.8419
Trigram	0.6588	0.9619
Quadrigram	0.9133	0.9937
Pentagram	0.9743	0.9987

(a) 基于真实用户数据的词汇丰富性评估

Evaluation Metric	Score
Average MST Edge Length	0.0642
Semantic Ratio	0.9672

(b) 真实用户评论的评论级语义多样性评估总结。

Figure 5: 词汇与语义多样性在真实用户评论中的评估。

评论级语义多样性分析 在评估由大约 250 万条评论组成的真实用户评论数据集的评论级别语义多样性时，由于规模原因，完整计算最小生成树（MST）在计算上证明是具有挑战性的。具体来说，生成并处理一个 250 万乘以 250 万的完整成对余弦距离矩阵是不可行的。因此，我们采用了一种简化但有效的方法，通过使用 $k=30$ 的最近邻方法来构建一个近似 MST。尽管这种方法不能生成一个精确的 MST，但它提供了一个有效且在计算上切实可行的大数据集内在语义多样性的近似值。

平均最小生成树（MST）边长衡量了 MST 中直接连接的评论之间的平均语义距离，并且只考虑不同的不为零的边。在此分析中，如图 5b 所示，平均边长为 0.064，这表明在 MST 中，每个评论（节点）平均连接到其他相对语义接近的评论。虽然这个值乍看之下可能显得较低，但重要的是要认识到 MST 固有地代表了连接所有节点所需的最小边集。因此，MST 的边长通常较短，因为其构建优先考虑的是最小语义距离，捕捉基本的语义关系而非最大差异。尽管如此，在自然的人类评论中，给定数据集中常会出现重叠或相似的评论。

然而，单独解释平均边长可能会在没有额外背景信息的情况下产生误导。例如，一个包含许多重复评论的数据集可能由于只对不同节点进行平均计算而导致一个具有欺骗性的高平均边长。因此，引入语义比率（在这个数据集中为 0.96）是至关重要的。这个比率量化了相对于总评论数量的不同语义嵌入的比例，从而验证了平均边长的代表性。一个高的语义比率（例如 0.96）证实大多数评论是不同的，增强了信心，即计算出的平均边长准确地反映了真正的语义多样性，而非来自大量重复的伪象。

在检查评论数据集时，我们观察到命名实体和名词提及提取揭示出的敏感信息类型往往聚集成五个常见类别：(1) 个人姓名，(2) 亲属或家庭成员提及（例如，“我的女儿”，“孙子”），(3) 身体特征和健康描述，如身高、体重或医疗状况，(4) 衣服尺寸和合身相关细节，以及 (5) 地理或地点特定的提及。这些类别反映了现实世界中的标识符，当被下游语言模型保留或再现时，可能会增加隐私风险。如预期的，具有高实体数和名词数量的评论通常较长，并提供丰富的上下文细节。这些评论经常密集提及个人经历，使其更可能包含可识别的内容。

实体密集型评论。在高实体数量的情况下，我们观察到一种一致的模式：敏感词语经常属于以下类别：(1) 个人姓名，(4) 衣服尺码，和 (5) 地理位置，并且部分涉及 (2) 亲属关系和 (3) 身体特征。例如，“十三岁”这样的年龄参考（通常标记为 DATE）和“160 磅”或“5 英尺 8”这样的身体指标（常标记为 CARDINAL）通常会被提取。这些短语虽然本身不是直接的身份识别信息，但在与其他上下文细节结合时，特别是在较长的评论中多个此类线索同时出现时，可能会增加重新识别的风险。

以代词为主的评论。另一方面，高名词密度的评论通常以大量使用代词和语法性指代自我或他人（例如“我”、“我的”、“他”、“他们”）为特征。尽管单个代词可能看似无伤大雅，但其反复使用通常表明叙事结构和特定参与者的介入，这可能隐含地揭示角色、关系和用户视角。当与周围文本上下文结合时，这些指代可能促使对评论者身份、角色或情况的联想推理，特别是当评论提到与特定个人的具体行为或互动时。具有高实体密度或高名词密度的评论往往非常简短，通常少于十个字。这与我们的预期一致，因为数据集中大约 30 % 的评论低于此长度阈值。然而，在这些简短且破碎的评论中，我们观察到 spaCy 的命名实体识别性能显著下降。在人工检查的 1,000 条高实体密度评论中，只有 10–20 % 确实包含敏感术语；其余评论主要是由于大写而触发的误报。例如，在“Cheaply made”这样的短语中，单词“Cheaply”被错误地标记为 ORG（组织）。这些错误源自 spaCy 对句法和语义上下文的依赖：其实体识别器利用来自邻近标记、句子结构和语言模式的上下文线索，而这些在线碎片化和非正式评论中往往缺失或退化。缺乏足够的上下文时，模型倾向于依赖诸如大写特征等表面特征，从而导致不可靠的实体预测。相比之下，名词密度高的评论问题相对较少。因为代词通常较少依赖于扩展上下文，它们被提取出的可靠性较高。然而，我们观察到偶尔将大写的形容词或感叹词错误归类为专有名词的现象——例如，在“Super Cute!”中，“Super”和“Cute”都被错误地标记为了 PROP。这些限制突出显示了将通用目的的自然语言处理模型应用于结构最小的非正式用户生成文本时的挑战。

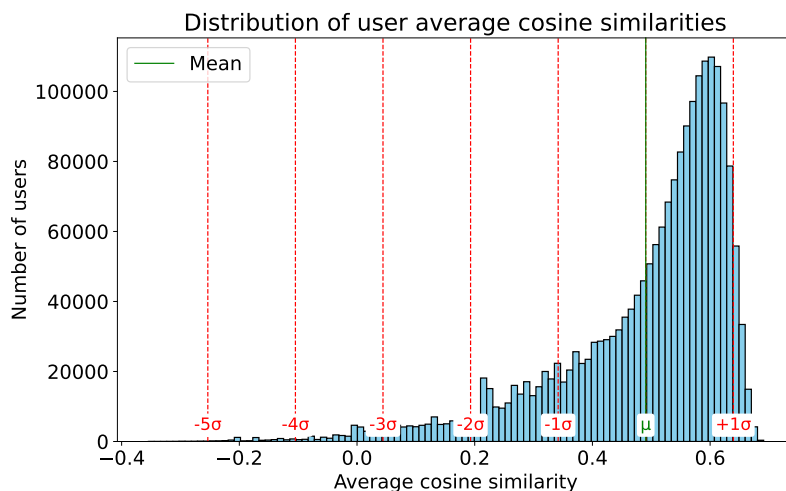


Figure 6: 用户平均余弦相似度的分布，垂直线标记出均值 (μ) 和标准差 ($\pm 1\sigma$ 、 $\pm 2\sigma$ 等)。相似度得分较低的用户（例如，超过 -2σ ）可能是潜在的风格离群者。

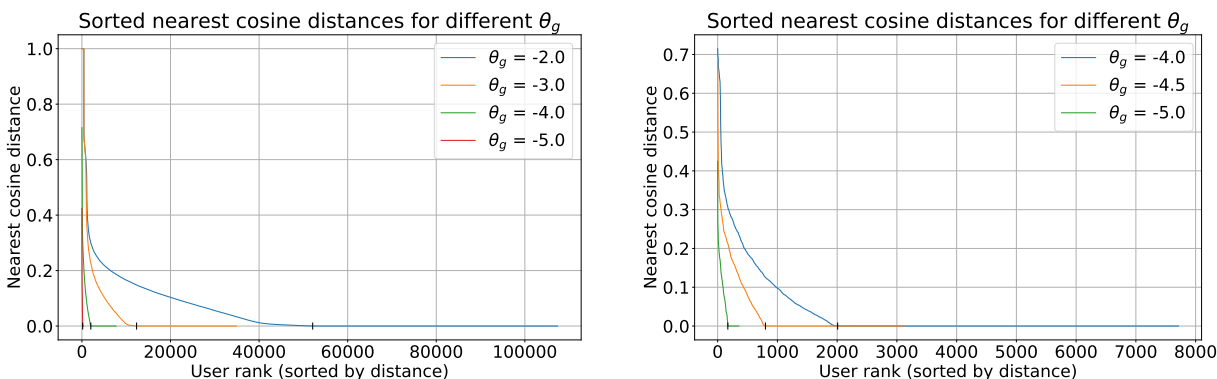
表格中 1 的示例被选取来代表四种不同情景的隐私相关内容：高实体数量、高名词数量、高实体密度和高名词密度。总体而言，这些示例捕捉了一系列隐私相关的模式——从包含详细个人信息的长篇、语境丰富的评论，到即使是几个字也可以透露关系或敏感特征的简短但密集的评价。其中包括对教育背景（例如，“夏威夷大学”、“学士学位”）、地理位置（例如，“夏威夷”）、服装尺寸（例如，“XL”、“大号/加大号”）、身高和体重等身体特征（例如，“5’8”、“110”）、医疗状况（例如，“足底筋膜炎”）、家庭关系（例如，“孙女”、“男朋友”），以及购买背景（例如，“9 到 9.5 码鞋”、“亚马逊”、“\$ 12.99 到 \$ 16.99”）的引用。

8.2

3. 用户级别的风格分析和异常分析

平均余弦相似度的分布 作为识别风格异常值的第一步，我们检查了每个用户与数据集中所有其他用户之间的平均余弦相似度得分的分布。图 6 展示了这种分布，每个值反映一个用户的写作风格与其他用户的相似程度。分布是左偏的，大多数用户聚集在较高的相似值附近，并且有一个延伸到较低相似度的长尾。这条尾巴包括与其他用户风格上有较大距离的用户，他们是进一步检查的候选者。

为了评估全球距离较远的用户的孤立程度，我们分析了他们在嵌入空间中的最近邻距离。图 7a 和图 7b 显示了在逐渐严格的全球阈值（即， $\theta_g \in [-2, -5]$ ）下，用户的最近邻距离的排序情况。对于每个全球阈值，排序后的最近邻距离曲线表现出两个部分的结构。第一部分包括从相对较高的余弦距离（通常接近 1.0）急剧下降到接近零的情况，随后是一个平坦的尾部，其中剩余用户与其最近邻居的距离非常小。随着全球阈值变得更加极端（例如， $\theta_g \in \{-4.0, -4.5, -5.0\}$ ），这种两部分的形状变得越来越明显。在我们的分析中，我们将每条曲线的初始急剧下降部分的用户视为最终的风格离群值，因为他们在全球和本地都是远离的——远离一般人群并且与任何附近的风格群体不紧密对齐。为了定义每条曲线的陡降区域和平坦区域之间的边界，我们使用本地距离阈值 $\theta_l = 10^{-4}$ 作为平坦化的截止。这个阈值是可调的，可以根据识别风格离群值所需的敏感度进行调节。



(a) 在 $\theta_g \in [-2, -5]$ 的范围内的全局阈值的最近邻距离曲线，说明了越来越严格的全局过滤效果。
 (b) 低阈值 $\theta_g \in \{-4.0, -4.5, -5.0\}$ 的曲线，突出了局部孤立用户和接近风格集群的用户之间的分离。

Figure 7: 在不同的全局阈值下，全局距离较远用户的最近邻距离曲线。

9

附录 D: LLM 生成的合成数据结果

9.1

一级提示

您是一个高级 AI 模型，负责生成合成产品评论。下面是几个产品评论的示例，每个示例都有多个属性。您的目标是生成新的合成产品评论，遵循相同的结构，并在评分和内容风格方面保持相似的分布。合成评论不应复制或泄露所提供示例中的任何具体信息。相反，它们应在遵循示例中观察到的模式和主题的同时展示创造力。

示例评论

评分, 评论, 用户 ID

< 在此处放置示例评论 >

说明

- 生成 $\{num_reviews\}$ 条新的产品评测。
- 每条评论应包括以下属性：评分、文本、用户 ID。
- 确保合成评论反映出示例评论中存在的评分（1.0 到 5.0）的多样性和分布。
- 不要复制示例评论中的任何具体细节或短语。
- 生成的评论应对各种产品具有合理性和相关性。
- 保持示例评论中的风格和语气。

请现在生成合成产品评论。

9.2

二级提示

您是一个先进的 AI 模型，负责生成合成的产品评论。以下是几个关于时尚和服装产品评论的示例，每个示例都有多个属性。您的目标是生成新的合成产品评论，这些评论应遵循相同的结构，并在评分和内容风格上保持相似的分布。合成评论不应复制或泄露所提供示例中的任何具体信息。相反，它们应在遵循示例中观察到的模式和主题的同时体现创造力。

Evaluation Metrics	Claude L1	Claude L2	ChatGPT L1	ChatGPT L2	ChatGPT L3	Real Data
Unigram (L_r / H_n)	0.1672 / 0.8474	0.1480 / 0.8847	0.1419 / 0.8653	0.1422 / 0.8640	0.2386 / 0.8813	0.0042 / 0.6176
Bigram	0.5184 / 0.9439	0.6657 / 0.9692	0.6768 / 0.9721	0.7025 / 0.9729	0.7942 / 0.9786	0.1696 / 0.8419
Trigram	0.7114 / 0.9780	0.8739 / 0.9904	0.9335 / 0.9957	0.9274 / 0.9951	0.9592 / 0.9967	0.6588 / 0.9619
Quadrigram	0.7939 / 0.9876	0.9488 / 0.9970	0.9857 / 0.9993	0.9798 / 0.9988	0.9929 / 0.9996	0.9133 / 0.9937
Pentagram	0.8372 / 0.9915	0.9777 / 0.9988	0.9991 / 1.0000	0.9925 / 0.9996	0.9993 / 1.0000	0.9743 / 0.9987
Avg MST Edge Length	0.000043	0.000018	0.000078	0.000026	0.000914	0.0642
Semantic Ratio	0.8740	1.0000	1.0000	0.9990	0.9980	0.9572
D_{sen} Score	0.8999	0.8938	0.9211	0.9234	0.9055	0.9036
Mean entity count	0.2040	0.5260	0.1636	0.2650	0.2619	0.7106
Mean nominal count	2.6180	5.8930	1.6534	3.8184	2.4506	7.6358
Mean entity density	0.0177	0.0206	0.0187	0.0172	0.0272	0.0269
Mean nominal density	0.2151	0.2295	0.1849	0.2291	0.2063	0.2569
1st percentile d_i^{nn}	1.485×10^{-4}	3.463×10^{-5}	2.887×10^{-4}	6.491×10^{-5}	0.0064	0.1847

Figure 8: 评估指标用于比较模型生成的数据和真实数据。

示例评论

评分, 评论, 用户-id

< 在此处放置示例评论 >

说明

- 生成 $\{num_reviews\}$ 篇关于时尚和服装产品的新评论。
- 每个评论必须包含的属性有：评分、评论、用户 ID。
- 反映评分（1.0 到 5.0）的真实多样性和分布。
- 避免复制示例评论中的细节或短语。
- 生成与各种产品相关的合理评价。
- 保持现实风格、语调和情感的多样性。

重要约束

- 确保评论反映每个评分的情感能够真实且平衡地分布，符合典型的客户评论模式。
- 保持高词汇多样性（独特的措辞，多样的词汇）。使用多样的语言、同义词和表达方式。避免重复使用单词、短语或句子结构。
- 确保语义的区别性（每个评论必须描述独特的体验或产品方面）。确保每个评论分别描述独特的体验、产品或各个方面，各自有显著不同。
- 避免生成看似不真实、通用或与典型客户反馈脱节的评论。

请现在生成合成的时尚和服装产品评论。

9.3